# Lung and colon cancer classification using medical imaging: a feature engineering approach

Aya Hage Chehade, Nassib Abdallah, Jean-Marie Marion, Mohamad Oueidat, Pierre Chauvet

# Lung and Colon Cancer Classification Using Medical Imaging : A Feature Engineering Approach

**Aya Hage Chehade** ( aya.hagechehade@etud.univ-angers.fr )
University of Angers: Universite d'Angers
**Nassib Abdallah**
University of Angers: Universite d'Angers
**Jean-Marie Marion**
University of Angers: Universite d'Angers
**Mohamad Oueidat**
Lebanese University: Universite Libanaise
**Pierre Chauvet**
University of Angers: Universite d'Angers

**Research Article**

**Keywords:** Lung Cancer, Colon Cancer, Histopathological Images, Machine Learning, Feature Engineering, Image Classification

**Posted Date:** January 3rd, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1211832/v1

# Lung and Colon Cancer Classification Using Medical Imaging : A Feature Engineering Approach

Aya Hage Chehade[1*], Nassib Abdallah[1,2], Jean-Marie Marion[1], Mohamad Oueidat[3] and Pierre Chauvet[1]

[1*]LARIS, SFR MATHSTIC, Univ Angers, France.
[2] LaTIM, INSERM, UMR 1101, Univ Brest, France .
[3]Faculty of Technology, Lebanese University, Lebanon.

*Corresponding author(s). E-mail(s):
aya.hagechehade@etud.univ-angers.fr;
Contributing authors: nassib.abdallah@univ-angers.fr;
marion@uco.fr; mohoueidat@yahoo.com; pierre.chauvet@uco.fr;

**Abstract**

Lung and colon cancers are the most common causes of death. Their simultaneous occurrence is uncommon, however, in the absence of early diagnosis, the metastasis of cancer cells is very high between these two organs. Currently, histopathological diagnosis and appropriate treatment are the only possibility to improve the chances of survival and reduce cancer mortality. Using artificial intelligence in the histopathological diagnosis of colon and lung cancer can provide significant help to specialists in identifying cases of colon and lung cancers with less effort, time and cost. The objective of this study is to set up a computer-aided diagnostic system that can accurately classify five types of colon and lung tissues (two classes for colon cancer and three classes for lung cancer) by analyzing their histopathological images. Using machine learning, features engineering and image processing techniques, the five models XGBoost, SVM, RF, LDA and MLP were used to perform the classification of histopathological images of lung and colon cancers that were acquired from the LC25000 dataset. The main advantage of using machine learning models is that they allow for better interpretability of the classification model since they are based on feature engineering; however, deep learning models are black box networks whose working is very

difficult to understand due to the complex network design. The acquired experimental results show that machine learning models give satisfactory results and are very precise in identifying classes of lung and colon cancer subtypes. The XGBoost model gave the best performance with an accuracy of 99% and a F1-score of 98.8%. The implementation and the development of this model will help healthcare specialists identify types of colon and lung cancers. The code will be available upon request.

**Keywords:** Lung Cancer; Colon Cancer; Histopathological Images; Machine Learning; Feature Engineering; Image Classification.

# 1 Introduction

According to the World Health Organization, cancer is considered one of the most common causes of mortality in the world. Cancer cells acquire autonomous growth, genetic instability and significant metastatic power. Among the most frequently affected organs, colon and lung cancers account for the highest number of deaths. Lung cancer accounts for 18.4% of cancer-related deaths, while colon cancer accounts for 9.2% of all cancer-related deaths worldwide [1]-[2]. The rate of simultaneous occurrence of lung and colon cancer is approximately 17%. Although this frequency is unlikely, but in the absence of an early diagnosis, cancer cells metastasis is very high between these two organs [3]. Currently, appropriate treatment and early diagnosis are the only way to reduce cancer mortality [4]. Indeed, the earlier a person is diagnosed, the better the management, and the greater the chance of recovery and survival of the patient.

Various tests such as imaging sets (x-ray, CT scan), Sputum cytology, and tissue sampling (biopsy) are done to look for cancer cells and rule out other possible conditions. While performing the biopsy, evaluation of the microscopic histopathology slides by experienced pathologists is essential to establishing the diagnosis [5] and defines the types and subtypes of cancers [6]. To automatically diagnose colon and lung cancers, this study relies solely on histopathological images. Histopathological images are widely used by health specialists for diagnosis, and they are very important in predicting patients' chances of survival. Traditionally, in order to diagnose cancer by examining histopathological images, health specialists have to go through a long process; however, it is now possible to perform this process in less time and effort with the technological tools available [3]. Recently, artificial intelligence technologies are known for their ability to examine data faster and make decisions.

Machine learning (ML) is a subfield of artificial intelligence (AI) that allows machines to learn a specific task through experience with the data sets to which they are exposed, without explicit programming [7]. Machine learning algorithms are used in biomedical applications for the prediction and classification of several types of signals and images. Deep learning (DL) algorithms

have been developed to enable machines to handle large-dimensional data like multidimensional anatomical images, and videos. DL is a subfield of machine learning that structures algorithms in layers to create an "artificial neural network", based on the structure and function of the human brain [8].

In previous research articles, most of the authors considered using DL to classify colon and lung cancer images at the same time. Some authors have focused on the lung cancer classification, while others have concentrated entirely on the classification of colon cancer.

There are few works for only the colon cancer classification. For instance, Bukhari et al. [9] used three convolutional neural networks architectures: ResNet-18, ResNet-30, and ResNet50. ResNet-50 achieved the highest accuracy of 93.91%, followed by ResNet-30 and ResNet-18 with an accuracy of 93.04% each.

To classify histopathological images of lung cancer into three-class, Hatuwal et al [10] used Convolution Neural Network (CNN). The classification result obtained in the was 97.2%. Nishio et al. [11] used homology-based technique and machine learning methods to classify lung tissue images into three classes. The overall classification accuracy obtained was 99.43%.

Masud et al. [12] classify colon and lung histopathological images using a deep learning-based method. They used domain transformations of two types to extract four feature sets for image classification. Then they combined the features of the two categories to arrive at the final classification results. They have achieved an accuracy of 96.33%. Mangal et al. [13] made a classification of colon and lung cancers based on histopathological images by applying a shallow neural network architecture. They achieve an accuracy of 97% and 96% in classifying lung and colon cancers respectively. Toğaçar [3] performed the classification of colon and lung cancers' histopathological images by training the images with the Darknet-19 model and then obtain the feature sets to which two optimization algorithms were applied to select the inefficient features. Then, the efficient feature sets, that have been created for each of the two optimization algorithms by distinguishing the ineffective features from the rest of the features in the set, were combined and classified using SVM classifier. He have obtained an overall accuracy of 99.69%.

The main objective of this study is to propose a medical diagnostic support system for lung and colon imaging. In other words, it is to set up an automated system that can accurately classify the subtypes of colon and lung cancer from histopathological images using machine learning, and to show that with feature engineering we can find powerful accuracy results.

The main advantage of using machine learning models is that they allow a better interpretability of the classification model since they are based on feature engineering; however, deep learning models are black box networks whose their working is very difficult to understand due to the complex design of the network. Indeed, in the medical and diagnostic field, feature engineering is crucial for doctors because it allows them to know the importance and impact of

each feature on the classification and identification of cancer subtypes, unlike deep learning models which are black box networks.

The article is organized as follows: Section 2 concerns the materials and methods used. Section 3 reports the experimental results. Section 4 provides a discussion of the results obtained. Finally, section 5 brings the conclusion.

# 2 Material and Methods

## 2.1 Lung and Colon Cancer Dataset

Lung and Colon Cancer Histopathological Image Dataset, published in 2020, is known as LC25000 dataset. LC25000 dataset contains 25,000 images of five classes of colon and lung tissues, 5,000 images of each class [14]. Each image is 768 $\times$ 768 pixels in size. The five types are Colon Adenocarcinoma, Benign Colonic tissue, Lung Adenocarcinoma, Benign Lung tissue, and Lung Squamous cell Carcinoma.

The most frequent type of colon cancer is Colon Adenocarcinoma, which accounts for over 95% of all cases of colon cancer. It is produced when an adenoma - a type of polyp - develops within the large intestine and eventually turning into cancer. Lung Adenocarcinoma, a type of cancer cells that represents for around 60% of all cases of lung cancers, usually grows in the glandular cells located in the outer part of the lung and then spreads to the alveoli within the lung. Lung Squamous Cell Carcinoma, which is the second most frequent type of lung cancer, develops in the airways or bronchi of the lungs and represents around 30% of all cases.

Sample of histopathological images of these five classes of colon and lung tissues that are collected from the LC25000 dataset are illustrated in figure 1.

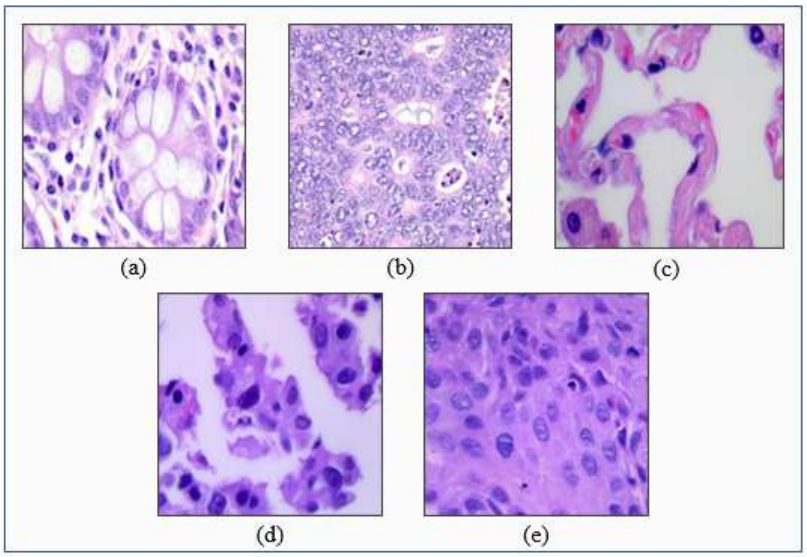


**Fig. 1** Sample images of: (a) Colon Benign tissue, (b) Colon Adenocarcinoma, (c) Lung Benign tissue, (d) Lung Adenocarcinoma, and (e) Lung Squamous cell Carcinoma collected from the LC25000 dataset.

## 2.2 Overview of the methodology

Figure 2 shows an overview of the methodology used for classifying colon and lung cancer subtypes on the basis of histopathological images. The RGB images have been introduced into the system. 2,500 images were used at all stages of our study, 500 images from each class. Images were resized to 200x200 pixels. Each image's contrast was enhanced using Unsharp Masking method, and then the image features were extracted. The Recursive Feature Elimination, which is a feature selection method, is used in order to select the most efficient features. Then, a machine learning algorithm classified the image on the basis of the selected features. 20% of the dataset was used as test data and 80% was devoted to training the data (randomly chosen). The machine learning algorithm is trained using the images features of the training set. Finally, image features of the testing set are used for assessing the performance of the model. The programming language used is Python with the implementation of the following libraries: numpy, pandas, matplotlib, tensorflow, scikit-learn, scikit-image and xgboost.

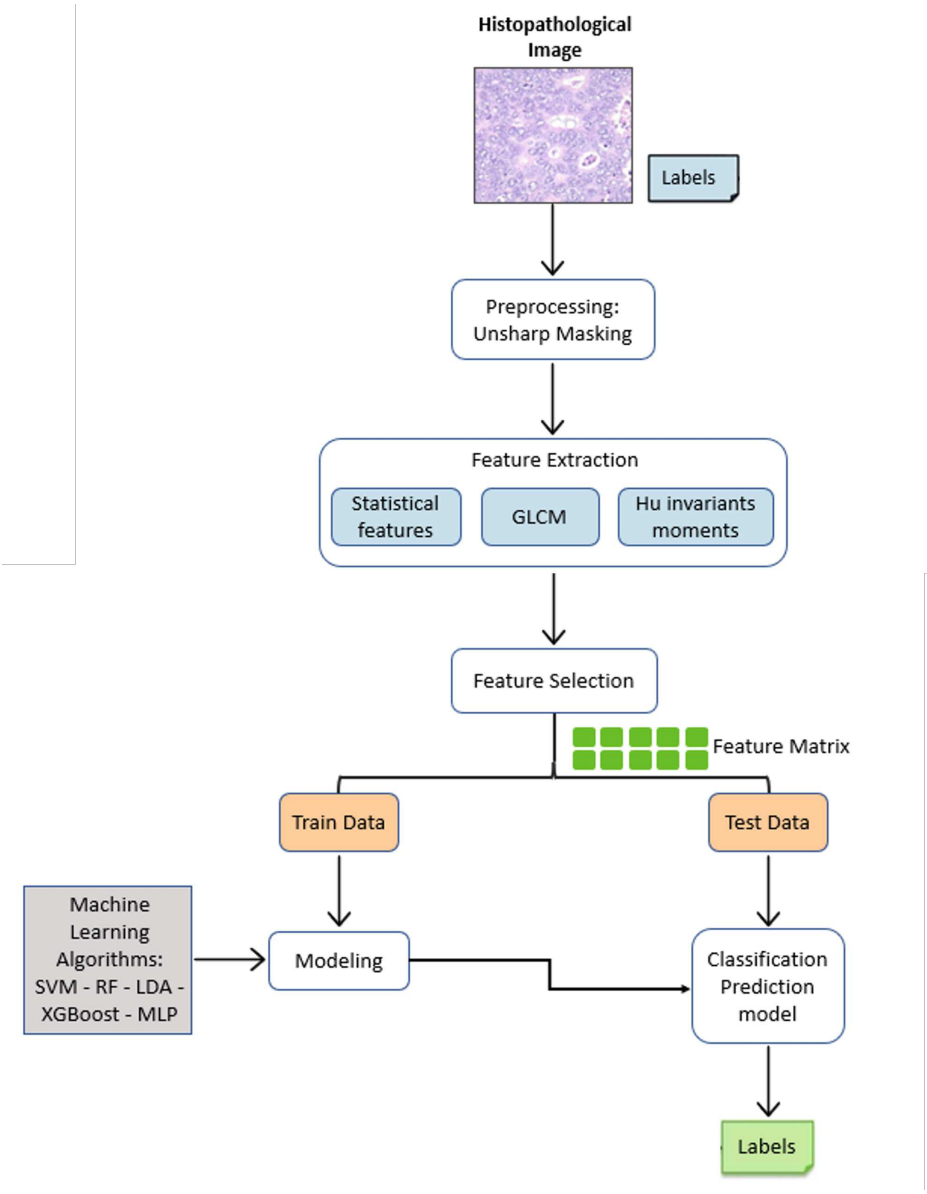


**Fig. 2** Overall flowchart of the methodology used for the classification of cancer subtypes from histopathological images.

## 2.3 Image sharpening

After image acquisition, the images must be pre-processed. Indeed, the preprocessing of the images is essential to improve the quality of the image and

extract important information from the images to make them more adequate for the learning algorithm.

The contrast of each image is enhanced using the Unsharp Masking (UM) which is an image sharpening method. Unsharp Masking enhances the contrast, and thus sharpens the original image, which can help emphasize texture and detail. The basic idea of the UM method is to subtract the original image by a blurred version of the image itself. The typical formula used for unsharp masking is as follows:

$$Sharpened = original + (original - blurred) \times amount \tag{1}$$

The outcome of the Unsharp Masking is conditioned by the radius and amount parameters. The blurring step could use any image filter method, but traditionally a Gaussian filter is used. The radius parameter in the unsharp masking filter refers to the sigma parameter of the Gaussian filter. The radius controls the degree of blurring of the original image, and therefore the dimension of the area encircling the edges that is concerned by the sharpening. The value of the enhancement effect is determined by the amount parameter which is the value of contrast added to the edges.

In our case, in order to choose the best parameters for the unsharp masking method, we carried out a sensitivity study on the parameters: We have tested radius values from 1 to 5, and amount values from 1 to 20. We have obtained that the best values of radius and amount are 2 and 5 respectively, since the models gave the best performance using these values. Therefore, these values are used in the rest of our study. Figure 3 represents the result of enhancing a histopathological image using the Unsharp Masking method under the indicated conditions.

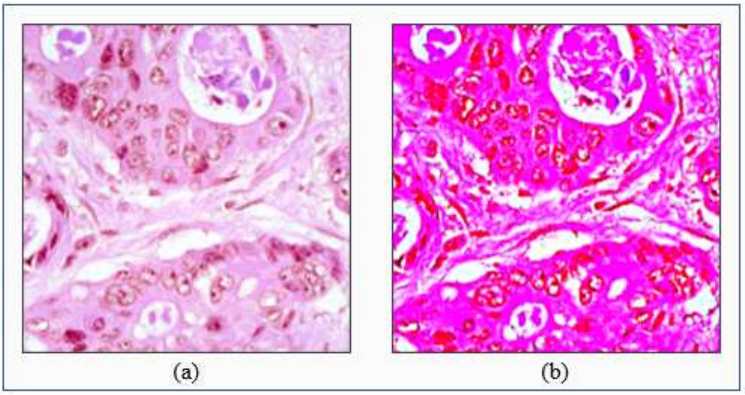


**Fig. 3** A sample of colon cancer image: (a) original image and (b) sharpened image using Unsharp Masking.

## 2.4 Feature extraction

Features are measured values that can be informative for a predictive analysis to classify the attribute. Features contained in histopathological images are essential for the diagnosis of the disease, and efficient features extraction is of high importance to improve the diagnostic accuracy and assist in cancer classification [15]. In this paper, we extracted 37 features, including first order statistics, GLCM and the Hu invariant moments. The computed features for each method are shown in Table 1.

**Table 1** Computed features for each feature extraction method.

| Feature Extraction methods | Computed features |
|---|---|
| First order statistical features | Mean, Standard deviation, Median, Percentile 25%, Percentile 50%, Percentile 75%. |
| GLCM | Contrast1, Contrast2, Contrast3, Contrast4, Dissimilarity1, Dissimilarity2, Dissimilarity3, Dissimilarity4, Homogeneity1, Homogeneity2, Homogeneity3, Homogeneity4, Energy1, Energy2, Energy3, Energy4, asm1, asm2, asm3, asm4, Correlation1, Correlation2, Correlation3, Correlation4. |
| Hu invariant moments | h1, h2, h3, h4, h5, h6, h7. |

### 2.4.1 First order statistics

The features obtained from the first-order statistics provide information about the distribution of brightness in the image. The first-order statistics used are: mean, standard deviation, median, percentile 25%, percentile 50% and percentile 75%.

### 2.4.2 GLCM

In biological imaging, the Gray Level Co-occurrence Matrix (GLCM) is a widely used method for texture analysis due to its ability to capture the spatial dependence of gray level values inside an image since the pixels are considered in pairs. The co-occurrence matrix is a second-order statistical characteristics of the changes in image brightness. It gives a description of the gray level variations between each pixel in the texture of the image and its neighboring pixels. Indeed, it is a tabulation of the frequency of different combinations of pixels brightness values (gray tone) which occur within an image [17].

The co-occurrence matrix is a function of two parameters: the distance (d) that is measured in number of pixels and their orientation ($\theta$). The orientation $\theta$ takes the values 0°, 45°, 90°, and 135°, which represent the four directions the horizontal, diagonal, vertical, and anti-diagonal, respectively.The occurrence of a gray level pattern can be represented by a relative frequency matrix $P_{\theta,d}(I_1, I_2)$ which describes the frequency of appearance of two gray level pixels $I_1, I_2$ in the window that are separated by a distance d in the $\theta$ direction [18].

The computed features are: contrast, dissimilarity, homogeneity, energy, angular second moment (ASM) and correlation, of which a group of four features is calculated for each.

$$Contrast = \sum_{I_1,I_2} \mid I_1 - I_2 \mid^2 P(I_1, I_2) \tag{2}$$

$$Dissimilarity = \sum_{I_1,I_2} P(I_1, I_2) \mid I_1 - I_2 \mid \tag{3}$$

$$Homogeneity = \sum_{I_1,I_2} \frac{P(I_1, I_2)}{1+ \mid I_1 - I_2 \mid^2} \tag{4}$$

$$Energy = \sum_{I_1,I_2} P(I_1, I_2)^2 \tag{5}$$

$$Correlation = \sum_{I_1,I_2} \frac{(I_1 - \mu_1)(I_2 - \mu_2)P(I_1, I_2)}{\sigma_1\sigma_2} \tag{6}$$

The contrast is a feature that measures the local level variations and it takes high values for high contrast images. The dissimilarity provides a measure of the randomness of pixels and takes low values if we have the same pixel pairs. Homogeneity is a measure that takes high values if we have similar pairs of pixels. The ASM is used to measure the smoothness of an image and takes a low value if the region is less smooth. Correlation measures the correlation between pixels in two different directions. Since these features depend on d and $\theta$, then their values differ if the image is returned. Thus we will have features that are invariant to rotation.

### 2.4.3 Hu invariant moments

The moment feature generally describes the geometric characteristics in the image area. Hu invariant moments are a set of seven numbers calculated using central moments that are invariant to image transformations. Due to the invariance to translation, rotation and scaling, Hu invariant moments are largely used in the field of image pattern recognition, classification, and target recognition [16]. Therefore, in this paper, we used the Hu invariant moments to represent the characteristics of histopathological images of colon and lung cancers.

## 2.5 Feature selection

Recursive feature elimination (RFE) is a feature selection method that eliminates the least important features, as well as dependencies and collinearity that may exist in the model, until the desired number of features is reached. RFE is popular because it is easy to implement and it is effective in selecting features from a training data set that are more relevant to predict the target variable.

Features are ranked using the feature_importances_ attributes of the model. RFE requires that a specified number of features be retained, but since the number of valid features is not known in advance, then to find the optimal number of features, cross validation is used with RFE to evaluate several subsets of features and select the best. RFECV performs RFE in a cross-validation loop to find the optimal number of features. The purpose of recursive feature elimination is to select features by considering recursively smaller feature sets. First, the classifier is trained on the initial feature set and the importance of each feature is obtained based on its contribution to the classification. Then, the features were sorted from high to low according to their importance, which results in a feature ranking. Lastly, the features that are least important are eliminated from the actual feature set. And then the updated features are used to re-train the model, and we obtained the classification performance using the new feature set. This process is repeated recursively on the reduced set until the desired number of features to be selected is reached. RFE needs several parameters such as estimator and scoring. A scoring function is a metric to evaluate the performance of the model such as accuracy, f1-weighted, mean squared error; in our study, accuracy is the metric used.

In this study, RFE tells us to keep only 12 of the 37 features. So, the models are trained only on these 12 features. We compared the feature non-selection and RFE method to look at the performance. The analysis performed with the two methods resulted in not using the RFE method and not reducing the feature vector, since the classification system was more efficient with the use of all features.

## 2.6 Classification

The features extracted from the 2500 images were used in all stages of our study. 20% of the dataset was used as test data and 80% was devoted to training the data (randomly chosen).

The features that are extracted from the images were fed into machine learning algorithms. The machine learning algorithms used are: Support Vector Machine (SVM), Random Forest (RF), Extreme gradient boosting (XGBoost), Linear Discriminant Analysis (LDA), and Multilayer Perceptron (MLP). These machine learning algorithms were trained using the image features of the training set.

In this study, the SVM hyperparameters were tested to select those that performed best with the experiment database; The SVM kernels:['linear', 'rbf'] and C: [1, 10, 100, 1000] were tested. We obtained that the best values of the SVM hyperparameters are a linear kernel and a regularization parameter C of 100, which were selected automatically by the model as they performed best, and a one-versus-one multi-class method was used.

The hyperparameters of the RF model were also tested; The n_estimators: [10, 50, 100, 300], and criterion: ['gini', 'entropy'] were tested. The best values of the RF hyperparameters that were selected are 300 trees and a gini criterion.

The default hyperparameters for the XGBoost algorithm are provided by the implementation of xgboost. The tree-based models (gbtree) which is the type of model to run at each iteration, is the general parameter selected for the XGBoost model. The maximum depth of a tree is 6.

Also, the default hyperparameters of the LDA algorithm are provided by the implementation of sklearn.discriminant_analysis. The svd (Singular value decomposition) solver is used since it does not compute the covariance matrix, so it is recommended for data with a large number of features.

The MLP classifier is composed of three hidden layers with 150, 100 and 50 neurons in each layer, using a 'relu' activation function. The Softmax activation function is used in the last layer of the network. The solver for weight optimization used is 'adam'. The other parameters for the MLP model, such as number of epoch value of 300, and minibatch value of 200 were selected.

## 2.7 Performance Evaluation

There are many metrics that are used to evaluate machine learning models. In this paper, the confusion matrix and associated metric parameters, such as: Accuracy, Precision, Recall and F1-score, are used for the measurement.

- Accuracy is a measure of the classifier's ability to accurately predict cases into their correct category. It is the proportion of valid results obtained or correctly classified samples from total samples.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

  Where TP, TN, FP and FN represent the True Positive, True Negative, False Positive and False negative values, respectively. True Positive (TP) indicates real disease, which means that the real value is positive, and it is classified positively i.e., that the person has the disease, and the test is positive. False negative (FN) indicates no disease while it exists, which means the actual value is positive while it is classified negatively, i.e., that the person has the disease, and the test is negative. False positive (PF) indicates a disease when it does not exist, which means that the true value is negative when it is classified positively. True Negative (TN) indicates the absence of the disease, which means that the true value is negative, and it is classified as negative, i.e., that the person is healthy, and the test is negative.

- Precision is defined as the ratio of correctly detected samples (true positives) to samples that have been detected as positive.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

- Recall, also called sensitivity, is the percentage of positive instances of a particular class that are correctly detected. It is defined as the ratio of true positive samples to the total number of positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

- F1-score is defined as the harmonic average of the precision and the recall.

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{10}$$

# 3 Results

This section presents the acquired results of the five classifiers that are being investigated in this paper. Indeed, the models were evaluated on the test data to determine their performance. In this study, in order to design the most efficient feature extraction methods, we made a comparison of the performance of the models with the use of statistical features, GLCM, Hu invariant moments and combinations between these different methods. Table 2 presents the accuracy results for the different models and groups of characteristics, for the same training and test database.

**Table 2** Comparison of the accuracy results for the different models and groups of characteristics.

| | XGBoost | SVM | RF | LDA | MLP |
|---|---|---|---|---|---|
| Statistical features | 87.8% | 83.6% | 87.2% | 77.2% | 87.4% |
| GLCM | 86.8% | 90% | 86.8% | 87.4% | 90.4% |
| Hu invariant moments | 65.8% | 62% | 69.2% | 62% | 60.2% |
| Statistical features + GLCM | 94.8% | 91.6% | 94.2% | 89% | 93.4% |
| Statistical features + Hu invariant moments | 91.6% | 87% | 90.2% | 82.2% | 85.6% |
| GLCM + Hu invariant moments | 90.2% | 89% | 88% | 85.8% | 84.4% |
| Statistical features + GLCM + Hu invariant moments | 95.6% | 95% | 94.6% | 91% | 92.2% |

An accuracy of 95.6%, 95%, 94.6%, 91% and 92.2% and F1-score of 96%, 95%, 95%, 91% and 92% were obtained respectively with the classifiers XGBoost, SVM, RF, LDA and MLP on the test data, using a combination of the three feature extraction methods. The performance of the classification models on the same test data is shown in Table 3. The acquired results show that the machine learning models perform satisfactorily and are highly accurate in identifying lung and colon cancer subtypes. The XGBoost model performed best with an accuracy of 95.6% and a F1 score of 96%. The confusion matrix for each technique and for the same dataset is shown in Figure 4. The confusion matrix represents the true label versus the predicted label of the images for the test data in given labeled categories. Table 4 shows the precision, recall and f1-score of the XGBoost model for the different categories of histopathological images on the test data.

To be able to compare our model with existing models in the literature, we did the same work with the same steps but using the 25,000 images of colon and lung cancer from the LC25000 database with 70% for training and 30% for testing. Table 5 presents a comparison of the achieved results of the classification of colon and lung cancer subtypes with other methods using the same dataset. Table 6 and Table 7 present a comparison of the achieved results for colon cancer classification and lung cancer classification respectively with other methods using the same dataset and using 90% for training and 10% for testing. Figure 5 illustrates the confusion matrix of the XGBoost model for the three types of classification. Precision, recall and f1-score of the XGBoost model for the different classes of colon and lung cancer with 70% for training and 30% for testing are shown in Table 8. Table 9 and Table 10 present the precision, recall and f1-score of the XGBoost model for the different classes of colon cancer and lung cancer respectively with 10% for testing.

**Table 3** Precision, Recall, F1 score and overall Accuracy of classification models on the same dataset of 2,500 images.

| Classifier | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| XGBoost | 95.6% | 95.8% | 96% | 95.9% |
| SVM | 95% | 95% | 95.2% | 95.1% |
| RF | 94.6% | 94.8% | 95% | 94.9% |
| LDA | 91% | 91.2% | 91% | 91% |
| MLP | 92.2% | 92.6% | 92.4% | 92.5% |

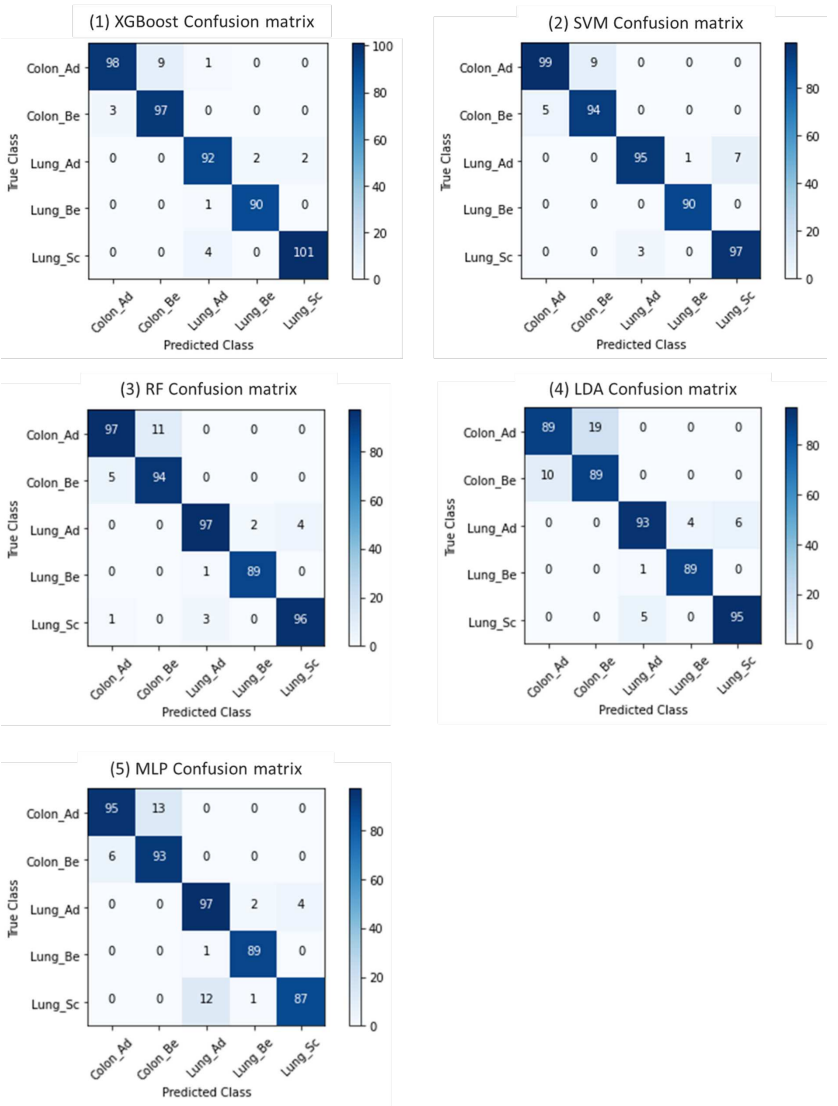


**Fig. 4** Confusion matrix of Colon and Lung cancer classification with different models: (1) XGBoost, (2) SVM, (3) RF, (4) LDA, and (5) MLP.

**Table 4** Precision, Recall, F1 score and overall Accuracy of the XGBoost model for the different classes of colon and lung cancer using 2,500 images with 20% for testing.

| Class | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Colon_Ad | 95±2% | 92±3% | 93±2% | |
| Colon_Be | 93±3% | 96±2% | 95±2% | |
| Lung_Ad | 91±5% | 90±4% | 90±4% | 93.8±2% |
| Lung_Be | 97±2% | 98±1% | 97±1% | |
| Lung_Sc | 93±4% | 93±3% | 93±3% | |

**Table 5** Comparison of the achieved results with other methods using the same dataset of colon and lung cancer.

| Reference | Cancer Type | Classifier | Test Rate | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|
| [3] | Lung and Colon | DarkNet-19 + SVM | 30% | 99.69% | - | - | - |
| [12] | Lung and Colon | CNN | 30% | 96.33% | 96.39% | 96.37% | 96.38% |
| Proposed model | Lung and Colon | XGBoost | 30% | 99% | 98.6% | 99% | 98.8% |

**Table 6** Comparison of the achieved results with other methods using the same dataset of colon cancer.

| Reference | Cancer Type | Classifier | Test Rate | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|
| [13] | Colon | CNN | 10% | 96.61% | - | - | - |
| Proposed model | Colon | XGBoost | 10% | 99.3% | 99.5% | 99.5% | 99.5% |

**Table 7** Comparison of the achieved results with other methods using the same dataset of lung cancer.

| Reference | Cancer Type | Classifier | Test Rate | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|
| [10] | Lung | CNN | 10% | 97.2% | 97.33% | 97.33% | 97.33% |
| [13] | Lung | CNN | 10% | 97.89% | - | - | - |
| Proposed model | Lung | XGBoost | 10% | 99.53% | 99.33% | 99.33% | 99.33% |

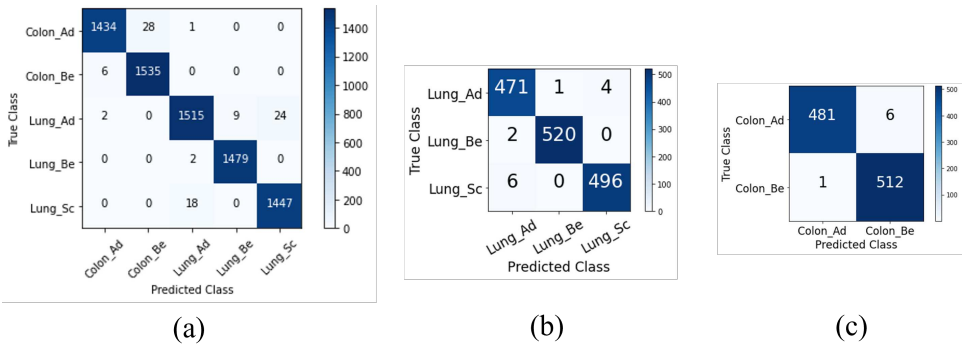


**Fig. 5** Confusion matrix of: (a) Colon and Lung cancer classification, (b) Lung cancer classification, and (c) Colon cancer classification, using all the images of the LC25000 dataset and with the XGBoost model.

**Table 8** Precision, Recall, F1 score and overall Accuracy of the XGBoost model for the different classes of colon and lung cancer using 25,000 images with 30% for testing.

| Class | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Colon_Ad | 99% | 98% | 99% | |
| Colon_Be | 98% | 100% | 99% | |
| Lung_Ad | 99% | 98% | 99% | 99% |
| Lung_Be | 99% | 100% | 100% | |
| Lung_Sc | 98% | 99% | 99% | |

**Table 9** Precision, Recall, F1 score and overall Accuracy of the XGBoost model for colon cancer with 10% for testing.

| Class | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Colon_Ad | 100% | 99% | 99% | 99% |
| Colon_Be | 99% | 100% | 99% | |

**Table 10** Precision, Recall, F1 score and overall Accuracy of the XGBoost model for the different classes of lung cancer with 10% for testing.

| Class | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Lung_Ad | 99% | 99% | 99% | |
| Lung_Be | 100% | 100% | 100% | 100% |
| Lung_Sc | 99% | 99% | 99% | |

# 4 Discussion

The 2,500 RGB images including 500 images of each class were fed into the system. The unsharp masking method was used to sharpen the images. A sensitivity study on the parameters resulted in the best values of the radius and amount parameters which are 2 and 5 respectively and which are used in the rest of our study. After that, three texture extraction methods including first-order statistical features, GLCM and invariant Hu moments were used for feature extraction. In order to design the most effective feature extraction methods, we compared the performance of the models with the use of these different methods and their combinations.

According to Table 2, the analysis performed with the different methods resulted in using a combination of the three feature extraction methods: statistical features, GLCM and Hu invariant moments, since the most efficient model is obtained with these three combined feature extraction methods. Indeed, we notice that by calculating the statistical characteristics with the XGBoost model, we have 87.8% of classification accuracy. By calculating the GLCMs, we have 86.8% classification accuracy. By combining these two groups of features, we obtain 94.8% of accuracy. And by adding the features of Hu invariant moments, we get 95.6% of accuracy classification. Hence the interest of using the three combined feature extraction methods. Therefore, a concatenation of the feature vectors extracted from these three methods resulted in the combined feature set with 37 features, which are the samples of the dataset in the training and classification steps.

The features extracted from the images were fed into the machine learning algorithms. 80% of the features (randomly chosen) are used to train the machine learning algorithm and the rest 20% are used as test data to evaluate the system performance. As shown in Table 3, the XGBoost model has the best accuracy of 95.6% and an F1-score of 96%. From Table 4, 8, 9 and 10, we can see that the XGBoost model works well in identifying different classes of colon and lung cancer subtypes. As shown in the confusion matrix in Figure 4, only 22 samples out of 500 images have been incorrectly classified with the XGBoost classifier. The Lun_Be class achieved the greatest classification result, while the Col_Ad class got the highest misclassification result.

Overall, it can be said that the ML models, especially the XGBoost model, followed by the SVM and RF models, are very accurate in identifying classes of lung and colon cancer subtypes, although there is still room for improvement. Therefore, the obtained results show that machine learning models can be used to classify histopathological images of colon and lung cancers with high reliability and precision.

In most recently published research articles, the authors have used deep learning to classify colon and lung cancers' histopathological images. Tables 5, 6 and 7 present a comparison of our results using the 25,000 images of the LC25000 dataset with those of articles in the literature. Indeed, these previous studies used deep learning, while our study used machine learning. Our study has proved that with feature engineering we can find results that are

competitive with Deep Learning approaches. XGBoost achieved an accuracy of 99% for the classification of colon and lung cancer subtypes, 99.3% for the classification of colon cancer, and 99.53% for the classification of lung cancer subtypes. We notice that the model for each type of organs are more efficient. In fact, our objective is not to compete with existing models, but to show the interest of machine learning and feature engineering models and to show that it is possible to find better results using models of machine learning.

The main advantage of machine learning models over those of deep learning is that in machine learning models, we have a view on the models since they are based on feature engineering which is the process of transforming raw data into useful features which help us to better understand our model and to increase its predictive power and thus allows a better interpretability of the classification model. Indeed, in the medical and diagnostic field, feature engineering is crucial for doctors to make life changing decisions because it allows them to know the importance and impact of each feature on the classification of cancer subtypes; unlike deep learning models which are black box networks whose their working is very difficult to understand and interpret because of complex network design [19]. Indeed, deep learning models take automatic decisions on its own without us being able to interpret what is going on inside the model, during the analysis at each level of the neural network.

Also, there are many more parameters and hyperparameters that can be learned in deep models than in machine learning models, and so a deep learning system can take a long time to train; It can take from a few hours to a few weeks! While feature engineering-based ML takes comparatively much less time to train, ranging from a few seconds to a few hours [19].

Additionally, machine learning algorithms are less complex than deep learning algorithms and can often run on conventional computers, while deep learning systems require much more powerful hardware and resources with very high performance due to the amount of data processed and the complexity of the mathematical calculations involved in the algorithms used. This need for power has led to increased use of graphics processing units (GPU) which are very expensive.

## 4.1 Model Explainability with SHAP

The SHAP method is used to explain the output of a machine learning model by computing the contribution of each feature to the prediction. Therefore, it allows to evaluate how the contribution of each feature affects the model [20].

The importance of SHAP features is calculated as the average of the absolute Shapley values. The idea of SHAP feature importance is that important features are those with great absolute Shapley values. Figure 6 illustrates the most important features that are selected and ordered according to their importance using the SHAP method for the previously trained random forest model for colon cancer prediction. The first order statistical features are the most relevant, followed by second order features such as correlation. Percentile

50% was the most relevant feature, which modified the absolute probability of predicted cancer by an average of 8 percentage points. Thus, on the medical side, specialists and doctors can interpret the variables and know which features are more important in identifying and classifying cancer subtypes.

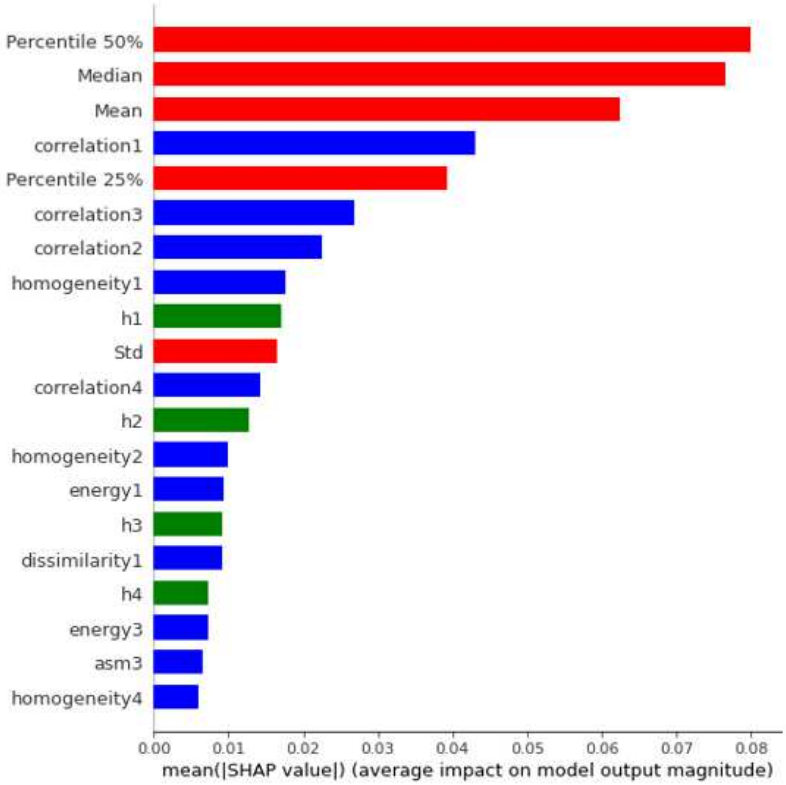


**Fig. 6** SHAP feature importance measured as the mean of the absolute Shapley values. Colors represent the groups of features. Percentile 50% is the most essential feature, modifying the absolute probability of predicted cancer by an average of 8 percentage points (0.08 on x-axis).

The SHAP Summary Plot shown in Figure 7 combines the importance of features with their effects. Each point on the graph represents a Shapley value for a feature and an instance. The x axis position is determined by the Shapley value. The horizontal location shows whether the effect of that value caused a higher or lower prediction. The features on the y-axis are ordered according to their importance. Each line (y-axis) on the graph points to the feature on the left and is colored according to the value of the feature - high values for that feature are red, and low values for that feature are blue. Values to the right have a ”positive” impact on the output, and the values to the left have a ”negative” impact on the output. Note that positive and negative refer to the

direction in which the output of the model is impacted, it has no guidance on the performance.

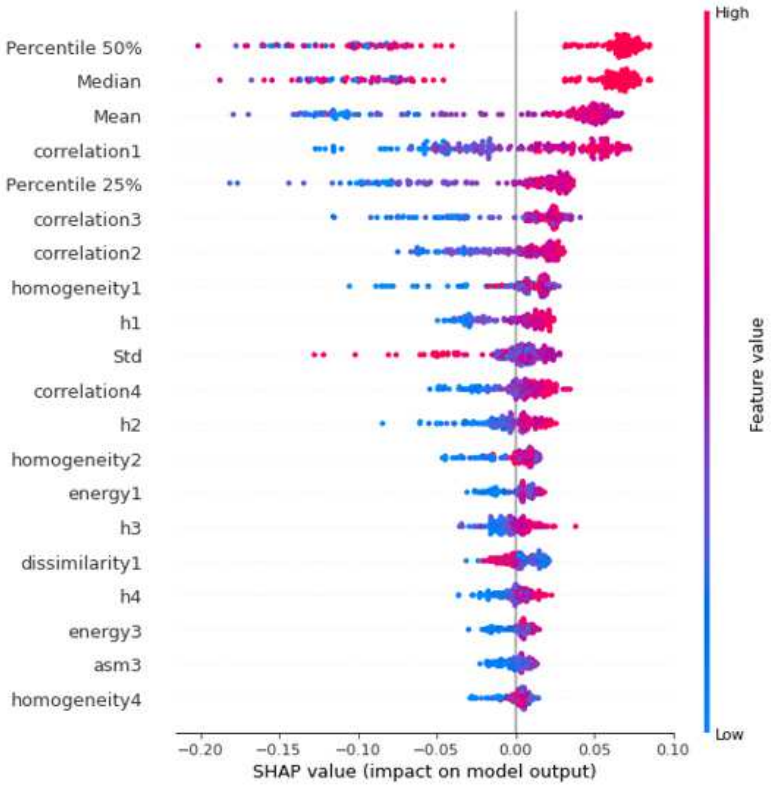


**Fig. 7** SHAP summary plot. High values for the feature are red, and low values for that feature are blue. Values to the right have a positive impact on the output, and the values to the left have a negative impact.

This plot allows us to visualize the impact of the feature, as well as how the impact of the feature varies with lower or higher values. For example, a large value of mean increases the risk of predicted colon cancer and a small value reduces the risk. The features presented such as mean, correlation1 and Percentile 25% have negative impacts for low values and positive impacts for high values. So medically, specialists can understand the variables and know how the values of each characteristic impact the identification of colon cancer subtypes.

Figure 8 presents the SHAP force_plot output for two patients from the colon cancer dataset. The prediction begins from the base value. The base value for the Shapley values is the mean of all predictions. Then the prediction is modified accordingly based on the value of each feature. Feature values that

increase predictions are in red, and their visual size shows the magnitude of the feature effect. Feature values that decrease predictions are in blue.

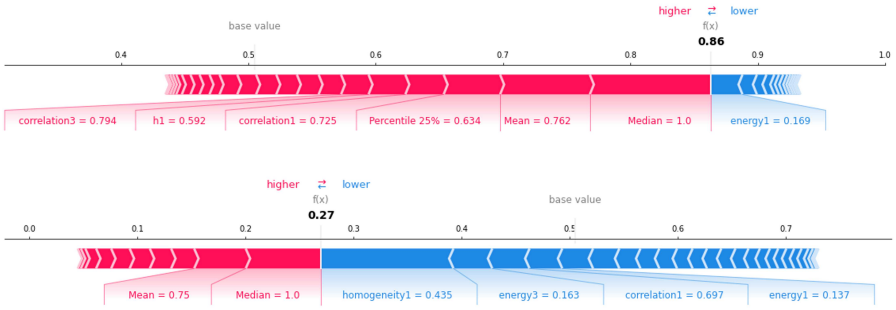


**Fig. 8** SHAP force_plot to provide an explanation of the predicted colon cancer probabilities for two patients. Each feature value is a force that decreases or increases the prediction.

The first patient has a high risk prediction of 0.86 of having colon cancer. Median, Mean, Percentile 25% increase his predicted risk of cancer. The greatest impact comes from the median feature. Although energy1 has a significant effect on decreasing the prediction. The second patient has a low predicted risk of 0.27. Features that increase risk are compensated by features that decrease risk such as homogeneity1.

Thus, having a justification for the prediction of a model would give specialists confidence regarding the validity of the model's decision. Indeed, in the medical field, decision-making processes must be transparent, and then it is important to explain the model predictions in order to support the specialists' decision-making processes.

# 5 Conclusion

In this paper, we presented machine learning models that are based on feature engineering for the classification of histopathological images of colon and lung cancers into five classes (three malignant and two benign).

We preprocessed the dataset using an image enhancement method known as unsharp masking. Three feature sets were extracted for the classification of images. The resulting features were then concatenated to create a combined feature set that was fed into the machine learning algorithms. The XGBoost model has the best classification performance in terms of accuracy, precision, and recall for distinguishing lung and colon cancer subtypes. XGBoost achieved an accuracy of 99% and an f1 score of 98.8%. SHAP method is used to provide an explanation of the output of a machine learning model and to show the contribution of each feature on the model. Using this method,

specialists can then understand which features of the histopathological image contributed to its classification as cancer. Unlike previous papers where the authors used deep learning which is a black box network that is very difficult to interpret and in the medical field specialists cannot understand what is happening inside the model.

Thus, the use of computer programs that are based on machine learning and feature engineering to analyze data and extract important information could be a very useful and crucial tool in the medical field for the immediate and accurate diagnosis of malignant tumors. Indeed, these programs will be able to provide significant help to specialists to better interpret features and know the importance and impact of each on the identification of colon and lung cancer subtypes. In the future, it is planned to explore other feature extraction techniques that provide relevant features for identifying colon and lung cancer subtypes from histopathological sections to improve model performance. It is also planned to evaluate the performance of our proposed approach on other histopathological images of colon and lung cancer to evaluate its efficacy.

## Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

## Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

## References

[1] Bray F., Ferlay J., Soerjomataram I., Siegel R.L., Torre L.A., Jemal A., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: a cancer journal for clinicians, 2018. 68, 6.394–424.

[2] Bermúdez A., Arranz-Salas I., Mercado S., López-Villodres J.A., González V., Ríus F., Ortega M.V., Alba C., Hierro I., Bermúdez D., Her2-Positive and Microsatellite Instability Status in Gastric Cancer—Clinicopathological Implications, Diagnostics 2021, 11, 944.

[3] Mesut Togaçar, Disease type detection in lung and colon cancer images using the complement approach of inefficient sets, Computers in Biology and Medicine, Volume 137, 2021, 104827, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104827.

[4] Sánchez-Peralta L.F., Bote-Curiel L., Picón A., Sánchez-Margallo F.M., Pagador J.B., Deep learning to find colorectal polyps in

colonoscopy: A systematic literature review, Artificial Intelligence in Medicine, Volume 108, 2020, 101923, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2020.101923.

[5] Travis W.D. et al., International association for the study of lung cancer/American thoracic society/European respiratory society international multidisciplinary classification of lung adenocarcinoma, Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer vol. 6, 2011, 244-85, doi: 10.1097/JTO.0b013e318206a221.

[6] Yu K.H., Zhang C., Berry G. et al., Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features, Nat Commun 7, 12474, 2016, https://doi.org/10.1038/ncomms12474.

[7] D. Bazazeh and R. Shubair, Comparative study of machine learning algorithms for breast cancer detection and diagnosis, 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016, pp. 1-4, doi: 10.1109/ICEDSA.2016.7818560.

[8] Schmidhuber J., Deep Learning in neural networks: An overview. Neural Network, 2015, 61, 85–117.

[9] Bukhari S.U.K., Asmara S., Bokhari S.K.A., Hussain S.S., Armaghan S.U., SHAH S.S.H., The Histological Diagnosis of Colonic Adenocarcinoma by Applying Partial Self Supervised Learning, medRxiv 2020, https://doi.org/10.1101/2020.08.15.20175760.

[10] Hatuwal B.K., Thapa H.C., Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images, International Journal of Computer Trends and Technology, Volume 68 Issue 10, 21-24, October 2020, doi: 10.14445/22312803/IJCTT-V68I10P104.

[11] M. Nishio, M. Nishio, N. Jimbo, K. Nakane, Homology-Based Image Processing for Automatic Classification of Histopathological Images of Lung Tissue, Cancers 2021, 13, 1192, doi: 10.3390/cancers13061192.

[12] Masud M., Sikder N., Nahid A.A., Bairagi A.K., AlZain M.A., A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework, Sensors 2021, 21, 748, https://doi.org/10.3390/s21030748.

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

[14] Borkowski A.A., Bui M.M., Thomas L.B., Wilson C.P., DeLand L.A., Mastorides S.M., Lung and Colon Cancer Histopathological Image Dataset (LC25000), arXiv:1912.12142v1 [eess.IV], 2019, https://arxiv.org/abs/1912.12142v1.

[15] S. Alinsaif and J. Lang, Texture features in the shearlet domain for histopathological image classification, BMC Med. Informat. Decis. Making, vol. 20, no. S14, pp. 1–19, Dec. 2020.

[16] M. Li et al., Research on the Auxiliary Classification and Diagnosis of Lung Cancer Subtypes Based on Histopathological Images, in IEEE Access, vol. 9, pp. 53687-53707, 2021.

[17] Madero Orozco H. et al., An Automated systems for lungs nodule classifications based on wavelet feature descriptors and support-vector-machines, Biomedical Engineering Online, vol. 14, no. 1, p. 9, 2015.

[18] Aggarwal N. and AgrawalR. K., First and Second Order Statistics Features for Classification of Magnetic Resonance Brain Images, Journal of Signal and Information Processing, Vol. 3 No. 2, 2012, pp. 146-153, doi: 10.4236/jsip.2012.32019.

[19] Dargan S., Kumar M., Ayyagari M.R. et al., A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning, Arch Computat Methods Eng 27, 1071–1092 (2020), https://doi.org/10.1007/s11831-019-09344-w.

[20] Christoph Molnar, Interpretable machine learning. A Guide for Making Black Box Models Explainable, 2019. https://christophm.github.io/interpretable-ml-book/.