

Improving Classification Accuracy Using Hybrid Machine Learning Algorithm on Malaria Dataset[†]

Rashke Jahan * and Shahzad Alam

Department of Computer Engineering, Jamia Millia Islamia, New Delhi; email: salam7@jmi.ac.in

* Correspondence: rashkecs57@gmail.com

† Presented at the 4th International Electronic Conference on Applied Sciences, 27 October–10 November 2023; Available online: <https://asec2023.sciforum.net/>.

Abstract: Machine Learning Algorithms are integrated into Computer-Aided Design (CAD) methodologies to support medical practitioners in diagnosing patient disorders. This research seeks to enhance the accuracy of classifying malaria-infected erythrocytes (RBCs) through the fusion of Machine Learning Algorithms, resulting in a hybrid classifier. The primary phases involve data pre-processing, segmentation, feature extraction, and RBC classification. This paper introduces a novel hybrid Machine Learning Algorithm, employing two combinations of supervised algorithms. The initial combination encompasses Stochastic Gradient Descent (SGD), Logistic Regression, and Decision Tree, while the second employs Stochastic Gradient Descent (SGD), Xgboost, and Random Forest. The proposed approach, implemented using Python programming, presents an innovative hybrid Machine Learning Algorithm. Through a comparative analysis between individual algorithms and the proposed hybrid algorithm, the paper demonstrates heightened accuracy in classifying malaria data, thus aiding medical practitioners in diagnosis. Among these algorithms, SGD, Logistic Regression, and Decision Tree yield individual accuracy rates of 90.63%, 92.23%, and 93.43% respectively, while the hybrid algorithm achieves 95.64% accuracy on the same dataset. The second hybrid algorithm, combining SGD, Xgboost, and Random Forest, outperforms the initial hybrid version. Individually, these algorithms achieve accuracy rates of 90.63%, 95.86%, and 96.11%. When the proposed hybrid algorithm is applied to the same dataset, accuracy is further enhanced to 96.22%.

Keywords: Content based image retrieval (CBIR); Malaria; Image Processing; Decision Tree Algorithm; SGD; Logistic Regression; Voting Classifier; Adaboost; Xgboost, Random Forest

Citation: Jahan, R.; Alam, S., Improving Classification Accuracy Using Hybrid Machine Learning Algorithm on Malaria Dataset. *Eng. Proc.* **2023**, *52*, x.

<https://doi.org/10.3390/xxxxx>

Academic Editor: Nunzio Cennamo

Published: date



Copyright: © 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Malaria is a severe illness triggered through a Plasmodium genus blood parasite and becomes a major cause of death in the globe [1]. Malaria is caused by female mosquitos Anopheles that directly infects the Red Blood Cells. The parasite enters the blood through the mosquito's saliva. The RBC in the blood is infected with the various kinds of malaria parasites. Malaria is caused by Plasmodium species parasites that spread through infected mosquito bites from individual to individual. It can also be transferred by transfusions of blood. The incubation period is eight to twelve days for the most severe type of malaria. It may be as lengthy as ten months in some rare types of malaria. Under a light microscope, standard malaria detection based on manual microscopic consideration will generally improve the opportunity of incorrect identification and late diagnosis. As an outcome, many scientists have suggested automated malaria diagnosis based on image processing strategy to give timely malaria parasite detection as well as enhancing malaria diagnosis accuracy. [2] proposed a methodology for extracting cell features that involves the use of K-means clustering to segment cell nuclei. This strategy yielded a segmentation error rate of only 6.46%. Author [3] established the Sobel Edge Detection approach as a means of differentiating the borders of the malaria parasite from the surrounding blood

corpuscles. Researcher [4] proposed a method for differentiating parasite components from white blood cells (WBCs) by employing adaptive thresholding on the histogram of the V-value in the HSV colour space. A total of 20 test photos were utilised, yielding an accuracy percentage of 60%. [5] employed adaptive histogram thresholding on slide pictures to morphologically detect red blood cells (RBCs) that are infected with malaria parasites. The authors [6] employed a combination of geomorphologic segmentation and color-specific information in order to identify parasites among densely populated blood smears. The researchers performed tests on a dataset consisting of 75 photos in order to evaluate the effectiveness of their methodology at the patch level.

In the latest decades, Malaria parasite detection has been one of the significant interesting research field because of the several proposed new automated detection approaches. In order to design an automatic detection scheme, the parasite of malaria direct infects the red blood cells(RBC), thus separating the red blood cells(RBC) from the artefacts and background in a microscopic image [7]. An author proposed that the Breast cancer classification has been suggested in a new system. A hybrid of Support Vector Machine (SVM) and K-Means are used by the current system [8]. The experimental outcome showed how the suggested algorithm is effective. [9] concluded that the hybrids algorithm's performance level (Decision tree and Artificial neural network) is better than that of the performance of the algorithms individually. Learning-based techniques are more effective for larger datasets. In this paper, we propose a novel method using Python programming, presents an innovative hybrid Machine Learning Algorithm, through a comparative analysis between individual algorithms and the proposed hybrid algorithm. This paper demonstrates heightened accuracy in classifying malaria data, thus aiding medical practitioners in diagnosis.

2. Methodology

By using the hybridization technique, the efficiency of individual algorithms can be improved. These are the steps, which is followed by the suggested scheme are image acquisition, preprocessing of images, segmentation of images, extraction of features, and the classification of images using five distinct kinds of machine learning algorithms i.e. Stochastic Gradient Descent (SGD), Decision Tree, Logistic Regression, Xgboost, and Random Forest.

Hybrid classifier's main objective is to increase accuracy of classification compared to other individual classifiers. In the hybrid classifier, the misclassifications of every individual classifier are often overcome, thus increasing the accuracy rate. The classifiers are combined for the ending classification of infected and uninfected erythrocytes(RBC) using the majority voting technique. It is an important and novel area of research as compared to discrete learning approaches.

2.1. Data acquisition

For the detection of malaria parasites, the data is collected from National Institute of Health(NIH). For machine learning, the dataset source for the suggested algorithm is available on kaggle.com. This dataset was created by NIH that has a 27,558 cell image malaria dataset with the same number of parasitized and uninfected cells. To train and test our model, this is an excellent dataset of labeled image. Training set of 24802 images and testing set of 2756 images have been created by dividing the Malaria cell data set in the ratio of 9:1 for the training and validation of the proposed. To detect and segment red blood cells, a level-set based algorithm was implemented.

The entire dataset used consists of images of normal Red Blood Cells smear sample as shown in figure-1 and the infected Red Blood Cell as shown in figure-2.

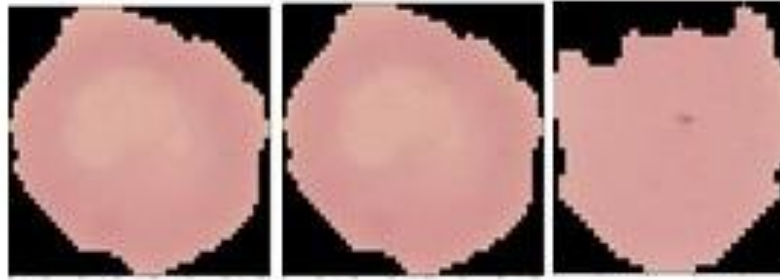


Figure 1. Normal Red Blood Cells microscopic images.

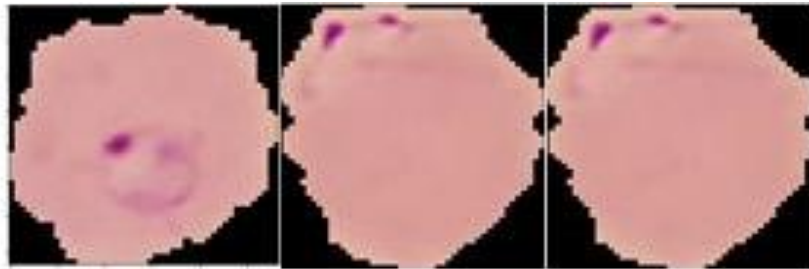


Figure 2. Malaria parasite infected Red Blood Cells microscopic images.

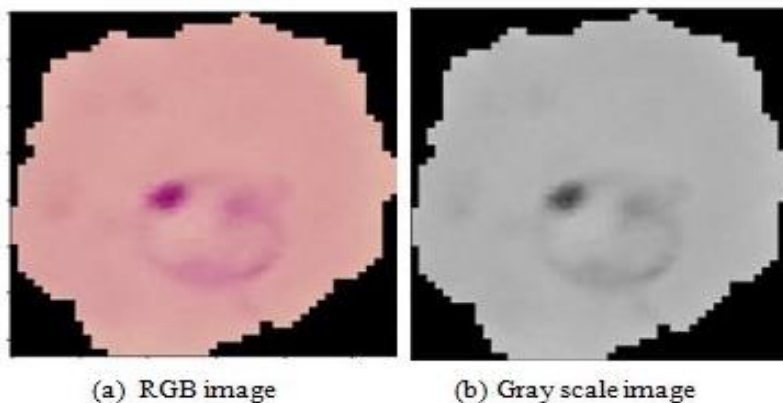
2.2. Preprocessing

Preprocessing is primarily implemented to enhance image quality and decrease picture differences that would unnecessarily complicate subsequent processing steps. First, the input images are resized to 256 / 256 images in this module. This re-dimensioned image is then transformed into a gray image. Then, by applying suitable contrast stretching boundaries, the grayscale is applied to make the infected cell or parasites darker and the unwanted background lighter. Contrast stretching methods and, in specific, histogram equalization are the most popular for enhancement. Color normalization methods, including the common use of grayscale colors, have been applied for illumination and staining differences.

2.3. RGB to Gray-scale Conversion

The transformation from an RGB image to a gray image includes easy matrix manipulation and is taken to decrease the quantity of learning and detection information involved. The RGB colour model is an additional colour model that adds different types of red, green and blue light.

In the RGB colour system, a colour is described by displaying the amount included of each red, yellow, and blue. The values of its red, blue and green parts must first be obtained to transform any RGB image into a gray-scale image of its luminance (fig. 3).



(a) RGB image

(b) Gray scale image

Figure 3. RGB and Gray scale image.

2.4. Feature Extraction

Extraction of feature is the method of identifying and classifying some of the specific points on an image [10].

2.4.1. Texture Based feature

Texture-based characteristics help identify the distinct Haralick textures of the infected and uninfected Red Blood Cell While determine the texture characteristics, [11] the matrix-based elements are determine using the following formulas.

$$\text{Contrast}(c) = \sum_{i,j} |i - j|^2 p(i,j) \tag{1}$$

$$\text{Correlation}(C) = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j} \tag{2}$$

$$\text{Energy}(E) = \sum_{i,j} p(i,j)^2 \tag{3}$$

$$\text{Homogeneity} = \sum_i \sum_j \frac{1}{1+(i-j)} p_d(i,j) \tag{4}$$

$$\text{ASM} = \sum_i \sum_j p_d^2(i,j) \tag{5}$$

2.4.2. Statistical Feature

Statistical values of the image are also called as the second order of the histogram. Statistical features of Haralick are determined using the following formulas.

$$\text{Mean: } \mu = \sum_{k=0}^{G-1} kp(k) \tag{6}$$

$$\text{Varian: } \sigma^2 = \sum_{k=0}^{G-1} (k - \mu)^2 p(k) \tag{7}$$

$$\text{Kurtosis} = \sigma^{-4} \sum_{k=0}^{G-1} (k - \mu)^4 p(k) - 3 \tag{8}$$

2.4.3. Hu Moment shape based features

In 1962, Hu-Moment proposed invariant to rotation, scaling and translation (RTS) invariant seven linked area features and also known as Algebraic Moment Invariants. Moments invariants are the area descriptor which is used to characterize the shape and size of an image. Moments are used to describe the properties of an object with respect to the area, position and orientation.

For an N×N images m moment of the image $f(x,y)$ is
 $m(i,j) = \sum_x x^i \sum_y y^j f(x,y)$

2.5. Classification

This is the final and important phase where the input image characteristics are categorized and compared to the characteristics of the infected Red Blood Cells image [12]. In this research, five types of Machine Learning Algorithm are used as a classifier ie. SGD, Random Forest, Decision Tree, Logistic Regression and Xgboost. The classification accuracies of Decision Tree, SGD, Logistic Regression, Xgboost, Random Forest and Hybrid of these three classifiers were compared. Compared to the other classifiers, the hybrid classifiers' accuracy is high.

Based on the outcomes of the classification acquired in the job where diagnostic accuracy is calculated using the following technique-

$$\text{Accuracy}(A) = \frac{(TP) + (TN)}{(TP) + (TN) + (FP) + (FN)} \quad (9)$$

$$\text{Recall}(R) = \frac{(TP)}{(TP) + (FN)} \quad (10)$$

$$\text{Precision}(P) = \frac{(TP)}{(TP) + (FP)} \quad (11)$$

2.6. Proposed system

The design of the suggested hybrid method is based on the classification of infected and uninfected erythrocytes using the microscopic images of thin smears in the blood. The suggested method's general block diagram is shown in Figure. The model consists of a microscopic image of six stages of thin blood smear.

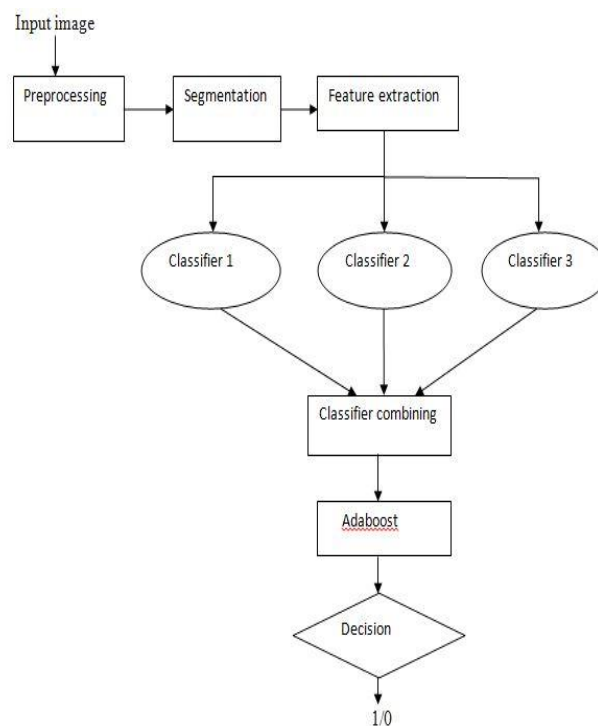


Figure 4. Pre-processing, segmentation, feature extraction, and endmost discrimination of infected and uninfected erythrocyte using hybrid classifier.

2.7. Hybrid classifier (SGD, Decision tree and Logistic Regression)

The method used to implement the hybrid algorithm using SGD, Decision tree and Logistic Regression is introduced, the individual is implemented whose output will give less accuracy to predict the result of the ensemble of the classifiers. The hybrid of classifier is the mixing of the determination of three classifiers to acquire the final classification outcome. The hybrid classifier's primary objective is to increase the accuracy of classification compare to separate classifiers. The misclassifications of every individual classifier are frequently overcome in the hybrid of classifier, thereby improving the rate of accuracy. The classifiers are coupled using the majority voting method for the final classification of infected and uninfected Red Blood Cells. Boosting techniques enhance the accuracy of any supervised algorithm for machine learning.

2.8. Hybrid classifier (SGD, Xgboost and Random Forest)

The approach used for implementing the hybrid algorithm using SGD, Xgboost and Random Forest, the individual is implemented, but accuracy to predict the result of the

ensemble of classifiers is more than the individual. For the final classification of infected and uninfected Red Blood C, the classifiers are therefore paired using majority voting techniques. And then the boosting method improves the accuracy of hybrid machine learning algorithm.

2.9. Classification based on hybrid classifier

Hybrid classifier pairs various classifiers' choice to achieve the final outcome of classification. Hybrid classifier's primary objective is to enhance classification accuracy. In the hybrid classifier, misclassifications of individual classifier are often overcome, thus improving the precision rate.

3. Results

This study set out to enhance the accuracy of the hybrid machine learning algorithm. Hence, this section shows the analysis, results, and discussion about individual and proposed hybrid machine learning classifiers' accuracies on selected malaria dataset. Here we have selected two combinations of supervised machine learning algorithms, the first is SGD, Logistic Regression, and Decision Tree algorithm and the second is SGD, Xgboost and Random Forest algorithm. The following results shown in the table are the outcomes of both ways of implementing our proposed hybrid. Using the combination of SGD, Xgboost and Random Forest average probabilities combination rule of hybrid is done, as it gives better performance on selected datasets.

Table 1. Results of hybrid of SGD, Logistic Regression and Decision Tree algorithm.

	Adb	Evc	Decision tree	Logistic Regression	Stochastic Gradient Descent
Accuracy	0.956459	0.94303	0.93541	0.912192	0.906386
Recall	0.969547	0.96458	0.94688	0.951841	0.946884
Precision	0.946750	0.92716	0.92847	0.885375	0.879605

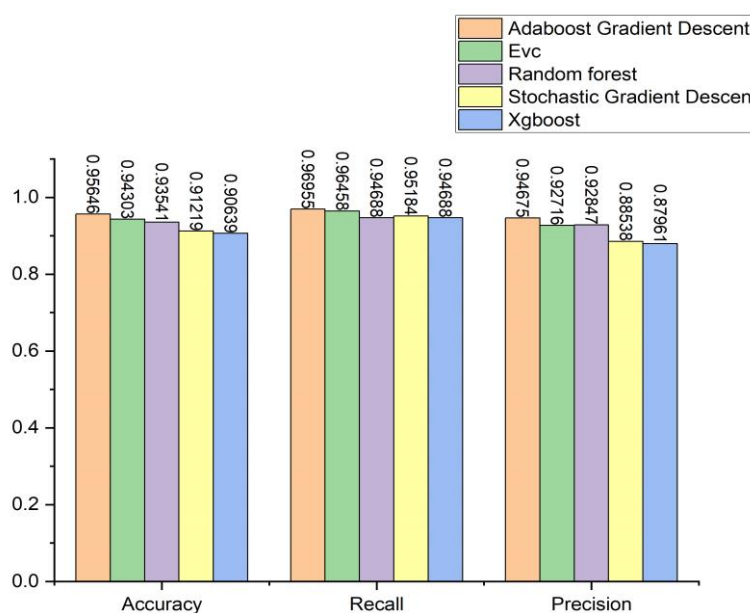


Figure 5. Comparison of individual classifier in terms of accuracy, precision and recall.

Table 2. Results of ensemble of Random Forest, Xgboost and SGD algorithm.

Adaboost	Evc	Random forest	Stochastic Gradient Descent	Xgboost
----------	-----	---------------	-----------------------------	---------

Accuracy	0.962264	0.95863	0.96117	0.906386	0.95863
Recall	0.977337	0.98087	0.97662	0.946884	0.97875
Precision	0.950413	0.94089	0.94907	0.879605	0.94270

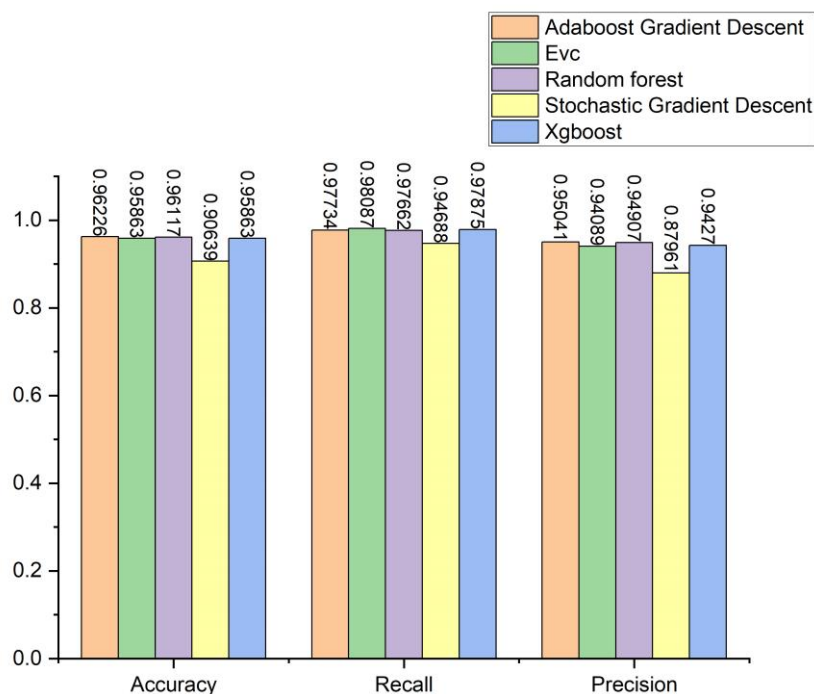


Figure 6. Comparison of individual classifier in terms of accuracy, precision and recall.

3.1. Comparative Analysis

The classification accuracies of Decision Tree, SGD, Logistic Regression, Xgboost, Random Forest and Hybrid classifiers were compared. The accuracy of the Hybrid classifier was the highest as compared to the other classifiers.

Table 3. Comparison of classification accuracy.

Sl No.	Model	Classification accuracy with malaria cell image dataset
01.	SGD	90.63%
02.	Logistic Regression	91.21%
03.	Decision Tree	93.54%
04.	Xgboost	95.86%
05.	Random Forest	96.11%
06.	Hybrid of SGD, Logistic Regression and Decision Tree	95.64%
07.	Hybrid of SGD, Xgboost and Random Forest	96.22%

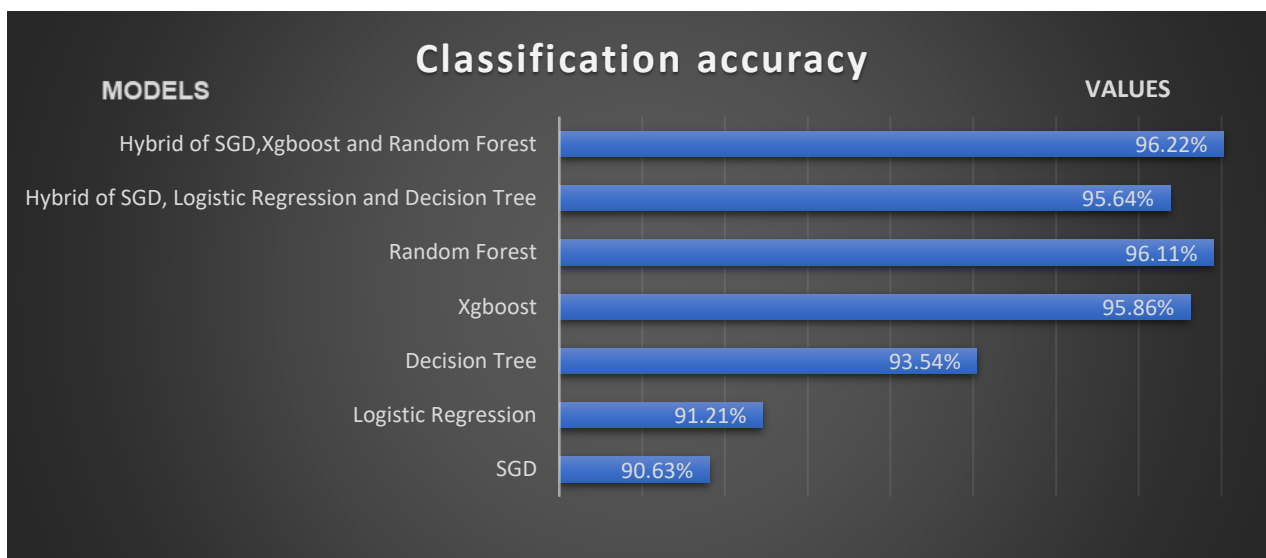


Figure 7. Performance Metrics of Models.

4. Conclusion

In this study, the proposed ensemble and hybrid algorithm demonstrate that hybrid machine-learning techniques perform better than the individual algorithms on the selected medical dataset of Malaria Cell. In our proposed system, there are two combinations of classifiers. First, the individual classifiers SGD, Logistic regression, and Decision Tree are combined and the second combination is of SGD, Random Forest and Xgboost using the majority voting technique for the final classification of infected and uninfected Red Blood Cells. These three algorithms SGD, Logistic Regression, and Decision tree detect the parasites individually and their accuracies are 90.63%, 92.23%, and 93.43%. The proposed hybrid algorithm is applied to the same dataset and then the accuracy is increased to 95.64%. The second proposed hybrid algorithm using combinations of SGD, Xgboost and Random Forest outperform over the proposed hybrid algorithm. They also detect the parasites individually and their accuracies are 90.63%, 95.86%, and 96.11%. And again the proposed hybrid algorithm applied on the same data set and the accuracy is increased to 96.22%. The result shows that the hybrid machine-learning algorithm is the key to improve classification accuracy. And the performance of the combination of second hybrid algorithms is better than the combination of the first hybrid algorithm.

Supplementary Materials: Not Applicable

Author Contributions: Conceptualization, methodology and editing by Shahzad Alam, and software, validation, data curation, writing—original draft preparation by Rashke Jahan.

Funding: Not Applicable

Institutional Review Board Statement: Not Applicable

Data Availability Statement: Not Applicable

Acknowledgments: Not Applicable

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kaur J. Malarial positive image retrieval using Content Based Retrieval Systems. :37–41.
2. Sarrafzadeh, Omid, & Alireza, M. D. (2015). Nucleus and cytoplasm segmentation in microscopic images using K-means clustering and region growing. *Advanced Biomedical Research*, 4. PMID:26605213

3. P. Rakshit and K. Bhowmik, "Detection of presence of parasites in human RBC in case of diagnosing malaria using image processing," 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), Shimla, India, 2013, pp. 329-334, doi: 10.1109/ICIIP.2013.6707610.
4. Kaewkamnerd S, Uthaiyibull C, Intarapanich A, Pannarut M, Chaotheing S, Tongshima S. An automatic device for detection and classification of malaria parasite species in thick blood film. BMC Bioinformatics. 2012
5. I. R. Dave and K. P. Upla, "Computer aided diagnosis of Malaria disease for thin and thick blood smear microscopic images," 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2017, pp. 561-565, doi: 10.1109/SPIN.2017.8050013.
6. Chakrabortya et al., J Health Med Informat 2015, 6:1 <http://dx.doi.org/10.4172/2157-7420.1000179>
7. Imran Razzak M, Informatics H. Automatic Detection and Classification of Malarial Parasite. Int J Biometrics Bioinforma. 2015;(9):1.
8. Ijaz M, Alfian G, Syafrudin M, Rhee J. Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest. Appl Sci. 2018;8(8):1325.
9. V AA, Asha A. Improving Prediction Accuracy Using Hybrid Machine Learning Algorithm on Medical Datasets. Ijser. 2018;9(10):1461-7.
10. Charpe K, Bairagi VK, Desarda S, Barshikar S. A Novel Method for Automatic Detection of Malaria Parasite Stage in Microscopic Blood Image. Int J Comput Appl. 2015;128(17):32-7.
11. V.K. B, K.C. C. Comparison of texture features used for classification of life stages of malaria parasite. Int J Biomed Imaging. 2016;2016.
12. Devi SS, Roy A, Singha J, Sheikh SA, Laskar RH. Malaria infected erythrocyte classification based on a hybrid classifier using microscopic images of thin blood smear. Multimed Tools Appl. 2018;77(1):631-60.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.