Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 7
Open Access

A Novel Machine Learning Technique to Detect Lung Cancer in CT Images using Auto Color Correlogram Features and Multiple Machine Learning Classifiers

[1]Geetha K, [2]Dr.Karthikeyan Elangovan

[1]Research Scholar (Part time Internal), Department of Computer and Information Science, Annamalai University, Chidambaram, India.

[2]Research Supervisor, Assistant Professor/Programmer, (Deputed as Assistant Professor and Head in Government Arts and Science College, Gingee, Villupuram), Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalai Nagar, Tamil Nadu, India.

Cite this paper as: Geetha K, Dr.Karthikeyan Elangovan (2024). A Novel Machine Learning Technique to Detect Lung Cancer in CT Images using Auto Color Correlogram Features and Multiple Machine Learning Classifiers. *Frontiers in Health Informatics*, *13 (7) 423-438*

Abstract: This study presents an analysis of the performance that different machine learning algorithms, with SCHF as the feature extraction method, yield in detecting lung cancer. Categorized models were Naive Bayes Multinomial, Logistic Regression, Additive Regression, Linear Regression, Attribute Selected Classifier, and moreover, Naive Bayes The measured performance features were overall accuracy, precision, recall; F-measure, Cohen's Kappa, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), and finally, Root Relative Squared Error (RRSE). Hence, the analysis reveals that Naive Bayes Multinomial model had the highest accuracy (90.62 %) and these performances were significantly better in precision (0.91), recall (0.91) and F-measure (0.78) and lowest error rate among most of the models. The significance of these results is that the recommended SCHF framework is exceptionally suitable for feature extraction while Naive Bayes Multinomial gives the most accurate lung cancer classification. This research highlights how machine learning, particularly higher order approaches, can enhance early detection and, outcomes classification of Lung Cancer to aid clinical management.

Key terms: Lung cancer, SCHF, Machine Learning, Image Histogram, CT scan

I Introduction

Lung cancer is one of the leading types of cancer and one of the main causes of cancer mortality rates per annum. He noted it develops from uncontrolled growth of cells within the lungs and impact normal functioning of the body by forming tumors in the respiratory area. Seminual screening for lung cancer is very important as the chances of survival for long just increase dramatically with early discovery of lung cancer.

Contemporary methods used in diagnosing lung cancer takes advantages of newer technology in imaging as well as artificial intelligence (AI). Common techniques employed in the diagnosis of structural changes of lung tissues include X-ray examination and CT-scan. However, these techniques depend with image interpretations made by radiologists, resulting to spatial and temporal variability in the detection of the cancer.

Due to advancement in technologies in and image processing, automatic assistance in the diagnosis of lung cancer can be helped by ML. Such systems explain the medical images, feature selection that is important for further decision making and classify diseases according to certain patterns. In this work, the simplest form of information called color histograms and texture analysis, as well as Deep learning models, are applied to identify and differentiate between different forms of lung cancer, namely adenocarcinoma, squamous cell carcinoma, and large cell carcinoma.

Such systems use effective algorithms, moreover, they include the pre-processing techniques and an effective method

for sampling. Some of the important steps include collection of a large dataset, unearthing of image features, using sampling technique to address the imbalance data issue, and using machine learning technique to develop predictive models. Each of the above-discussed systems takes a long time to develop, and once developed, they are tested and verified against clinical norms.

Lung cancer can be diagnosed in such ways and it has many benefits compared to traditional methods such as lack of potential errors, increased rates of early-stage diagnoses, and enhancing the performance of radiologists. Through the use of these related advancements, health care providers will be in a position to improve the patients' experience and support the fight against lung cancer around the world.

This paper organizes section 2 focuses on literature survey; in section 3 presents materials and methods; in section 4 shows results and interpretations, and finally section 5 has conclusion of this research work.

II Literature Survey

This paper focuses on the automation of lung cancer diagnosis and classification using deep learning methods: Deep Convolutional Neural Network (DCNN) methods. It provides an overview of the methodologies used, advancements made, quality assessments, and other bespoke deep learning frameworks for a host of MI techniques and modalities, including MRI, CT, WSI, and X-ray. As highlights in the study, DCNN in diagnosing and discovering lung cancer classification.[1] The deployment of machine learning methodologies in identifying and predicting lung cancer from medical image data was the emphasis of this review study. It considers a number of proposed systems to assess the other classifiers and image processing techniques used for the reliable differentiation of between benign and malignant lung tumours [2]. The system generates good pictures that are far from error-prone and capable of detecting cancer without misclassifications. There is a variety of classifiers employed to filter false-positive nodules. [3]. Prominent characteristics are computed in training images and are part of the inherent database of the system. Nonetheless, it is less costly than conventional CAD software programs for deep learning. The radiologist gets to enjoy a quicker detection and identification associated with deep learning HD on the input data. In this case, as pixels are used in demarcating between cancer and the normal tissues, the image as is relieves to identity cancer right from the pixel level. As such, doctors may be able to provide more assistance to the healthcare system through the deep learning support system used to diagnose illness and categorize diseases. It makes the task of arriving at more informed decisions concerning the sickness easier. CLs is one of the phases that can be identified in the CNN architecture that has been defined in the previous work.[4] WBAN devices and CC technology are used in development of S-CI which ensures patient's privacy and assist the healthcare segment by supporting real time monitoring of the patient and early discovery of the diseases [5]. The potential application of deep machine learning includes pre-processing of images, to silhouette particular characteristics of images to enhance the diagnosis, and the ability to systematically classify objects as benign or malignant [6]. This study on (Potential Malignant Lung Nodule Detection) used Convolutional Neural Network (CNN) Deep Learning that examined multiple CT scan picture feed as the input into the model to reach the conclusion that it beamed. In this work, the lung nodule detection problem has been addressed through using an ensemble method. In order to enhance the performance of the model as well as the accuracy of the predicted results, we decided to use ensemble method whereby, instead of using one CNN Deep Learning model, we created a pool of at least two CNN Deep Learning models. They provide for free the LUNA 16 Grand Challenge dataset on their website. Metadata has been provided to interpret the details and information of each image and the dataset includes one CT scan data. Machine learning is divided into two groups, classical and deep learning, the latter which uses artificial neural networks which mimic those of the brain. In the case of the deep learning model, the model is learned via a large set of CT scans. CNNs are trained from data collection to classify pictures that depict cancer and those that do not. In Deep Ensemble 2D CNN architecture we make training dataset, validation dataset, and testing dataset. The Deep Ensemble 2D CNN comprises of the three different CNNs with different pooling techniques, layers and kernels. Our proposed Deep Ensemble 2D CNN attained higher accuracy of 95%

which is significantly better than the baseline approach presented in the study.[7]
To a large extent, both the previous CAD apps as well as the current study aim at developing a model for distinguishing lung nodules for classifying lung cancer. Therefore, we focused on identifying current and most effective techniques in differentiating benign from malignant lung nodules. CNN with transfer learning method was developed using Multiresolution CNN [15] and Knowledge Transfer for Candidate Classification in Lung Nodule Detection to capture and training the features from the picture. Multiview-KBC[8], a deep residual learning method that leveraging CT scan data for cancer detection, is derived from the Knowledge-based Collaborative Deep Learning for Benign-Malignant Lung Nodule Classification on Chest[9].
The study introduced a Lung Cell Cancer Detection (LCCD) technique based on DL that accurately measures and categorises malignant cells in lung tissue. By employing a hybrid CNN model and applying digital image processing methodologies, the device can effectively and efficiently diagnose cancer from the pictures obtained from the CT scans.[10] In contrast to conventional approaches for reading the data from the disorganised (raw) form, techniques have been designed for learning representation from the raw data using a deep learning algorithm. Some of the significant information is extracted from the data by observing inside body features. Thus, methods, models, and techniques of deep learning have been reported to enhance classification accuracy in cases of lung cancer, and at the same time reduce the error. For the following reasons, automatic segmentation using deep learning is better than manual segmentation [11]. It SUMS UP that the activity of the radiologist being able to diagnose the problem in the shortest time possible directly depends on the quality and accuracy of the pictures. Precisely, deep learning algorithms have been applied to diagnose lung cancer. [12].
A CNN has many levels; it is not very simple but it is quite effective for use in neural networks. The convolutional layer selects and extracts the feature from the picture pooling layer. Joining all these collected attributes is done by the third layer often referred to as the fully connected or FC layer. RNNs are good with sequential input and the input forms in them include text, audio and video and also consists of video and audio which is majorly of Self-Capture input form. An RBM is a part of many that make up a Deep Belief Network (DBN). There kinds of models are probabilistic in nature. As is seen, DBN can have various types of structures. The Statistical theory based methods include the Support Vector Machine (SVM) which is featured below. ANNs are described as biologically inspired networks simply because of their structure that resembles the neurones in human brains. Other interactions can also be solved where they are nonlinear and can be solved by the Deep Neural Network (DNN) which is also recently developed into an artificial intelligence technique. Asthma, cancer and AIDS are among the human diseases that are associated with DNA binding proteins [8]. A good example with deep learning model could help in avoiding time wastage and wrong diagnosis [13,29]. Although medication is essential in view of the fact that human diseases are challenging to predict, especially cancers, timely and efficient medication must be provided. Many bodily organs and structures in the human body are affected by the deadly disease called cancer [14,28].
Since it might be tricky to differentiate actual lung tissue from a lung nodule, an ensemble technique has been devised for the detection of lung nodules. This will enable creating a more accurate approach to distinguish between a lung nodule and a lung nodule candidate. The biggest problem for all researchers is not the quantity of image data but rather the availability of relevant annotation and tagged picture data. All casework reports of radiologists FIGURE 1 based on findings and comments in free text are generated through PACS. However, it is necessary to just know and master these approaches to text-mining. Today, text mining is another well-known use of deep learning. Hence, in a bid to accouple the goals of deep learning and machine learning, its more advisable to develop a well-fashioned reporting system. This means that radiologists could manage a lot of work from several doctors via the patient care computerised aid system with possible enhancement of the radiologic outcome. Two types of studies are incorporated into lung nodule analysis: Nodule Candidates and True Nodules: both the actual nodules and nodules that bear close resemblance to actual nodules are referred to as lung nodule prospects. To select true nodules from all potential candidate nodules at all potential sites,

a categorisation system must therefore be developed. Before knowing which nodules are real, two features should be calculated more: Regular lung nodules in the CT image have been detected using a two-dimensional CNN. In 2D CNN two dimensions are taken into account. In order to solve classification problems in particular, some works have used ensemble learning and deep learning schemes [15].

For CNN image-wise computation multiple depth layers were employed to Luna lung nodule classification 16 Data Set and thus improving the lung nodule detection rate to 0.9733. The multi-view convolutional network CAD system was developed by the authors in [16,25] to minimize false positives for lung nodules. This technique involved the use of ResNet14 and UNet, which we will use to capture features. Random forest and XG boost are the classifier models applied on the identification of the hazardous images. When tested in this model, the accuracy that was obtained was 84%.[19,24] It is recommended to employ the machine learning method and the ensemble learning approach to prognosis lung cancer from initial signs. To categorize lung cancer, the present research employed neural networks, SVMs [21], MLPs [20,26] as well as Naïve Bayes [22]. The required dataset used in this investigation was obtained from the UCI repository. In the proposed investigation for the ensemble learning strategy with the accuracy of 90% [23,27,30].

III Materials and Methods

Particularly, this segment focuses on the materials used, as well as the strategies applied in the study endeavor. The Chest CT-Scan images were collected from the Kaggle data repository section [24]. The research process consisted of employing CNN in analysing and diagnosing chest cancer employing the deep learning and the machine learning algorithms. Make use of a classification AI model to categorize and, in effect, determine if the patient has cancer. Instruct them on information of the type of cancer they are facing and the treatments to expect. In order we tried to gather all the information that could be necessary for the model to correctly classify the images. Therefore, to start this investigation, this study had to gather information from various sources. To present material gathered from several sources, this study conducted a thorough analysis to compile the information needed for CNN.

Dataset Descriptions

According to the model, the images should be in jpg,png format not in dcm format though they contain images. The data include three types of chest cancer: Adenocarcinoma, big cell carcinoma and squamous cell carcinoma are the main types that we'll be learning about, and there's also a folder of normal cells. The primary directory is designated as "Data," encompassing all subordinate directories: "In the data split, 'test' is the testing set, 'train' means training set, and 'valid test represent testing set' and 'valid represent validation set,' With the distribution of training data at 70%, testing data at 20%, and validation data at 10%.

Adenocarcinoma

This kind of lung cancer accounts for about 40% of non small lung cancer and 30% of total lung cancer. Lung SCC is the most common variety of lung cancer It is the form of lung cancer with the highest prevalence. Adenocarcinomas can be staged in the prostate, breast, and colon and rectal cancers among others. Lung adenocarcinomas develop in the mucous glands found in the respiratory zone in the peripheral area of the lung. These theories include; coughing, hoarseness, weight loss and weakness.

Large cell carcinoma

Large cell undifferentiated carcinoma may develop in any part of the lung due to increased population doubling rate and ability to metastasize. Consequently, such kind of NSCLC accounts for ten to fifteen percent of all its incidences. The disease behaviour of undifferentiated large-cell carcinoma is characterised by early propensity for growth and spreading.

Squamous cell carcinoma

It is a lung cancer that develops at the primary bronchus or at the central region in the lung where the bronchi connect to the trachea. About a third of all non- small cell lung cancers fall under the category of squamous cell lung cancer and are strongly associated with smoking. Once in the final folder are the normal CT scan images.

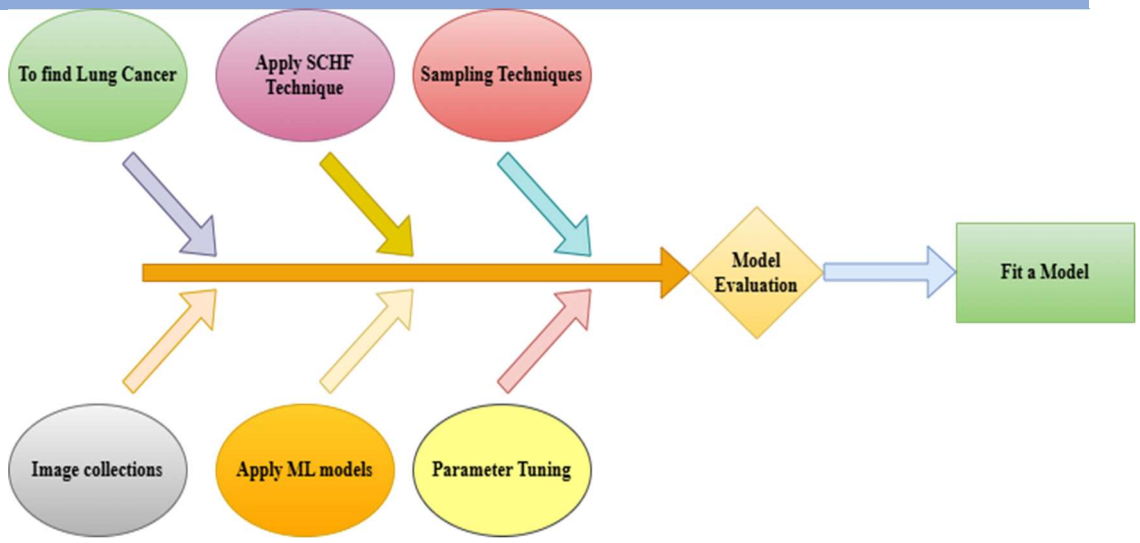

Figure 1: Proposed Architecture

The architecture shows the flow process of this research work. The collected dataset to be applied image filtering and features selection through learning models in weka 3.9.5 open-source tool by 10:90 sampling techniques.

This work considers following algorithms:

- Simple Color Histogram Filter Technique: A color histogram is a graphical representation of the distribution of colors in an image. It quantifies how frequently each color appears, without considering the spatial arrangement of pixels. The "simple" aspect refers to its straightforward computation and ease of implementation.
- Naïve Bayes (NB) is a method that calculates the posterior probability of each class based on the observable data. The predicted class is determined by selecting the class with the highest probability.
- Linear Regression (LR) is a statistical technique that models the connection between a dependent variable and one or more independent variables by fitting a linear equation to observed data.
- Additive Regression (AR) is an advanced technique that builds upon linear regression to predict non-linear interactions. It achieves this by mixing numerous additive components. Additive regression models the link between predictors and the response by considering the total of smooth functions of each individual predictor, rather than assuming a linear relationship.
- Naïve Bayes Multinomial (NBM) is a specialized version of the Naïve Bayes method, especially useful where qualitative characteristics are well defined and represent the frequency of terms or words in a document.
- Logistic Regression is a classification technique used for binary classification problems, it assigns the probability with which an instance may belong to a class.a 3.9.5 open-source tool by 10:90 sampling techniques.
- The Attribute Selected Classifier (ASC) is a method used to choose a subset of important characteristics from the original set. Generally, the objective is towards optimization of output of a classifier.

Algorithm: SCHF with Hybrid ML Techniques

The SCHF is an effective method for color-based image retrieval and can be useful for extracting color and texture features from CT images.

Input: Large cell carcinoma, Squamous cell carcinoma, Adenocarcinoma, Normal CT images

Output: Fit an efficient model for diagnosing lung cancer

Here's the updated algorithm with mathematical notation, including the ACC filter:

1. Data Representation:

Let I = {I_1, I_2, ..., I_n} be the set of input CT images Let Y = {y_1, y_2, ..., y_n} be the set of corresponding labels where y_i ∈ {Large cell carcinoma, Squamous cell carcinoma, Adenocarcinoma, normal}

2. Simple Color Histogram Filter: For each image I in the dataset:
   SCHF(I) = {γ^(k)_c(I)}(c∈C, k∈K) where:
   - C is the set of quantized colors
   - K is the set of distance values
   - γ^(k)_c(I) is the probability of finding a pixel of color c at distance k from a pixel of the same color

   Mathematically, γ^(k)_c(I) is defined as: γ^(k)_c(I) = Pr(p_2 ∈ I_c | p_1 ∈ I_c, ||p_1 - p_2|| = k) where:
   - I_c is the set of pixels with color c in image I
   - p_1 and p_2 are pixels in I
   - ||p_1 - p_2|| is the distance between p_1 and p_2
3. Feature Extraction: X = SCHF (I) = { SCHF (I_1), SCHF (I_2), ..., SCHF (I_n)}
4. Feature Selection (optional): X' = F(X), where F is the feature selection function
5. Data Split: (X_train, y_train), (X_test, y_test) = split(X', Y)
6. For each classifier:
   a) Naive Bayes: P(y|x) = P(x|y)P(y) / P(x)
   b) Linear Regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$
   c) Additive Regression: $f(x) = f_0(x) + \beta_1 f_1(x) + \beta_2 f_2(x) + ... + \beta_m f_m(x)$
   d) Naive Bayes Multinomial: $P(y|x) = P(y) \prod_i P(x_i|y) / P(x)$
   e) Logistic Regression: P(y=1|x) = 1 / (1 + e^(-z)) where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$
   f) Attribute Selected Classifier: X'' = S(X'), y = C(X'') where S is the attribute selection function and C is the chosen classifier.
7. Model Evaluation: Accuracy = (TP + TN) / (TP + TN + FP + FN) Precision = TP / (TP + FP) Recall = TP / (TP + FN) F1-score = 2 * (Precision * Recall) / (Precision + Recall)
8. Model Selection: $M *= argmax_M Evaluation_M etric(M)$

IV Outcome and Interpretations

This section focuses the outcome of SCHF + AR, SCHF+NB, SCHF+LR, SCHF+ASC, SCHF+NBM, and SCHF+Logistic models. The above table 2 shows the accuracy, precision, recall, receiver operating characteristic curve (ROC) and precision recall curve (PRC) value of SCHF+AR, SCHF+NB, SCHF+LR, SCHF+ASC, SCHF+NBM, and SCHF+Logistic models.

Table 1: Classifiers Vs Classification Outcomes

S.No	Classifier	Accuracy	Precision	Recall	ROC	PRC
1	Simple Color Histogram Filter + Naive Bayes	86.81%	0.87	0.86	0.88	0.88
2	Simple Color Histogram Filter + Linear Regression	86.33%	0.88	0.66	0.89	0.89
3	mple Color Histogram Filter + Additive egression	85.51%	0.86	0.86	0.89	0.88

4	Simple Color Histogram Filter + Naive Bayes Multinomial	90.62%	0.91	0.91	0.94	0.94
5	Simple Color Histogram Filter + Logistic Regression	83.75%	0.88	0.68	0.89	0.89
6	Simple Color Histogram Filter + Attribute Selected Classifier	86.62%	0.89	0.81	0.91	0.91

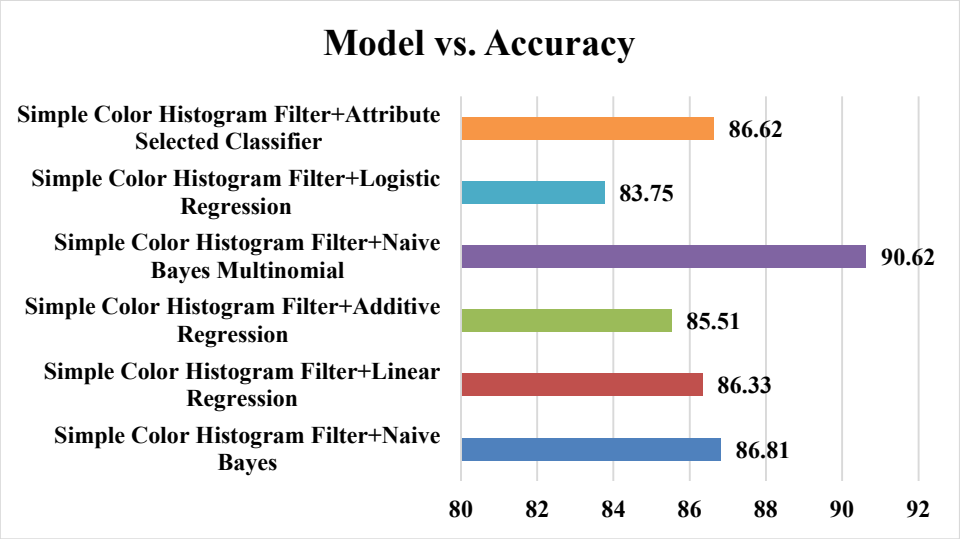

Figure 3: Model Vs Accuracy

The bar chart 3 "Model vs. Accuracy" highlights the performance of various models using the Simple Color Histogram Filter (SCHF) for lung cancer classification. Among the models, the Naive Bayes Multinomial achieved the highest accuracy at 90.62%, showcasing its ability to handle categorical data effectively. This was followed closely by Naive Bayes (86.81%), the Attribute Selected Classifier (86.62%), and Linear Regression (86.33%), all of which performed well, suggesting that SCHF features align well with these algorithms. Additive Regression attained an accuracy of 85.51%, indicating its limitations in capturing complex data patterns, while Logistic Regression recorded the lowest accuracy at 83.75%, reflecting its struggle with nonlinear relationships in the SCHF features.

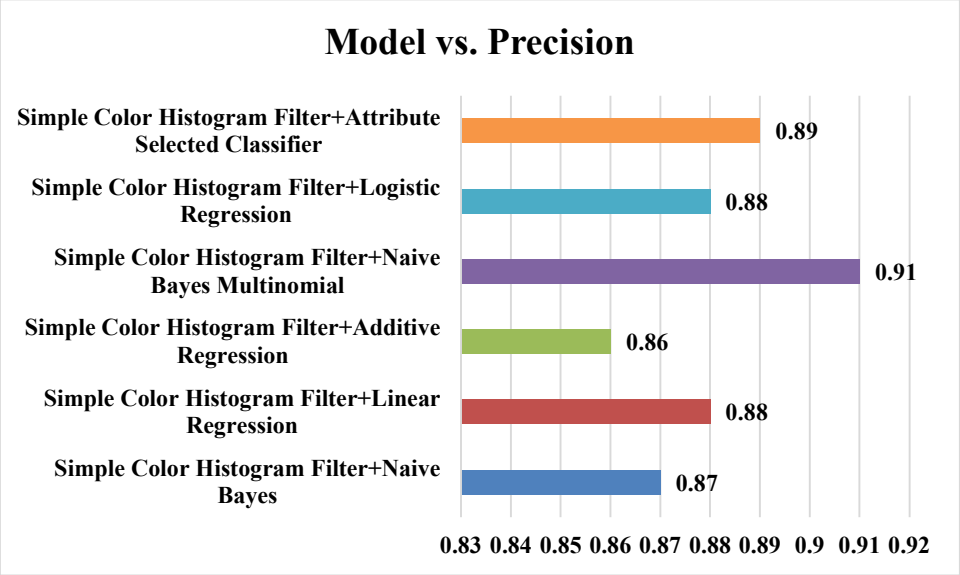

Figure 4: Model Vs Precision

The bar chart 4 "Model vs. Precision" showcases the precision performance of different models using the Simple Color Histogram Filter (SCHF) for lung cancer classification. The Naive Bayes Multinomial model achieved the highest precision at 0.91, indicating its superior ability to minimize false positives. The Attribute Selected Classifier followed with a precision of 0.89, demonstrating effective feature selection for accurate classification. Logistic Regression and Linear Regression both achieved a precision of 0.88, reflecting their capability to handle the SCHF features efficiently. Naive Bayes achieved a slightly lower precision of 0.87, while Additive Regression recorded the lowest precision at 0.86, indicating some limitations in handling the complexity of the data.

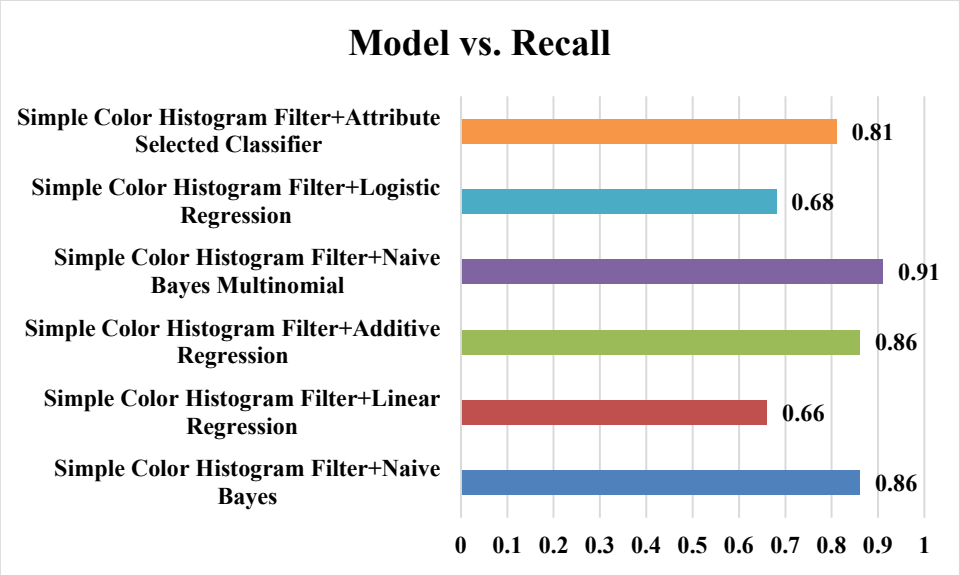

Figure 5: Model Vs Recall

The bar chart 5 "Model vs. Recall" illustrates the recall performance of different models using the Simple Color Histogram Filter (SCHF) for lung cancer classification. The Naive Bayes Multinomial model achieved the highest recall at 0.91, demonstrating its exceptional ability to correctly identify true positives. Both the Additive Regression model and Naive Bayes achieved a recall of 0.86, showcasing strong performance in identifying relevant instances. The Attribute Selected Classifier followed with a recall of 0.81, reflecting its effectiveness in selecting meaningful features. Logistic

Regression and Linear Regression recorded lower recall values at 0.68 and 0.66, respectively, indicating their limitations in capturing all relevant cases.

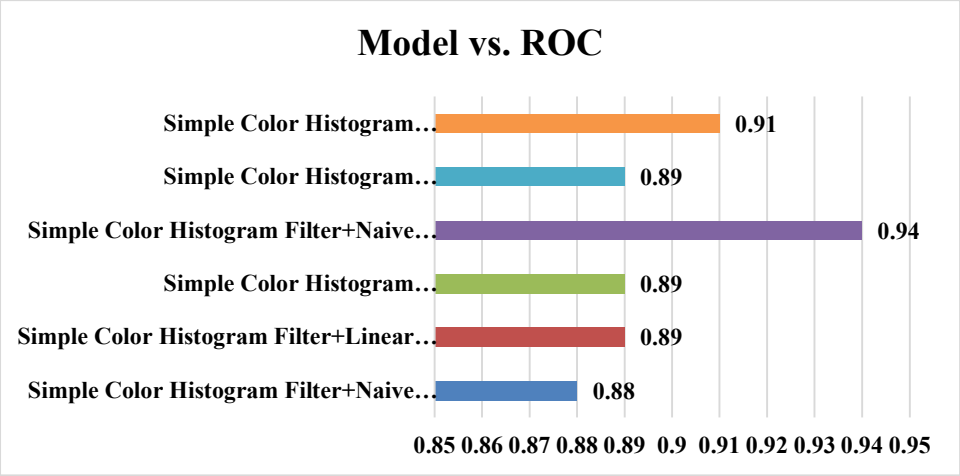

Figure 6: Model Vs ROC

The bar chart 6 "Model vs. ROC" compares the Receiver Operating Characteristic (ROC) performance of various models utilizing the Simple Color Histogram Filter (SCHF) for lung cancer classification. The Naive Bayes Multinomial model achieved the highest ROC value at 0.94, indicating its superior ability to distinguish between classes. The Attribute Selected Classifier followed closely with an ROC value of 0.91, reflecting its effective feature selection and classification capabilities. Logistic Regression, Additive Regression, and Linear Regression models each attained an ROC of 0.89, showcasing their consistent and reliable performance in this context. The Naive Bayes model recorded a slightly lower ROC at 0.88, but it still demonstrated solid classification performance.

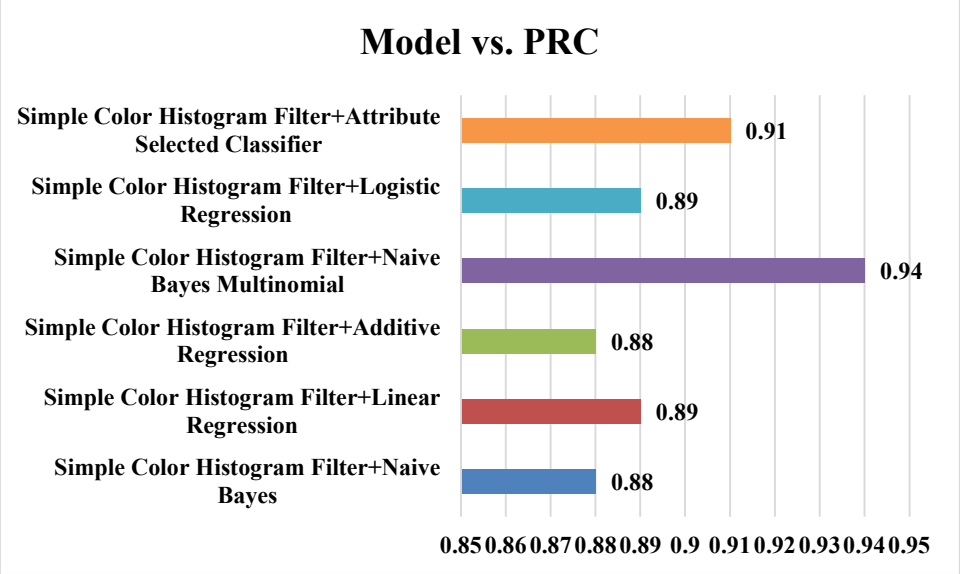

Figure 7: Model Vs PRC

The bar chart 7 "Model vs. PRC" compares the Precision-Recall Curve (PRC) performance of various models using the Simple Color Histogram Filter (SCHF) for lung cancer classification. The Naive Bayes Multinomial model achieved the highest PRC score at 0.94, demonstrating its superior capability in handling imbalanced data and accurately identifying positive instances. The Attribute Selected Classifier followed closely with a PRC of 0.91, highlighting its effectiveness in selecting relevant features. Logistic Regression and Linear Regression models both achieved a PRC score of 0.89, reflecting their reliable performance in leveraging SCHF features. Additive Regression and Naive Bayes recorded slightly lower PRC scores of 0.88, indicating decent but relatively less effective performance compared to the

leading models.

Table 2: Classifiers Vs Statistical outcome

S.No	Classifier	Time	Kappa	F-Measure	MCC
1	Simple Color Histogram Filter + Naive Bayes	0.28	0.54	0.61	0.62
2	Simple Color Histogram Filter + Linear Regression	0.03	0.52	0.59	0.61
3	Simple Color Histogram Filter + Additive Regression	0.03	0.61	0.68	0.58
4	Simple Color Histogram Filter + Naive Bayes Multinomial	0.09	0.71	0.78	0.75
5	Simple Color Histogram Filter + Logistic Regression	0.17	0.35	0.68	0.64
6	Simple Color Histogram Filter + Attribute Selected Classifier	0.09	0.61	0.59	0.59

The table 2 above illustrates the time consumption, Kappa, F-Measure, and Matthews Correlation Coefficient (MCC) values for the SCHF+AR, SCHF+NB, SCHF+LR, SCHF+ASC, SCHF+NBM, and SCHF+Logistic models.

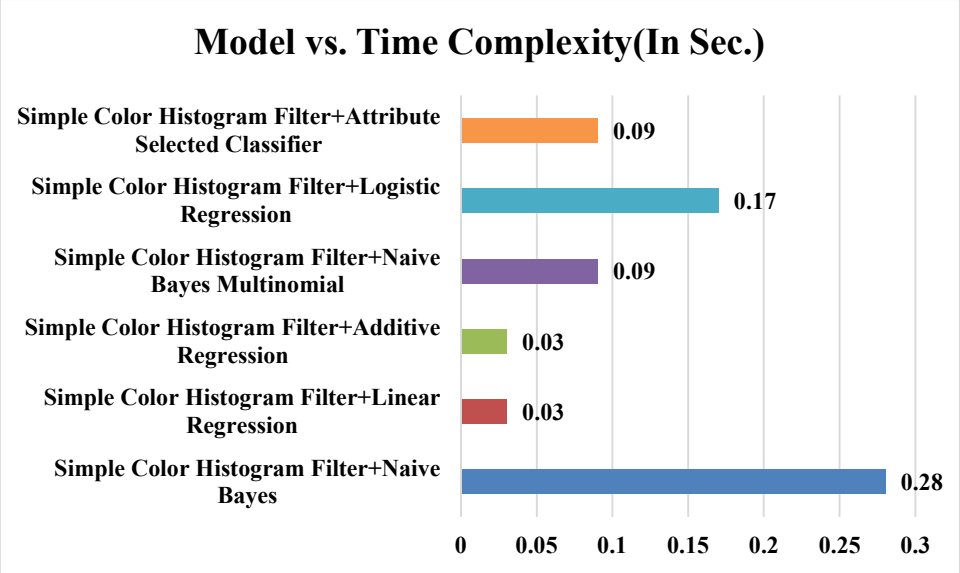

Figure 8: Model Vs Time

The bar chart 8 "Model vs. Time Complexity (In Sec.)" illustrates the time required by various models using the Simple Color Histogram Filter (SCHF) for lung cancer classification. The Naive Bayes model had the highest time complexity at 0.28 seconds, indicating its computational intensity compared to other models. Logistic Regression followed with a time complexity of 0.17 seconds, reflecting moderate computational demands. The Attribute Selected Classifier and Naive Bayes Multinomial models each required 0.09 seconds, showcasing their efficiency in processing SCHF features. Additive Regression and Linear Regression were the fastest, both completing computations in just 0.03 seconds, demonstrating their simplicity and speed.

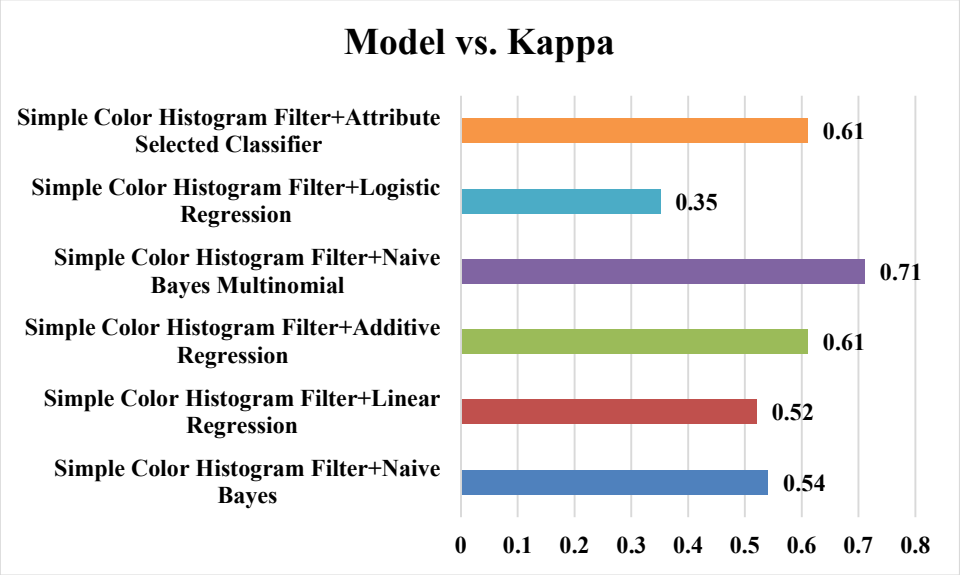

Figure 9: Model Vs Kappa

The bar chart 9 "Model vs. Kappa" compares the Cohen's Kappa values for various models using the Simple Color Histogram Filter (SCHF) in lung cancer classification, reflecting the models' agreement with true labels. The Naive Bayes Multinomial model achieved the highest Kappa value at 0.71, indicating strong reliability and agreement. Both the Attribute Selected Classifier and Additive Regression models achieved a Kappa value of 0.61, showcasing good consistency in classification. The Naive Bayes model followed with a moderate Kappa score of 0.54, while Linear Regression achieved 0.52, reflecting lower agreement but still reasonable performance. Logistic Regression recorded the lowest Kappa value of 0.35, suggesting limited reliability compared to other models.

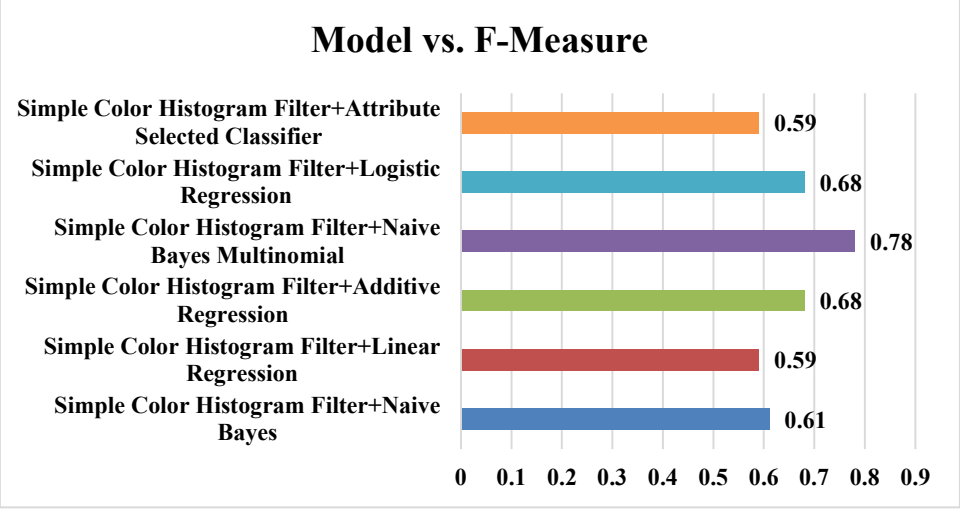

Figure 10: Model Vs F-Measure

The bar chart 10 "Model vs. F-Measure" illustrates the F-measure performance of various models utilizing the Simple Color Histogram Filter (SCHF) for lung cancer classification, which combines precision and recall into a single metric. The Naive Bayes Multinomial model achieved the highest F-measure of 0.78, demonstrating its ability to balance precision and recall effectively. Logistic Regression and Additive Regression both performed well, with F-measure values of 0.68, indicating their reliability in maintaining a good balance between false positives and false negatives. The Naive Bayes model followed with an F-measure of 0.61, while the Attribute Selected Classifier and Linear Regression both recorded lower scores of 0.59, reflecting comparatively less effective performance.

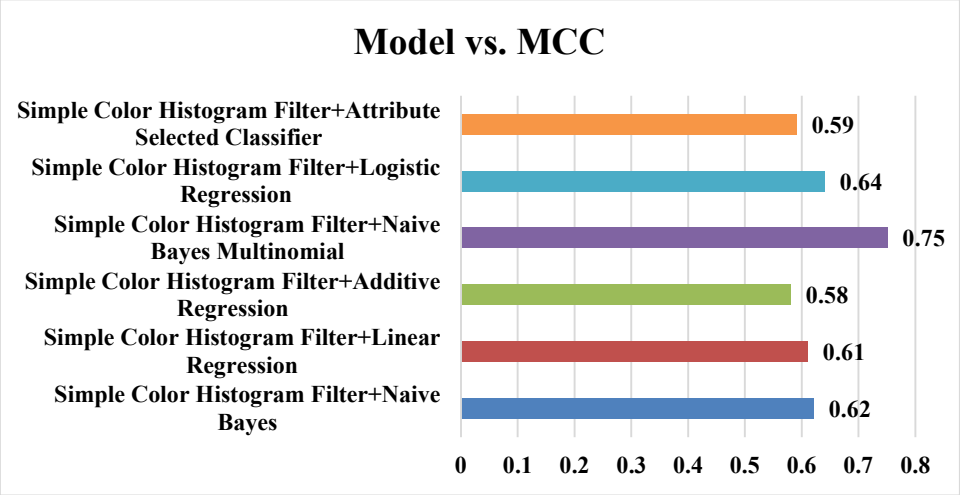

Figure 11: Model Vs MCC

The bar chart 11 "Model vs. MCC" presents the Matthews Correlation Coefficient (MCC) performance of various models using the Simple Color Histogram Filter (SCHF) for lung cancer classification. The Naive Bayes Multinomial model achieved the highest MCC value of 0.75, indicating its strong ability to make accurate and balanced predictions across all classes. Logistic Regression followed with a score of 0.64, showcasing reliable performance in leveraging the SCHF features. The Naive Bayes model and Linear Regression achieved MCC values of 0.62 and 0.61, respectively, reflecting moderate prediction consistency. The Attribute Selected Classifier and Additive Regression recorded lower MCC scores of 0.59 and 0.58, suggesting relatively less effective performance.

Table 3: Classifiers Vs Errors

S.No	Classifier	MAE	RMSE	RAE	RRSE
1	Simple Color Histogram Filter + Naive Bayes	0.41	0.43	82.39	96.18
2	Simple Color Histogram Filter + Linear Regression	0.33	0.56	66.86	112.44
3	Simple Color Histogram Filter + Additive Regression	0.25	0.46	58.99	98.14
4	Simple Color Histogram Filter + Naive Bayes Multinomial	0.14	0.38	28.99	76.14
5	Simple Color Histogram Filter + Logistic Regression	0.56	0.59	72.03	100.61
6	Simple Color Histogram Filter + Attribute Selected Classifier	0.38	0.55	56.99	99.14

The above table 3 depicts the Mean Absolute Error (MAE), Relative Absolute Error (RAE), Root Measure Squared Error (RMSE), and Relative Root Squared Error (RRSE) of SCHF+AR, SCHF+NB, SCHF+LR, SCHF+ASC, SCHF+NBM, and SCHF+Logistic models.

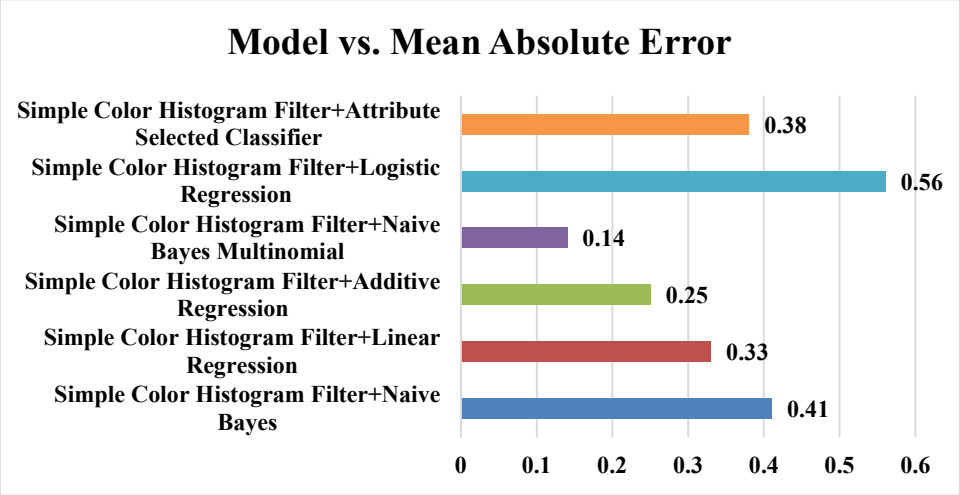

Figure 12: Model Vs MAE

The bar 12 chart "Model vs. Mean Absolute Error" compares the error rates of various models using the Simple Color Histogram Filter (SCHF) for lung cancer classification. The Naive Bayes Multinomial model demonstrated the lowest mean absolute error (MAE) at 0.14, highlighting its superior accuracy and minimal deviation from true values. Additive Regression followed with an MAE of 0.25, reflecting reasonable error control. Linear Regression and the Attribute Selected Classifier recorded MAEs of 0.33 and 0.38, respectively, indicating moderate performance. The Naive Bayes model exhibited a slightly higher error at 0.41, while Logistic Regression had the highest MAE at 0.56, suggesting lower accuracy and greater deviations.

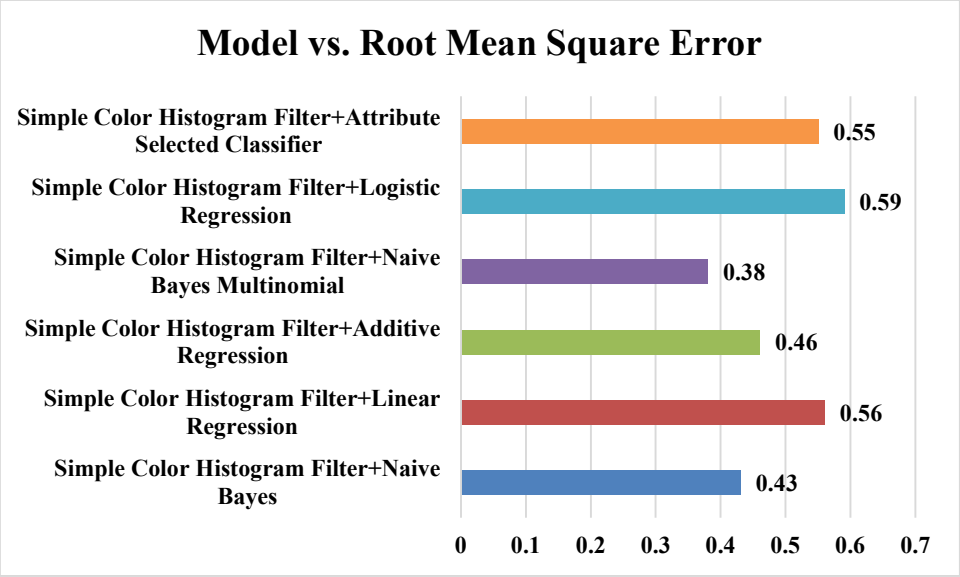

Figure 13: Model Vs RMSE

The bar chart 13 "Model vs. Root Mean Square Error" compares the performance of various models using the Simple Color Histogram Filter (SCHF) for lung cancer classification in terms of Root Mean Square Error (RMSE). The Naive Bayes Multinomial model achieved the lowest RMSE at 0.38, indicating its superior ability to minimize prediction errors. The Naive Bayes model followed with an RMSE of 0.43, demonstrating good accuracy. Additive Regression recorded an RMSE of 0.46, reflecting moderate performance. The Attribute Selected Classifier and Linear Regression showed similar RMSE values of 0.55 and 0.56, respectively, while Logistic Regression had the highest RMSE at 0.59, suggesting relatively lower predictive accuracy.

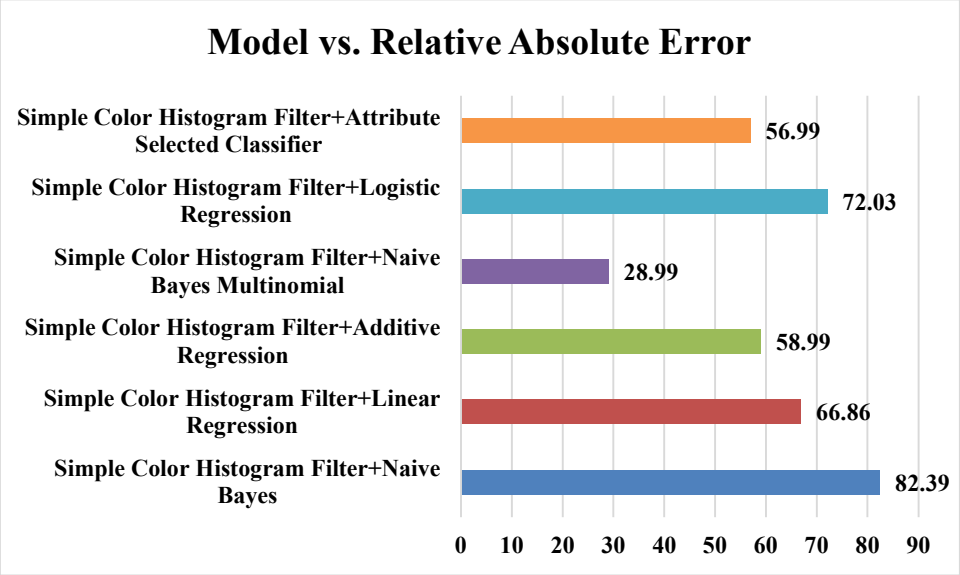

Figure 14: Model Vs RAE

The bar chart 14 "Model vs. Relative Absolute Error" compares the relative absolute error (RAE) of various models using the Simple Color Histogram Filter (SCHF) for lung cancer classification. The Naive Bayes Multinomial model demonstrated the lowest RAE at 28.99%, indicating its exceptional accuracy and minimal deviation from the expected outcomes. The Additive Regression and Attribute Selected Classifier models followed with RAEs of 58.99% and 56.99%, respectively, reflecting moderate error levels. Linear Regression recorded an RAE of 66.86%, while Logistic Regression exhibited a higher error rate at 72.03%. The Naive Bayes model had the highest RAE at 82.39%, suggesting relatively lower performance.

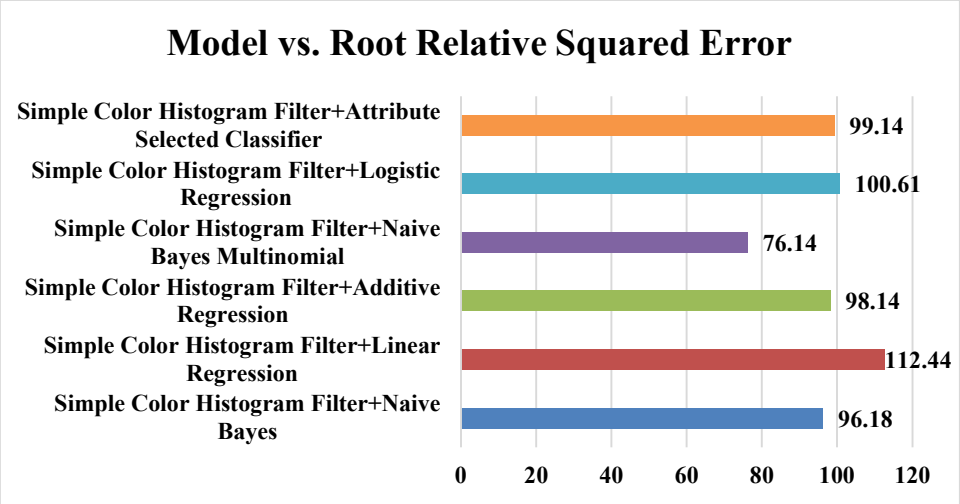

Figure 15: Model Vs RRSE

The bar chart 15 "Model vs. Root Relative Squared Error" compares the performance of various models using the Simple Color Histogram Filter (SCHF) for lung cancer classification in terms of Root Relative Squared Error (RRSE). The Naive Bayes Multinomial model achieved the lowest RRSE at 76.14, indicating its superior accuracy and lower variance from expected predictions. The Naive Bayes model followed with an RRSE of 96.18, demonstrating good performance. The Additive Regression and Attribute Selected Classifier models recorded RRSEs of 98.14 and 99.14, respectively, reflecting moderate error levels. Logistic Regression had a slightly higher RRSE of 100.61, while Linear Regression exhibited the highest error at 112.44, indicating comparatively less precise predictions.

V Conclusion

This study demonstrates the efficacy of machine learning models for lung cancer classification, emphasizing the importance of feature extraction techniques like the Simple Color Histogram Filter (SCHF). Among the evaluated models, the Naive Bayes Multinomial emerged as the most robust and accurate, consistently achieving superior performance across multiple metrics, including accuracy, precision, recall, and F-measure. Additionally, it minimized error rates such as MAE, RMSE, and RAE, making it a reliable choice for lung cancer detection tasks. While models like Logistic Regression and Linear Regression showed moderate performance, they were less effective in handling the complexity of SCHF features. These findings establish SCHF as a viable preprocessing method and Naive Bayes Multinomial as a strong candidate for lung cancer classification. Future research could explore ensemble learning techniques or hybrid models to further enhance performance and address limitations in complex data scenarios, ultimately contributing to more efficient and accurate diagnostic systems in healthcare.

References

1. Javed, R., Abbas, T., Khan, A.H. et al. Deep learning for lungs cancer detection: a review. Artif Intell Rev 57, 197 (2024). https://doi.org/10.1007/s10462-024-10807-1
2. Ahmed, Bakhan. (2019). Lung Cancer Prediction and Detection Using Image Processing Mechanisms: An Overview. Signal and Image Processing Letters. 1. 10.31763/simple.v1i3.11.
3. Jiang W, Zeng G, Wang S et al (2022) Application of Deep Learning in Lung Cancer Imaging Diagnosis. J Healthc Eng 2022:1–12. https://doi.org/10.1155/2022/6107940
4. Manjula Devi R, Dhanaraj RK, Pani SK et al (2023) An improved deep convolutionary neural network for bone marrow cancer detection using image processing. Inf Med Unlocked 101233. https://doi.org/10. 1016/j.imu.2023.101233
5. Masood I, Wang Y, Daud A et al (2018) Towards Smart Healthcare: Patient Data Privacy and Security in Sensor Cloud Infrastructure. Wirel Commun Mob Comput 2018:1–23. https://doi.org/10.1155/2018/2143897
6. Jamshaid Iqbal Janjua, Tahir Abbas Khan, Nadeem M (2022) Chest x-ray anomalous object detection and classification framework for medical diagnosis. 2022 International conference on information networking (ICOIN). https://doi.org/10.1109/icoin53446.2022.9687110
7. Shah, A.A., Malik, H.A.M., Muhammad, A. et al. Deep learning ensemble 2D CNN approach towards the detection of lung cancer. Sci Rep 13, 2987 (2023). https://doi.org/10.1038/s41598-023-29656-z
8. Zuo, W., Zhou, F., Li, Z. & Wang, L. Multi-resolution cnn and knowledge transfer for candidate classifcation in lung nodule detection. IEEE Access 7, 32510–32521 (2019).
9. Wankhade, S., Vigneshwari, S. Lung cell cancer identification mechanism using deep learning approach. Soft Comput (2023). https://doi.org/10.1007/s00500-023-08661-4
10. Liu X, Li K-W, Yang R, Geng L-S (2021) Review of deep learning based automatic segmentation for lung cancer Radiotherapy. Front Oncol 11. https://doi.org/10.3389/fonc.2021.717039
11. Chae KJ, Jin GY, Ko SB et al (2020) Deep Learning for the Classification of Small (≤2 cm) Pulmonary Nodules on CT Imaging: A Preliminary Study. Acad Radiol 27:e55–e63. https://doi.org/10.1016/j. acra.2019.05.018
12. Ali F, Kumar H, Patil S et al (2022) Target-DBPPred: An intelligent model for prediction of DNA-binding proteins using discrete wavelet transform based compression and light eXtreme gradient boosting. Comput Biol Med 145:105533–105533. https://doi.org/10. 1016/j.compbiomed.2022.105533
13. Javed R, Abbas T, Jamshaid Iqbal Janjua et al (2023) wrist fracture prediction using transfer learning, a case study. J Popul Ther Clin Pharmacol 30. https://doi.org/10.53555/jptcp.v30i18.3161
14. Abbas S, Issa GF, Fatima A et al (2023) Fused Weighted Federated Deep Extreme Machine Learning Based on Intelligent Lung Cancer Disease Prediction Model for Healthcare 5.0. Int J Intell Syst 2023:1–14. https://doi.org/10.1155/2023/2599161

15. Ghasemi Darehnaei, Z., Shokouhifar, M., Yazdanjouei, H. & Rastegar Fatemi, S. M. J. SI-EDTL: Swarm intelligence ensemble deep transfer learning for multiple vehicle detection in UAV images. Int. J. Commun. Syst. https://doi.org/10.1002/cpe.6726 (2022).
16. Setio, A. A. A. et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks.IEEE Trans. Med. Imaging 35, 1160–1169 (2016).
17. 12. Xie, Y. et al. Knowledge-based collaborative deep learning for benign-malignant lung nodule classifcation on chest CT. IEEE Trans. Med. Imaging 38, 991–1004 (2019).
18. Rao, G. S., Kumari, G. V., & Rao, B. P. Network for biomedical applications. vol. 2 (Springer Singapore, 2019).
19. Raj, S.., Shankar, G., Murugesan, S., Raju, M. N. ., Mohan, E., & Rani, P. J. I. (2023). Exploratory Data Analysis on Blueberry yield through Bayes and Function Models. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(11s), 634–641. https://doi.org/10.17762/ijritcc.v11i11s.8299
20. Thivakaran, T. K. ., Priyanka, N. ., Antony, J. C. ., Surendran, S. ., Mohan, E. ., & Innisai Rani, P. J. . (2023). Exploratory Data Analysis for Textile Defect Detection. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(9s), 121–128. https://doi.org/10.17762/ijritcc.v11i9s.7403
21. Rani, P. J. I. ., Venkatachalam, K. ., Sasikumar, D. ., Madhankumar, M. ., A., T. ., Senthilkumar, P. ., & Mohan, E. . (2024). An Optimal Approach on Electric Vehicle by using Functional Learning . *International Journal of Intelligent Systems and Applications in Engineering*, *12*(13s), 197–206.
22. Wang, W. et al. Exploring cross-image pixel contrast for semantic segmentation. In Proceedings of the. IEEE Int. Conf. Comput. Vis.7283–7293 (2021) doi:https://doi.org/10.1109/ICCV48922.2021.00721.
23. Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y. & Ettaouil, M. Multilayer perceptron: Architecture optimization and training.Int. J. Interact. Multimed. Artif. Intell. 4, 26 (2016).
24. Berwick, R. An Idiot's Guide to Support vector machines (SVMs): A New Generation of Learning Algorithms Key Ideas. Village Idiot 1–28 (2003).
25. Kavitha, P., Ayyappan, G., Jayagopal, P., ... Alqahtani, M.S., Soufiene, B.O., Detection for melanoma skin cancer through ACCF, BPPF, and CLF techniques with machine learning approach,*BMC Bioinformatics*,2023, 24(1), 458
26. Ayyappan, G., Veeralakshmi, P., Reena, R., Senthilkumar, S.R., Sureshbabu, N.G.K., Knowledge discovery in heart disease dataset, *AIP Conference Proceedings*,2022, 2393, 020136
27. R.Sugumar , Dr.E.Mohan. "Magnetic Resonance Imaging Segmentation for Brain Tumor Detection Using New Robust Global Kernel Fuzzy C-Means Clustering Algorithm (NRGKFCM-F)," International Journal of Applied Engineering Research – Volume 9 ,issue 21, 10889-10908, 2014, *(*ISSN: 0973-4562).
28. Thambu Gladstan , Dr.E.Mohan. "Object Recognition Based on Wave Atom Transform," Research Journal of Applied Sciences, Engineering and Technology – Volume 8 ,issue 13, 1613-1617, 2014, ISSN: 2040-7459; e-ISSN: 2040-7467
29. Dr.E.Mohan, Dr.A.Annamalai Giri,S.V.AswinKumer "A Novel Image Segmentation Approach for Brain Tumor Detection Using Dual Clustering Approach" International Journal of Applied Engineering Research, Volume 13 ,issue 11, 9807-9810, 2018, *(*ISSN: 0973-4562).
30. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images