I.J. Image, Graphics and Signal Processing, **2023, 3, 1-13**
Published Online on June 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2023.03.01

COVID-19 and Malaria Parasite Detection and Classification by Bins Approach with Statistical Moments Using Machine Learning

Hrishikesh Telang
Syracuse School of Information Studies/Information Systems, Syracuse, 13244, United States
E-mail: hmtelang@syr.edu
ORCID iD: https://orcid.org/0000-0003-4400-6896

Kavita Sonawane*
St. Francis Institute of Technology/Computer Engineering, Mumbai, 400103, India
E-mail: kavitasonawane@sfit.ac.in
ORCID iD: https://orcid.org/0000-0003-0865-6760
*Corresponding Author

Received: 17 July, 2022; Revised: 12 August, 2022; Accepted: 02 September, 2022; Published: 08 June, 2023

Abstract: This work introduces the novelty as an application of histogram-based bins approach with statistical moments for detecting and classifying malaria using blood smear images into parasitized and uninfected cell images and the rising disease of COVID-19/Normal lung images. Proposed algorithms greatly vary as compared to the previous work. This work aims to improve accuracy in detection and classification and reduce feature vector dimensionality. It focuses on detailed image contents extracted into 8 bins by considering the significance of the R, G, and B color component relationship in the formation of each pixel. The texture features are represented by the first four moments for each of the three colors separately. This leads to the generation of 12 features vectors, each of size 8 components for each image in the database. Feature dimensionality reduction is achieved by applying different feature selection techniques to obtain desired optimum feature space. The comprehensive feature analysis presented here identifies many useful findings in order to validate the contribution of each image content uniquely in detection and classification. The proposed approach experimented with two image datasets: the malaria dataset obtained from the National Library of Medicine (NLM) and the lung image dataset acquired from the Radiography Database from Kaggle. The performance of work presented here is evaluated and compared with previous work with the same set of parameters, namely precision, recall, F1 score, and the AUC. We have achieved and improved the performances compared to previous work and also achieved better results even for the COVID-19 dataset.

Index Terms: Index, 8 bins approach, malaria, COVID-19, statistical moments, accuracy, precision, recall, ROC-AUC, feature engineering, data analysis, feature selection, dimensionality reduction.

1. Introduction

Our previous research paper [1] proposed and applied a novel Bins Approach to segregate pixels into bins, using simple histogram thresholding of the image efficiently, thereby reducing feature vector size without disturbing the feature components and focusing on image histograms to extract the image features. This approach considers the count of pixel values but not the intensities. Furthermore, we tried to analyze the image texture by computing the statistical properties of the R, G, B planes of an image [1]. During our study, we were also motivated to address the recent threats and challenges posed by the COVID-19 pandemic worldwide, beleaguering people's lives and leading to countless deaths. Researchers have come up with some prediction systems to contribute for the need of an hour [28] The disease caused by the novel coronavirus transformed the very primacy of healthcare and several other aspects of education, travel, politics, and social conditions. At the time of writing this paper, the pandemic had spread rampantly across 223 countries, with over 116,874,912 confirmed cases and 2,597,381 deaths, with the United States of America (USA) soaring at an all-time high, peaking at nearly 28.7 million confirmed cases and about 521,625 deaths (March 9, 2021)

[2]. Thus, in furtherance to the previously acquired results, we performed advanced forms of feature engineering in this research and attempted to evaluate its performance over the COVID-19 radiography database images.

Color has always been regarded as one of the most visually appealing aspects of an image [3]. In the past, many researchers have used color histograms as the image representation for feature extraction and analysis [3,29]. Every colored pixel of a respective RGB plane holds a specific intensity value between 0 to 255. One such form of feature engineering was the process of extracting and analyzing these pixel intensity contents using suitable statistical moments such as Mean, Standard Deviation, Skewness, and Kurtosis [4]. In this research work, we segregate the respective intensity values of each RGB channel of the image into each of the eight bins and perform all of the moment calculations as mentioned earlier with the hope of achieving improvement in the output. This technique tries to extract the core image details which cannot be seen by naked eyes and ensures that a significant contribution of every small intensity value of each pixel is retained and stored in one of the eight bins.

Another aspect of feature engineering is dimensionality reduction, which plays an essential role in reducing the computational complexity and, in turn, processing time. The application of bins approach generates 12 different types of features, where each feature vector is of size 8 components, leading to a total of 96 feature components for one image. For the new medical images of COVID-19 detection and classification, we have also evaluated the performance using the feature vector, i.e., counting pixels into eight bins. So along with 96 components, the COVID-19 dataset is experimented with 96+08, i.e., 106 components for the Lung Images. Finally, we have also conducted data analysis and data visualization to understand the relationship between the feature components, identifying patterns, and deducing the inferences for the results achieved.

2. Literature Review

The system of employing statistical parameters dates nearly a decade ago when Kekre et al. gave importance to R, G, and B intensities in every pixel; these intensities have been improved further by modifying the histograms using a LOG function, followed by Bins Technique. Feature vector strength in image retrieval is evaluated using three parameters Precision-Recall Crossover Point (PRCP), Longest String (L.S.), and Length of String for Retrieving all Relevant (LSRR) [4]. It was observed that the LOG function performs far better compared to the original histogram [4]. A similar approach was also evaluated in YCbCr color space compared to RGB color space while also focusing on the texture contents of the image. Thus, 2000 BMP RGB images were converted to YbCbCr, one with scaled variation and another as the original [3,27]. The process is later computed using the four statistical moments mentioned earlier [3]. Even moments proved far better than odd, with maximum precision and recall values at 56% as an average of 200 query images [3]. Overall, YCbCr was found to be better than RGB [3]. In the subsequent work, Kekre et al. performed a slight variation in which the average RGB values of all pixels were segregated in respective bins [5]. Before arriving at 8 bins, 64, 27 bins were proposed and implemented by dividing the histogram into 4 and 3 parts, respectively [5]. (E.D.) and Absolute Distance(A.D.). It was observed that A.D. performs far better than E.D. [5]. In another research paper, Kekre et al. executed the study of multiple similarity measures and elaborated on selecting the most appropriate similarity measure for CBIR. It applied Correlation Distance (CD) in the form of '$\cos\theta$' between two vectors by extracting the first four moments into 27 bins by dividing color histograms into three parts [6]. Finally, with particular focus on the color histogram technique, Kekre et al. computed R, G, and B histograms, using 256 bins data as the feature vector. [7] It was then followed by reducing the size of the feature vector to 32 and 16 by performing linear grouping of 8 and 16 histogram bins, respectively. It was further delineated from the results that 256 bins performed better than 32 and 16 bins [7].

Biomedical signal/image processing has recently gained traction over the past few years. Alva et al. proposed an Alzheimer's Disease (A.D.) The detection System takes a T1-weighted MRI scan as the input image, which then does feature extraction using hybrid feature vectors (HFV) and later adopts a classifier to determine whether the patient has A.D. or not [8]. To pass images for feature extraction, pre-processing techniques such as orientation, linear and non-linear transformations, normalization, segmentation, and image smoothing are used. The proposed HFV method is a conglomeration of the techniques mentioned above to produce an accuracy of 93% SVM, 92% LDA, 89% KNN, and 92% Logistic Regression. Alva et al. also proposed a review on techniques for ear biometrics [9]. Typically, in this technique, the ear image is captured from a sensor, which then performs pre-processing, followed by feature extraction, which can uniquely identify the ear images [9]. In 2017, Nezhadian et al. proffered melanoma skin cancer detection focusing on extracting color and texture features [10]. Each of the components was initially evaluated using SVM and another when combined [10]. The RGB components were first segregated, zero values were removed from the matrix, and statistical moments were calculated for each RGB component [10]. The proposed system offered an accuracy of 97%, a specificity of 97%, and a sensitivity of 96% when all features were incorporated, providing the best accuracy than the previous techniques mentioned in the paper [10]. In 2017, Olugboja et al. invented a malaria parasite detection model by first cleaning up images and converting them to gray-scale images [11]. The best classifier result obtained from these images was a Gaussian SVM with a true positive rate of 99.8% and 99.2% for Linear SVM [11]. Naveen et al. designed a Breast Cancer Prediction Model Using Ensemble Machine Learning Models [12]. To achieve this, the features are normalized within a standard scale to keep them in range. Next, ensemble learning models with crossover validation and bagging techniques were built, and finally, the prediction results and the confusion matrix and classification reports

were evaluated accordingly. This later deduced that KNN and Decision Trees gave the highest accuracy with training and testing set in the ratio of 90:10. In 2018, Cai et al. focused on a lung prognosis prediction model based on SVM [13]. Due to the lack of sample data for lung cancer patients, a significant challenge is the unbalanced data categories of corresponding samples. To avoid a biased judgment of classification, we oversample the input parameters using the Borderline-SMOTE algorithm. In 2019, Krishnani et al. proposed a comprehensive approach to Predict Coronary Heart Diseases (CHD) using supervised ML classifiers such as RNN, DT, and KNN [14]. The pre-processing steps include handling missing values, feature selection, normalization, standardization, and oversampling, followed by classification and prediction. It also showed a comparative study of the classifier performance based on its accuracy. The proposed classifiers suggest that upon performing 10-fold cross-validation to generate randomness in the data, RNN was the best prediction model and gave the highest performance (accuracy=96.71%) among KNN (91.49%) and DT (92.1%).

Similarly, in 2019, Celik et al. employed classification methods such as LR, SVM, Extra Trees (E.T.), Gradient Boosting (G.B.), and RNN to predict Parkinson's Disease [15]. It takes input data of 1208 speech sets, from which 26 features of Parkinson's patients and non-patients were extracted in the forms of jitters, shimmers, pitch, periods of pulses, voice breaks, etc. The training and testing samples in the ratio of 80:20, LR, SVM, and E.T. offered an accuracy of 76.03%, 75.49%, and 73.71%, respectively. In 2018, Sarwar et al. proposed a diabetes prediction model using ML classifiers such as KNN, Naive Bayes (N.B.), SVM, DT, LR, and RNN. With the input dataset of 768 instances and nine features, the following methods were incorporated: data pre-processing, feature selection, the division into training and testing sets (70:30), classification, and prediction. The best accuracy obtained was SVM and KNN, at 77%, while NB gave 74% and DT and R.F. gave 71%. Finally, we also referred to our previous paper, which performed the same operations mentioned above for detecting Malaria parasites in blood smear images [1]. It followed the process of performing image cleansing and later incorporating bins approach, color moments, and texture moments, analyzed the performance separately using RNN, KNN, and SVM, and completed the fusion of all the features using RFE. After the fusion was performed, RNN produced an accuracy of nearly 96%, with SVM at 94.6% and KNN at 95.2%.

Evaluating performance metrics, thereby choosing the most appropriate classifier for our model, can be an exacting process. Statistical tests are thus designed to address this problem using hypothesis testing to discover hidden patterns and intricacies that cause the performance to impact the system in a certain way. We have used L1 and L2 Regularization Techniques such as Lasso and Ridge Regression, as suggested by Muthukrishnan et al. in their paper [16]. With the dataset used by Efron et al., they wanted to check which variables were getting selected using Lasso, Ridge, and OLS. Upon reviewing the median MSE being low for Lasso, it is observed that this regression predicts the model with more accuracy as it shrinks non-significant values to zero. Pushpalatha et al. incorporated the concept of a filter-based approach to classifying ten different similar-looking foodgrains [17]. Initially, 66 GLCM texture features were extracted from ten different types of foodgrains, and statistical features for each color, such as red, green, and blue, were removed. For the feature selection, they used a Correlation-based feature selection combining with Best First Search (B.F.), Genetic Search (G.S.), Greedy Step Wise (GSW), Linear Forward Search (LFS), and Subset Size Forward Search (SSFS). The model's evaluation with these reduced features was performed using six different classifiers: Bayes Net, Naïve Bayes, IBK, Kstar, and Random tree. Besides, we were also intrigued by the significance of Recursive Feature Elimination for Handwritten Recognition proposed by Zeng et al., and we added that into our algorithm [18].

The outbreak of COVID-19 caused such severe catastrophic consequences on human life and the world economy that many researchers worked in this field to acquire suitable outcomes for lung classification. However, the research performed since 2020 has only been focused on Deep Learning Techniques. Karhan et al. devised the classification of x-ray images using Deep Learning [19]. They applied the pre-trained ResNet50 model of CNN to extract deep features on chest x-ray images. For the results, about 99.5% accuracy has been achieved. Similarly, Kumar et al. adopted a Deep Learning framework, majorly focusing on SMOTE and ML classification [20]. First, images are pre-processed through image cropping and resizing. Later, the ResNet152 model was trained to classify Pneumonia and Normal X-ray images. To balance the datasets amongst the classes compared to COVID-19, SMOTE was used, later followed by fitting these datasets using machine learning classifiers. The best results obtained from this approach were from Random Forest with the Accuracy, Sensitivity, Specificity, F1-score, and AUC of 0.973, 0.974, 0.986, 0.973, and 0.997 0.977, 0.977, 0.988, 0.977, and 0.998 for XGBoost, respectively. One of the different approaches in this domain was Thepade et al., who performed COVID-19 identification by computing Luminance Chroma features [21]. The proposed method comprises training and testing phases; the former process converts the input x-ray image into three color spaces: YCrCb, Kekre-LUV, and CIE-LUV. The results obtained were promising: ExtraTree + RandomForest + SimpleLogistic produced 90.42% accuracy, followed by ExtraTree + RandomForest + NaiveBayes at 87.92% and ExtraTree + RandomForest + SimpleLogistic at 87.08% for classification of COVID-19, pneumonia, and normal class classification. Nurrahma et al. used different supervised machine learning methods such as SVM, DT, Neural Network, and the dataset will be passed with a selection feature and another without one [22]. Each classifier's performance was compared statistically using the ANOVA test to check classifier significance with and without feature selection. With the threshold value = 0.05, it is observed that the one without feature selection is more statistically significant (p-value < 0.05) than the one without (p-value > 0.05).

3. Proposed Method

This research work is primarily an extension of the previous research presented in the paper [1]. The entire experimentation designed and implemented in our last paper was aimed to detect and classify the malaria parasite accurately. After skimming through the detailed results and discussion, the following points have been summarized regarding the executed approaches [1].

The previous work [1] has presented the effectiveness of the bins approach and its applicability in the medical domain, with a particular focus on malaria detection and classification. The three methods experimented along with Bins are:

a. Convolutional Neural Networks (CNN)
b. Bins approach using 8 bins by employing the count of pixels as the feature vector
c. Computing the statistical properties of the entire image separated in R, G, and B color planes and
d. Grey-level co-occurrence matrix (GLCM) approach

With the application of Bins approach, only the count of pixels (b) as feature vectors is evaluated and compared with the algorithms mentioned above (a, c, and d) to check its effectiveness for malaria parasite detection and classification. The study proved that the performance of bins approach is truly appreciable that can be observed and analyzed clearly with a positive set of results in the paper [1]. Thus, a further extension to this work is planned in the same direction with the following objectives:

a. **Objective 1**: Further Improvement in the results using extended bins approach with statistical moments for malaria.
b. **Objective 2**: Application of this extended bins approach in another medical domain, i.e., rising COVID-19 infection -detection and classification with normal cases.
c. **Objective 3**: Feature vector dimensionality reduction to be achieved through feature selection and feature engineering: Detailed analysis concerning all feature components of bins approach and its significant contribution in achieving the accuracy for detection and classification.

I. Objective 1: Proposed application of Extended Bins Approach with statistical moments for Malaria parasite detection and classification

In the previous paper [1], as mentioned in sections IV -B and Fig. 2 [1], we had used the count of pixels extracted into 8 bins to form the feature vectors. In this technique, we neither focused on the intensity values being collected into each bin nor gave importance to the individual color component of those counted pixels into them. The modification applied here is the same as shown in [23]. We compute the first four moments by considering the intensities, namely mean, standard deviation, skewness, and kurtosis for those counts of pixels into each of the eight bins for all three R, G, and B color components [1, 7]. This leads to the generation of 4 moments into three colors, i.e., 12 types of feature vectors where the size of each feature vector is eight bins means 8 components. This feature extraction process is applied to the entire dataset, and feature databases are obtained and kept ready before using the classifiers given in Section I. For Each Image under the feature extraction process, Pixel Pi based on its three color intensities (R_i, G_i, and B_i), gets counted into one of the eight bins addressing from 000 to 111, i.e., Bin0 to Bin7 we compute the first four statistical moments given as follows:

$$\textit{Feature Vectors for R component: } R_{mean}, R_{standard\_Dev}, R_{skewness}, R_{kurtosis}$$
$$\textit{Feature Vectors for G component: } G_{mean}, G_{standard\_Dev}, G_{skewness}, G_{kurtosis}$$
$$\textit{Feature Vectors for B component: } B_{mean}, B_{standard\_Dev}, B_{skewness}, B_{kurtosis}$$

II. Objective 2: Application of modified Bins approach for COVID- 19

As discussed in Section I, the same approach is applied for COVID-19 detection and classification using x-ray images. In today's scenario, healthcare professionals face much confusion in performing accurate detection, identification, or classification of Normal or COVID-19 cases. With this proposed approach, we have tried to resolve this issue to some reasonable extent. The comprehensive system works best with the minor color information variations in the image, presenting core details of the textural information in the image. We had to convert our gray images into RGB images by the pseudo- coloring approach to take advantage of the same.

Pseudo coloring algorithm [24]:

The pseudo color image transformation is a process that is employed to convert a gray-scale image to a pseudo-colored RGB image by mapping each intensity value to color with respect to a lookup table and a function [24]. To

achieve this, the gray-scale image is first passed through three different image transformations [25]. The equations used in our project for each of the transformations are deduced in Equations (1), (2), and (3).

$$R = a\,|\sin(bx)| \tag{1}$$

$$G = a\,|\sin(bx + c)| \tag{2}$$

$$B = a|\sin(bx + 2c)| \tag{3}$$

Where a = 255, b = (2 *pi)/255, c = pi/5 are constants.

The pseudo coloring algorithm colorizes the image based on each gray-scale pixel value which is then mapped to the color ranges of RGB values using the transformation function [25-27]. The false color of a pixel is then calculated by mapping it to the lookup tables created for each RGB channel [26, 27]. These channels are then merged to form the pseudo-RGB image. Typically, this lookup table oscillates through the RGB color table from the previous process to create 768 unique colors of the image [26, 27]. With this image obtained, we produced a high contrast lung image. Since high contrasts lay lesser emphasis on the contents of the image, it was necessary to suppress the intensity values through the process of Histogram Equalization (HE) without disturbing the color balance of the image. RGB images are preferably converted to the YCbCr color space as it is specially designed for digital images. On performing HE, the image is converted back to RGB. The algorithm has been summarized in Table 1 as follows:

Table 1. Algorithm of Pseudo coloring Technique

a) Grayscale to pseudo-RGB transformation:
Step 1: *Set the value of three defined constants a = 255, b = (2 *pi)/255, c = pi/5.*
Step 2: *Create empty array needed for the lookup tables*
Step 3*: Using for loop, for every iterable i between 0 to 255,*
- *i) Compute bx = b*i.*
- *ii) Perform transformation on the r channel using the formulae* $R = a\,|\sin(bx)|$, $G = a\,|\sin(bx + c)|$, *and* $B = a\,|\sin(bx + 2c)|$
- *iii) Append the value of each of this pixel transformation to the three empty arrays corresponding to the respective channels in Step 2, thereby creating a matrix.*

Step 4: *With the input images of COVID-19 and Normal images found using the os, sys and glob Python libraries, we first read each of them in grayscale.*
Step 5*: With the Lookup table (LUT) function in cv2 library, we apply it to each of the matrices created in Step 3.*
Step 6: *Merge the channels and write the new image in new separate files.*

b) RGB to Histogram Equalization:
Step 1: *Clean and resize the noisy image.*
Step 2: *Convert the RGB into YCrCb color space.*
Step 3: *Split each of the channels and perform Histogram Equalization of the intensity plane Y.*
Step 4: *Merge the channels and convert the image into RGB image.*

Based on the algorithm mentioned above, we demonstrate such an example with a sample COVID-19 input image:

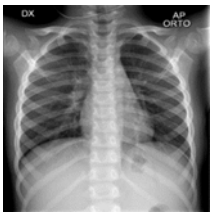


Fig. 1. Sample COVID-19 input image

As per the process of Grayscale to pseudo-RGB transformation, the following intermediate image is obtained:

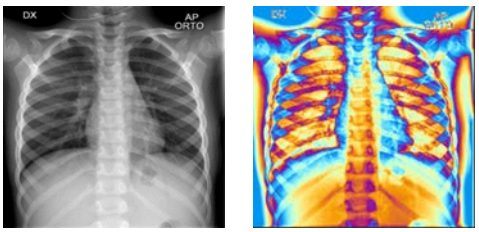


Fig. 2. Conversion of Gray-scale Lung image to pseudo-RGB transformation

As per the process of RGB to Histogram Equalization, on performing Step 1, Step 2, Step 3, and Step 4 together, we get:

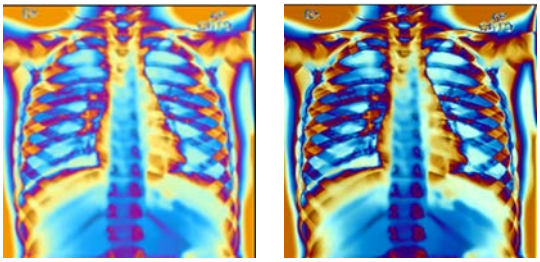


Fig. 3. Conversion of Gray-scale Lung image to YCbCr color space followed by HE and transformation to the pseudo-RGB image

As mentioned about the bins approach on malaria in section I, this dataset also undergoes the extended bins feature extraction process that leads to generating the same set of 12 types of feature vector each of size eight components. Thus, we can prepare 12 feature vector databases for 2484 COVID-19 and Normal images in the dataset.

III. Objective 3: Feature Engineering with Bins Approach:

This part of our analytical study is giving clear insights into the feature vectors. In addition, this produces a precise analysis concerning each feature component and its contribution to the results obtained based on the feature selection.

Feature Selection:

Given the significant input feature representation of an image, it is often beneficial to project the data to a lower-dimensional subspace to capture the real essence of the data and fit a better predictive model for classification. One such measure is choosing the right feature selectors to score a particular statistical test to filter the features for our input data. The three powerful feature selection techniques are filter methods (typically using statistical tests), wrapper methods, embedded methods, and hybrid methods. For our research, we used Univariate ROC-AUC as a filter method, Lasso (L1) and Ridge (L2) Regularization as embedded methods and RFE, Feature Selection by Random Shuffling (FSRS), and Recursive Feature Addition (RFA) as a hybrid method.

Classification:

This is the final step of our proposed model. It is the process that incorporates the use of machine learning principles to identify a variable that belongs to a particular class label, either between two or amongst a group of labels. Concerning the previous base paper proposed earlier, we perform a binary predictive model and extend it to other classification algorithms beyond Random Forests (RNN), K-Nearest Neighbors (KNN), and Linear SVM (LSVM), such as Logistic Regression (LR), Decision Trees (DT), Kernel SVM (KSVM), and Naïve Bayes (N.B.) classifier. The motive behind using the four classifiers in furtherance to the previous ones is that we wanted to check how the selection of particular features impacts the ROC-AUC curve of the respective algorithms and the overall performance before and after 10-fold cross-validation.

4. Experimental Setup

To fulfill the desired objectives mentioned in section III, we have experimented with the proposed extension of bins approach for two different medical application areas that are most crucial and challenging to identify and diagnose: malaria blood smear images and chest images of COVID-19 and Normal ones. The former has been obtained from the Lister Hill National Center for Biomedical Communication (LHNCBC), a part of the National Library of Medicine (NLM) [1]. It consisted of 13,779 segmented cell images, each of parasitized and uninfected cell images [1]. The latter was procured from the Radiography Database of 1143 samples of COVID-19 and 1341 samples of Normal. This dataset of images was collected from Kaggle. We performed all of the aforementioned processes using Python 3 over the Windows 10 operating system. Our previous study [1] used the same set of performance evaluation parameters, i.e., accuracy, precision, recall, and F1 measure throughout the classification and feature selection process.

5. Results and Discussions

On performing classification and feature selection of the statistical parameters of each of the color channels of the image, the following are the results obtained with 10-fold cross-validation for accuracy parameter:

1. Results obtained through Malaria blood smear images:

The following Fig. 4 suggests a range of accuracies obtained from those without passing through the feature selection process and the ones with the process. The ones relatively performing the worst are KSVM and N.B. Overall,

it can be delineated from the figure that LSVM provides the best accuracy, followed by LR and RNN and KNN interchangeably. It can also be perceived that Ridge Regularization and RFE generally provided the best accuracy for available classifiers. However, not all classifiers perform the correct guessing of class labels when looked through the lenses of the ROC-AUC scores. As per the convention of the measures obtained from the ROC-AUC curve, the higher the AUC, the better the model's performance in classifying the class labels. Following this rule, an AUC score > 0.9 is outstanding, between 0.8 and 0.9 is excellent, and smaller than or equal to 0.5 is worse than a random classifier. The following figure depicts such an aberration in values.

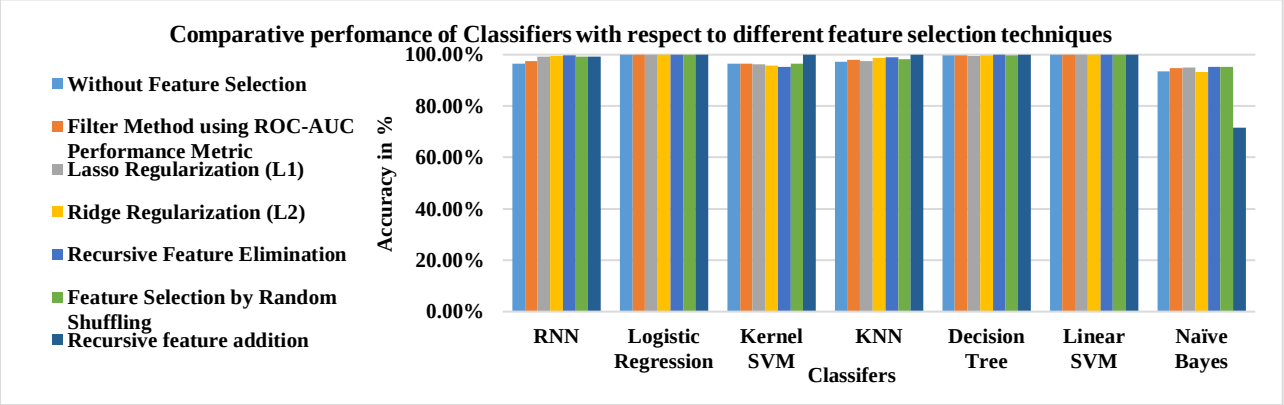


Fig. 4. A comparative analysis of different feature selection techniques performed over supervised classifiers

It is evident from Fig. 5 that KNN provides the worst classifier performance, with N.B. intermittently providing the worst values using specific feature selector algorithms. This is typically the case because KNNs are always sensitive to the size and scale of the feature vectors used in a dataset, making training computationally expensive and inefficient. Similarly, NB typically makes a probabilistic assumption of independent predictors, unlike real-world use cases, making specific estimations by a predefined hypothesis and not by a standard event likelihood. On the other hand, the rest of the classifiers are robust, distinguishing between class outputs across all feature selectors. Based on this study, we further decided to use only the best classifiers RNN, LR, KSVM, and DT, and remove the remaining classifiers, i.e., KNN and N.B., for our Lungs dataset.

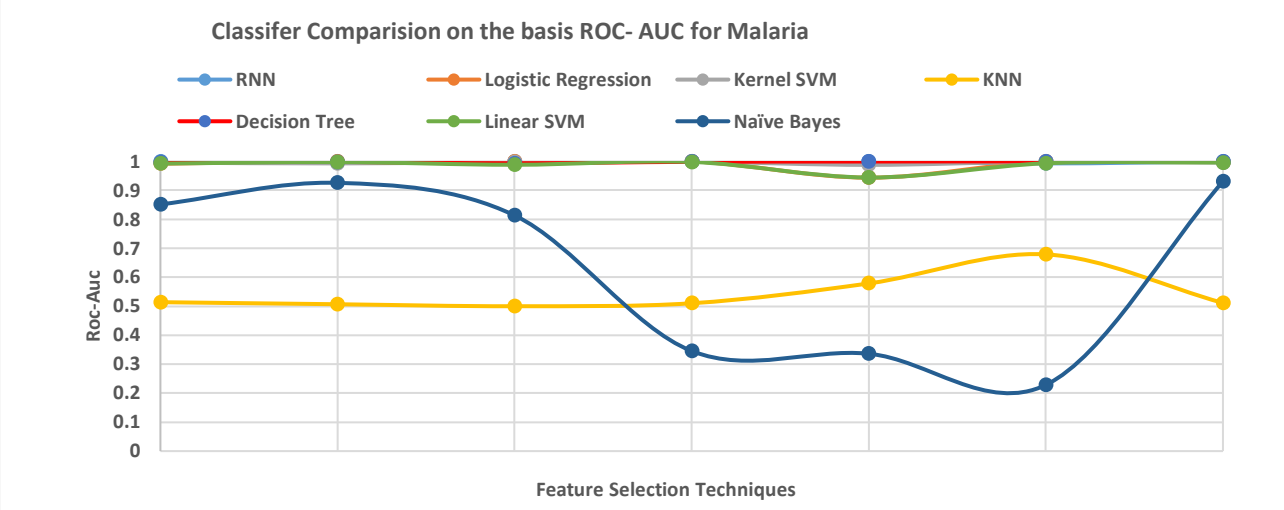


Fig. 5. ROC-AUC Performance Metric Using Extended Bins approach with Statistical Moments for Malaria parasite detection and classification

2. Results obtained through COVID-19/Normal images:

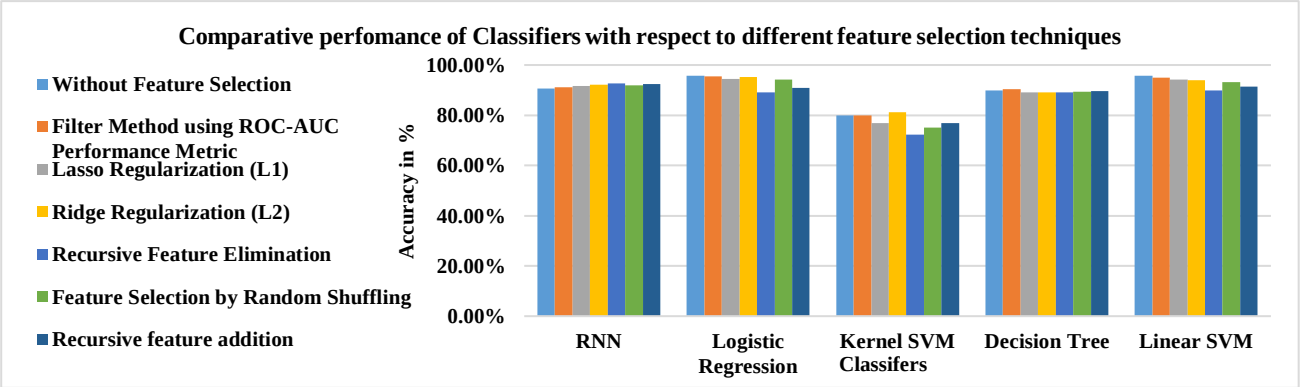


Fig. 6. Comparative analysis of the accuracy (in percentage) using Extended Bins approach with Statistical Moments for COVID-19/Normal Lung Image classification

The following Fig. 6 describes the range of accuracies obtained across different classifiers like used before. Overall, RNN, LR, and DT show improved accuracy, whereas KSVM and LR show relatively insufficient accuracy. In addition, it can be noticed that many of the feature selectors reduce the accuracy of the model rather than increasing it, with some exceptions. For example, Ridge regularization shows an accuracy of 81.30%, which is better than the one without (79.92%).

Similarly, Lasso, Ridge, RFE, and RFA show improved accuracy than the ones without them. On the other hand, classifiers like LR, DT, and LSVM do not show any significant improvement in accuracy, except when the Filter method in DT improved the accuracy by 1%. Finally, no dominant feature selector provides the best classification across all models; thus, they are individualistically regarded best for a specific classifier and an appropriate feature selector.

Fig. 7 describes the performance of the respective classifiers using the ROC-AUC version, thereby providing a different scenario: LSVM (the line shown in green) renders an accuracy of 0.5 nearly for all algorithms of feature selectors as compared to RNN, LR, and DT, peaking at 0.8 and 0.9 AUC score. Similarly, KSVM has proved inefficient at classification for three feature selectors, although the performance is relatively better than LSVM. Thus, we can conclude that LSVM and KSVM are unsuitable for predicting some of the class labels given the input feature vectors as they are somewhat haphazardly classified.

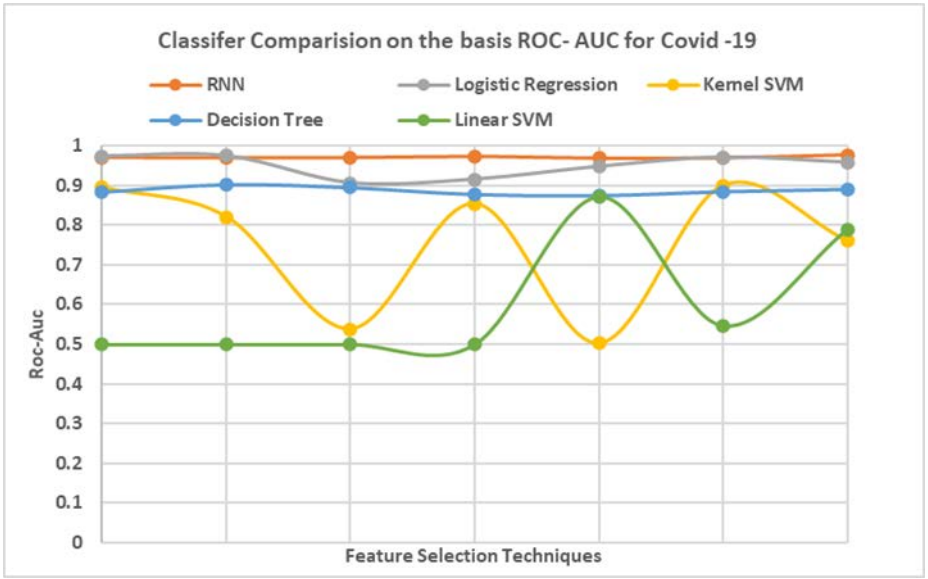


Fig. 7. ROC-AUC Performance Metric Using Extended Bins approach with Statistical Moments for COVID-19/Normal Lung Image classification

3. Feature Vector Engineering: Feature Selection:

Based on the results obtained using the best classifiers and the role of feature selectors, we thought of focusing on a detailed analysis of feature vector components as to how they contribute to obtaining the best results. A detailed study of the accuracy results obtained for all performance evaluation parameters concerning feature selection techniques for all different data sets is given in Fig. 4 and Fig. 6. Its study with respect to ROC-AUC analysis is shown in Fig. 6 and Fig. 8. The observations made are as follows:

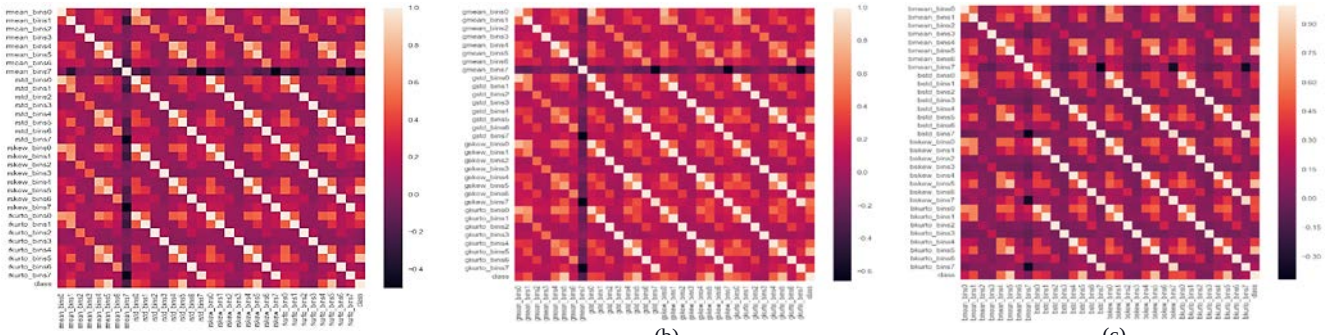
a) Malaria Parasite Dataset (Fig. 5 and Fig. 6):

The features selected by each classification model are as follows: All the feature selectors, *Lasso and Ridge Regression, RFE, RFA, and selection by FSRS* are the most efficient selectors that provided the highest possible accuracy for respective classification models, although with some being lesser than the other. The features selected by each of the classification models are also prominent: in the ROC-AUC performance metric feature selection approach, each of the statistical moments along with its respective bins 0, 1, 4, and 5 were predominantly selected.

In Lasso Regularization, bins 5 and 7 were majorly selected from the mean, standard deviation, and skewness of each color plane. In contrast, Ridge Regularization selected bins 0 and 7 of standard deviation, skewness, and kurtosis of green and blue planes. RFA selected 3 features, namely: gstd_bins0, gskew_bins0, and bmean_bins5, whereas RFA selected gskew_bins0, gstd_bins0 and gkurto_bins0. Finally, FSRS selected mean, skewness, and kurtosis majorly from bins 0, 4, and 5 of each of the three color planes. We can conclude that Bins 0, 4, 5, and 7 have prominent features predominantly extracted from green and blue planes more than red planes; however, the selected number is somewhat arbitrary. However, when we look at the features selected by Lasso, Ridge, and the other hybrid techniques such as RFA and RFE, green planes were chosen more than blue. Of these planes, *skewness* and *kurtosis* were the most favorably chosen moments. To sum up, the *green* channel of the malaria parasite images contributed to the prominent features as much as, if not more than, that of the *blue* channel. Fig. 8 (a) (b) (c) illustrate such a phenomenon deduced through our findings using a correlation matrix.

b) Lungs Image Dataset (Fig. 7 and Fig. 8):

The features selected by each classification model are as follows: It was generally perceived that Bins 2 and 7 were dominant across most moments set during the ROC-AUC performance metric, with Bins 0, 3, 4, and an occasional 6 used interchangeably amongst them. During Lasso Regularization, bin values of 0, 5, 6, 7 of the red plane, bins 0, 6, 7 of the green plane, and Bins 0, 3, 5 of the Blue plane were selected. Similarly, for Ridge Regularization, bins 3, 4, 6, and 7 were chosen dominantly across all three planes. RFE took moments of the mean and standard deviation of all three planes along with dominant bin values 4 and 5. RFA selects dominant bin values 0, 1, 3 from the planes, whereas FSRS selects Bins 5, 6, 7. Overall, the *red* and *green* plane features of the lung images are dominant in predicting the class labels, especially with the moments of *mean* and *standard deviation.*



(a) (b) (c)

Fig. 8. (a) (b) (c) Correlation graphs for Red plane, Green plane, and Blue Plane for Malaria Parasite Detection and Classification

From Fig. 8. (a) (b) (c), we can observe the correlation matrices. It can be discerned that there is a high correlation between two moments of the following degree: the mean has a high correlation with its standard deviation, and so on. As a result of such highly positive correlated values, either one of them or none of them is selected for feature selection. However, the correlation reduces from an odd moment to a next odd moment, that is, mean to skewness or standard deviation to kurtosis within the same bin value. As an example of a red plane, rskew_bins3 and rkurto_bins3 correlate 0.9996 as much as rskew_bins3 has with rstd_bins3 of 0.9992. However, rkurto_bins2 has a correlation of 0.5509 with rmean_bins2, as with varying bins such as rkurto_bins4 and rmean_bins5 of correlation coefficient 0.5032. Another pattern to note is the varying variance levels across the means of each channel compared to standard deviation, skewness, and kurtosis, which were relatively more stable. Features with higher variance are generally poor predictors for classification unless coupled with features having the same degree of variance. Therefore, results for precision and recall are presented only for the best classifier.

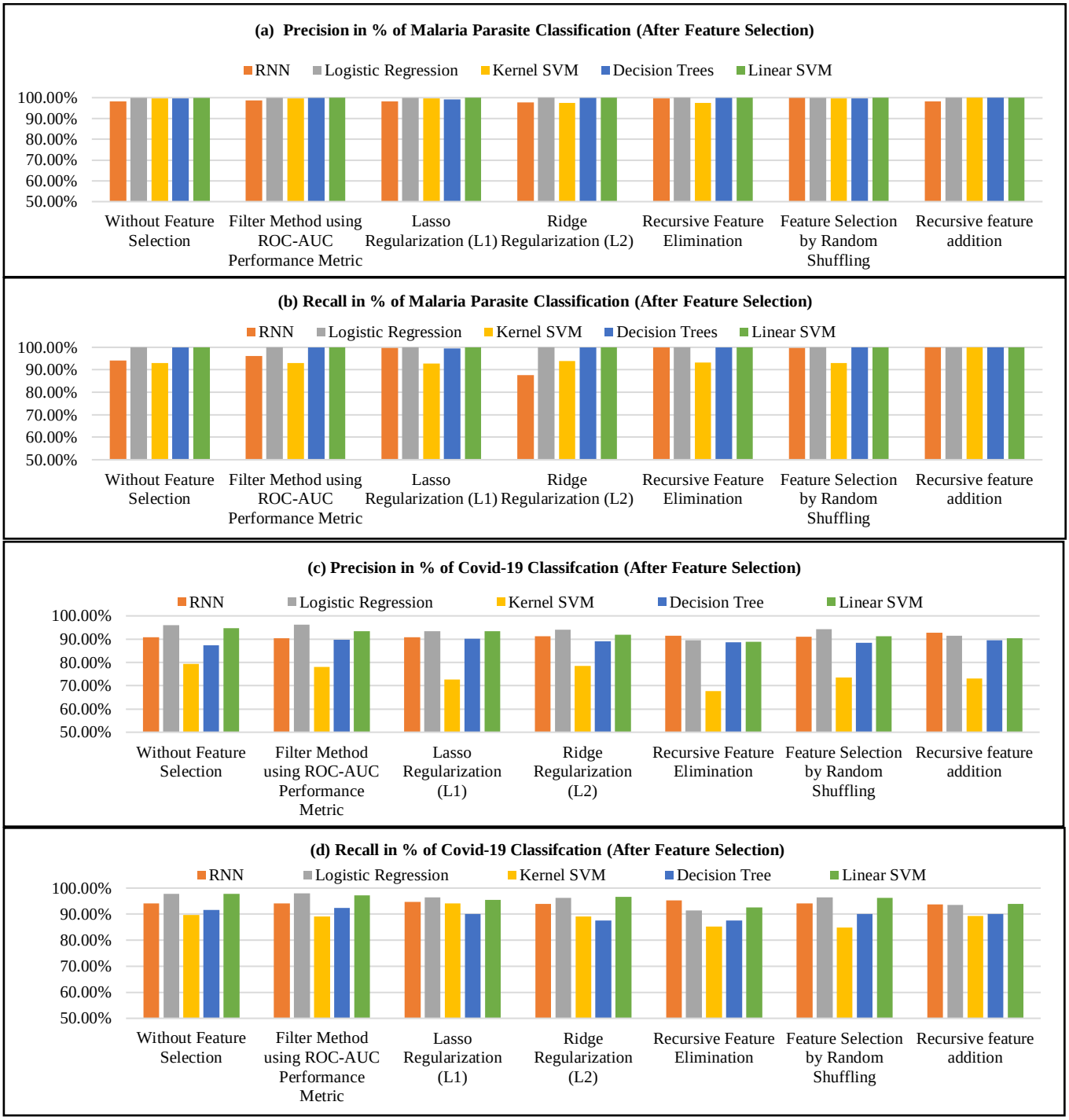


Fig. 9. Precision and Recall of Malaria Parasite Classification and COVID-19 classification (after Feature Selection)

With regards to the comparative study of the choice of feature selectors used, overall, KSVM can be the worst choice for COVID-19/Normal Image Classification, even if it is the moderately better one for Malaria Parasite Classification. We can also describe from Fig. 9 (a) that the Precision values are in the range of 95-100% for most of the classifiers, the maximum being LSVM at the upper mark. This shows that our system was able to identify blood smear images as parasitic for Malaria classification correctly. However, the same is not the case with the Recall values in Fig. 9 (b), whereby the true positives of the Malaria blood smear images are not being classified correctly for some models such as KSVM in some cases, RNN too.

Next, we can infer from Fig. 9 (c) that most of the feature selectors are not improving the precision of the model substantially; if anything at all, an iota of difference of barely 1%-2% can be observed in these graphs. This is incrementally followed in RNN and DT LR, and LSVM reduces in terms of precision values. Overall, the trends in Fig. 9 (c) show that our system was able to classify COVID-19 images correctly for COVID-19/Normal Image Classification. Additionally, in Fig. 9 (d), Recall values for KSVM work more modestly than the precision values, depicting that the classifier models can identify the relevant data with correct labels, but not as rigorously as Malaria

Image Classification. Recall values are also much better than the precision values of the same classification. The Filter method uses the ROC-AUC performance metric best, followed by FSRS, Lasso, and Ridge Regularization.

We could also perform vector contribution in best results by analyzing the malaria dataset feature vector databases and their respective KDE plots with further experimentation. The following Fig. 10 represents the KDE plot, typically done to estimate the probability distribution of a quantitative value across a typical scale for each of the three channels. These diagrams illustrate both the parasitized and uninfected cell plots in the same diagram for each moment about the respective color planes. First of all, the graphs containing no densities are not selected during feature selection. Secondly, the predominant features chosen by the most efficient classifiers as mentioned above are typically the value of the bin of 0 and 7, as they provide a more extended scale of feature value distributions. Thirdly, the primary reason behind such efficient classification of the parasite classification lies in the absence of bin values between Bins 1 to 6 for uninfected cell images against the parasitized ones; the distribution of parasitized cell images has a more significant distribution at Bin Values 1, 4 and 5. This makes the identification of color pixel values across respective color moments much more straightforward and more accessible.

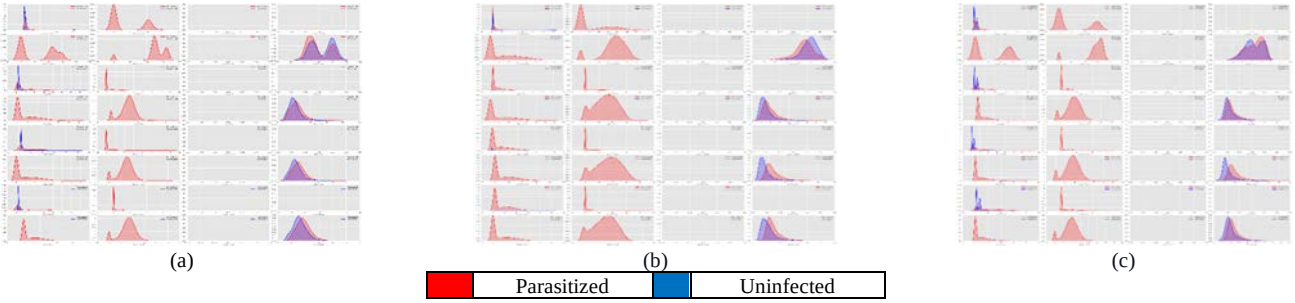


Fig. 10. (a) (b) (c) KDE Plots for each of the R, G, and B planes of Malaria Parasite Classification

6. Conclusion

In this research work, we predominantly focused on the intensity of each color pixel of an image by segregating them into respective bins based on the binary-coded values obtained (i.e., the equivalent of decimal integers Bins 0 to Bins 7) due to C.G. partitioning of R, G B histograms of an image into two parts. The novel idea proposed in our system is to evaluate whether this balanced segregation of image contents into these eight bins as feature components bring about a monumental impact on classification accuracy for two different diseases, i.e., Malaria and COVID-19 infections of our parasite system using appropriate feature engineering and selection techniques. The results and discussion concerning three objectives set for the proposed work highlight the following findings:

As per the first objective, we performed a comparative analysis of the results obtained from Malaria Parasite Classification using the Extended Bins Approach with statistical moments such as mean, standard deviation, skewness, and kurtosis; with this, we received an accuracy in the range of 95-99% and best value 99.9% for LSVM and LR classifiers. This proved to be much better than the results obtained using CNN (94%) and the GLCM, statistical moments of the entire image, fusion with bins approach (96.1%) simply by using the count of pixels as features in our previous paper [1].

The second objective was also to testify the performance of our proposed model to address the rising COVID-19 cases around the world with the help of the Lungs Radiography Image Database. Using these images as the input, we have pre-processed the gray-scale images (Lung X-ray images) into an RGB image during our experimentation by performing pseudo-color image processing. We performed transformation on each of the channels using trigonometric functions followed by the process of Histogram Equalization. Performance of bins with four moments on COVID-19 detection and classification has been proved with better outcomes in terms of accuracy in the range from 79.92% to 95.85% and 95.85% as the best result for LR classifier. For other parameters like precision and recall, we achieved 96.32% and 98.06%, respectively, for the LR classifier. Fulfillment of the third objective ensures the dimensionality reduction along with the best results. The adoption of feature selection processing techniques further reinforced the robustness of our model by choosing only those optimal combinations of feature input predictors that improved the performance of the model, thereby reducing its computational cost.

Detailed analysis of feature vector databases shown in Fig. 11 (a) (b) (c) with the feature predictors to infer their contribution with positive impact identified that bins 1 and 7 for malaria parasites and bins 2,3,5,6,7 for COVID/Normal Lung classification are found to be the predominant bins. The color component analysis states that green and blue color components are better in malaria and red and green for COVID-19 detection and classification. The study of statistical moments identifies that skewness and kurtosis for Malaria Parasite and mean, and standard deviation for COVID/Normal Lung Classification and Classification were most favorably chosen, leading to better results.

Finally, we can conclude that this comprehensive study with rigorous experimentation combines the spatial domain image processing with advanced machine learning techniques and is effective in two different medical domain

applications, i.e., for malaria parasite and COVID-19 detection and classification. Further, this work can be analyzed in all variants of COVID-19 with Normal and other image processing-based health care domains. Furthermore, we can perform a multi-class classification using other lung infected images such as Viral Pneumonia, etc.

References

[1] H. Telang and K. Sonawane, "Effective Performance of Bins Approach for Classification of Malaria Parasite using Machine Learning," 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2020, pp. 427-432, doi: 10.1109/ICCCA49541.2020.9250789.

[2] 2021. WHO Coronavirus (COVID-19) Dashboard. [online] Available at: <https://covid19.who.int/> [Accessed 9 March 2021].

[3] H. B. Kekre, Kavita Sonawane, "Performance Evaluation of Bins Approach in YCbCr Color Space with and without Scaling", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-3, Issue-3, July 2013.

[4] H. B. Kekre, Kavita Sonawane, "Performance of Histogram Modification by LOG Function for CBIR using Statistical Parameters of Bins Contents", International Journal of Electronics Communication and Computer Engineering, Volume 3, Issue 6, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

[5] H. B. Kekre, Kavita Sonawane, "Image Retrieval Using Histogram Based Bins of Pixel Counts and Average of Intensities", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 10, No. 1, 2012

[6] H. B. Kekre, Kavita Sonawane, "Effect of Similarity Measures for CBIR using Bins Approach", International Journal of Image Processing, 2012H. B. Kekre, Kavita Sonawane.

[7] H. B. Kekre, Kavita Sonawane, "Histogram Bins Matching Approach for CBIR Based on Linear grouping for Dimensionality Reduction", I.J. Image, Graphics and Signal Processing, pp 68-82, 2014

[8] M. Alva and K. Sonawane, "Hybrid Feature Vector Generation for Alzheimer's Disease Diagnosis Using MRI Images," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-6, doi: 10.1109/I2CT45611.2019.9033826.

[9] M. Alva, A. Srinivasaraghavan and K. Sonawane, "A Review on Techniques for Ear Biometrics," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2019, pp. 1-6, doi: 10.1109/ICECCT.2019.8869450.

[10] F. K. Nezhadian and S. Rashidi, "Melanoma skin cancer detection using color and new texture features," 2017 Artificial Intelligence and Signal Processing Conference (AISP), Shiraz, Iran, 2017, pp. 1-5, doi: 10.1109/AISP.2017.8324108.

[11] A. Olugboja and Z. Wang, "Malaria parasite detection using different machine learning classifier," 2017 International Conference on Machine Learning and Cybernetics (ICMLC), Ningbo, China, 2017, pp. 246-250, doi: 10.1109/ICMLC.2017.8107772.

[12] Naveen, R. K. Sharma and A. Ramachandran Nair, "Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models," 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, 2019, pp. 100-104, doi: 10.1109/RTEICT46194.2019.9016968.

[13] Z. Cai, Z. Yu, H. Zhou and Z. Gu, "The Early Stage Lung Cancer Prognosis Prediction Model based on Support Vector Machine," 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), Shanghai, China, 2018, pp. 1-4, doi: 10.1109/ICDSP.2018.8631657.

[14] D. Krishnani, A. Kumari, A. Dewangan, A. Singh and N. S. Naik, "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 367-372, doi: 10.1109/TENCON.2019.8929434.

[15] E. Celik and S. I. Omurca, "Improving Parkinson's Disease Diagnosis with Machine Learning Methods," 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 2019, pp. 1-4, doi: 10.1109/EBBT.2019.8742057.

[16] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 2016, pp. 18-20, doi: 10.1109/ICACA.2016.7887916.

[17] K. R. Pushpalatha and A. G. Karegowda, "CFS Based Feature Subset Selection for Enhancing Classification of Similar Looking Food Grains- A Filter Approach," 2017 2nd International Conference On Emerging Computation and Information Technologies (ICECIT), Tumakuru, India, 2017, pp. 1-6, doi: 10.1109/ICECIT.2017.8453403.

[18] X. Zeng, Y. -W. Chen and C. Tao, "Feature Selection Using Recursive Feature Elimination for Handwritten Digit Recognition," 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kyoto, Japan, 2009, pp. 1205-1208, doi: 10.1109/IIH-MSP.2009.145.

[19] Z. KARHAN and F. AKAL, "Covid-19 Classification Using Deep Learning in Chest X-Ray Images," 2020 Medical Technologies Congress (TIPTEKNO), Antalya, Turkey, 2020, pp. 1-4, doi: 10.1109/TIPTEKNO50054.2020.9299315.

[20] Rahul Kumar, Ridhi Arora, Vipul Bansal, Vinodh J Sahayasheela, Himanshu Buckchash, Javed Imran, Narayanan Narayanan, Ganesh N Pandian, and Balasubramanian Raman, "AI-based Diagnosis of COVID-19 Patients Using X-ray Scans with Stochastic Ensemble of CNNs", TechRxiv, 2020.

[21] S. D. Thepade, P. R. Chaudhari, M. R. Dindorkar and S. V. Bang, "Covid19 Identification using Machine Learning Classifiers with Histogram of Luminance Chroma Features of Chest X-ray images," 2020 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2020, pp. 36-41, doi: 10.1109/IBSSC51096.2020.9332160.

[22] Nurrahma and R. Yusuf, "Comparing Different Supervised Machine Learning Accuracy on Analyzing COVID-19 Data using ANOVA Test," 2020 6th International Conference on Interactive Digital Media (ICIDM), Bandung, Indonesia, 2020, pp. 1-6, doi: 10.1109/ICIDM51048.2020.9339676.

[23] H. B. Kekre and K. Sonawane, "Use of equalized histogram C.G. on statistical parameters in bins approach for CBIR," 2013 International Conference on Advances in Technology and Engineering (ICATE), 2013, pp. 1-6, doi: 10.1109/ICAdTE.2013.6524727.

[24] T. LI and H. ZHU, "Research on Color Algorithm of Gray Image Based on a Color Channel," 2020 Chinese Control And Decision Conference (CCDC), 2020, pp. 3747-3752, doi: 10.1109/CCDC49329.2020.9164375.
[25] Selvapriya, B. & Raghu, B. (2018), "A color map for pseudo color processing of medical images," International Journal of Engineering and Technology (UAE), 7. 954-958.
[26] Zare, Mohammad & Jampour, Mahdi & Farrokhi, Issa. (2011). "A heuristic method for gray images pseudo coloring with histogram and RGB layers". 10.1109/ICCSN.2011.6014949.
[27] Govind Haldankar, Atul Tikare and Jayprabha Patil, "Converting Gray-Scale Image to Color Image" Proceedings of SPIT-IEEE Colloquium and International Conference, Vol. 1, pp. 189-192, Mumbai, India.
[28] Ahmed Hassan Mohammed Hassan, Arfan Ali Mohammed Qasem, Walaa Faisal Mohammed Abdalla, Omer H. Elhassan, "Visualization & Prediction of COVID-19 Future Outbreak by Using Machine Learning", International Journal of Information Technology and Computer Science, Vol.13, No.3, pp.16-32, 2021.
[29] Hanan A. Al-Jubouri, "Integration Colour and Texture Features for Content-based Image Retrieval", International Journal of Modern Education and Computer Science, Vol.12, No.2, pp. 10-18, 2020.

Authors' Profiles

Mr. Hrishikesh Telang is a student who recently received his B.E. (First Class with Distinction) in Computer Science from Mumbai University (2020). He is highly consistent in his academic track record and is well revered amongst his faculty from the Department of Computer Science. From 2018 to 2019, Hrishikesh was a Research Intern at the R&D cell of his college, pursuing a Water Conservation project from the Internet of Things (IoT). Later, he pursued an internship in Data Science and Business Analytics between August 2020 to September 2020. Since December 2020, Hrishikesh has been working as a Research Assistant under the tutelage of Dr. Kavita Sonawane towards projects pertaining to medical science. His research interests include Image Processing, Machine Learning, Deep Learning, Natural Language Processing, Recommendation Systems, Data Analytics, Data Visualization, Data Mining and Warehousing, and Cloud Data Management and has actively pursued various self-projects to further his knowledge and understand core concepts. Hrishikesh is a prospective M.S. student in Information Management with a Certificate in Advanced Studies (C.A.S) in Data Science at the Syracuse School of Information Studies (2023).

Dr. Kavita Sonawane has received her Bachelor of Computer Engineering from Nagpur University in 2001 and Masters of Computer Engineering from University of Mumbai in 2008. She also received her Ph.D. degree in Computer engineering from NMIMS University, Mumbai, Maharashtra in 2015. From 2008 to 2009, she was a Research Assistant with Dr. H. B. Kekre (Professor of IIT-B). She has total 20 years of teaching experience in Computer Engineering Department. Currently, she is working as Professor and Head of Computer Engineering Department, at SFIT affiliated to the University of Mumbai, Maharashtra, India. Her research interests include Image Processing, Image analysis and retrieval, Medical Image Processing, Computer Vision, Machine Learning and Data analytics. She is an approved Ph. D guide at the University of Mumbai. She has more than 50 research papers published in various Peer reviewed Scopus/SCI indexed international Conferences and Journals to her credit.

How to cite this paper: Hrishikesh Telang, Kavita Sonawane, "COVID-19 and Malaria Parasite Detection and Classification by Bins Approach with Statistical Moments Using Machine Learning", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.15, No.3, pp. 1-13, 2023. DOI:10.5815/ijigsp.2023.03.01