engineering proceedings

MDPI

Proceeding Paper

Improving Classification Accuracy Using Hybrid Machine Learning Algorithms on Malaria Dataset †

Rashke Jahan * and Shahzad Alam

Department of Computer Engineering, Jamia Millia Islamia, New Delhi 110025, India; 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 preprocessing, 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 Algorithms on Malaria Dataset. *Eng. Proc.* **2023**, *56*, 232. https://doi.org/10.3390/ASEC2023-15924

Academic Editor: Alessandro Bruno

Published: 8 November 2023

1. Introduction

Malaria is a severe illness triggered via a Plasmodium genus blood parasite and becomes a major cause of death in the world [1]. Malaria is caused by female Anopheles mosquitos that directly infect the red blood cells. The parasite enters the blood through the mosquito's saliva. The RBC in the blood is infected with various kinds of malaria parasites. Malaria is caused by parasites belonging to the Plasmodium species and is spread via 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 types 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 observations will generally provide an opportunity to remedy incorrect identifications and late diagnoses. As a result, many scientists have suggested an automated malaria diagnosis based on an image processing strategy to give timely malaria parasite detection as well as to enhance malaria diagnosis accuracy. Ref. [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%. In [3], the author established the Sobel edge detection approach as a means of differentiating the borders of the malaria parasite

from the surrounding blood corpuscles. In [4], the researcher 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 color space. A total of 20 test photos were utilized, yielding an accuracy percentage of 60%. Ref. [5] employed adaptive histogram thresholding on slide pictures to morphologically detect red blood cells (RBCs) that are infected with malaria parasites. The authors of [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 recent decades, malaria parasite detection has been one of the most significant interesting research fields because of the several proposed new automated detection approaches. In order to design an automatic detection scheme, it should be acknowledged that the parasite of malaria directly infects the red blood cells (RBCs), thus separating the red blood cells (RBCs) from the artefacts and background in a microscopic image [7]. Another study has proposed the use of a breast cancer classification system as a new system to classify mammograms as normal or abnormal. A hybrid of support vector machines (SVM) and k-means are used by the current system [8]. The experimental outcome showed how the suggested algorithm is effective. Ref. [9] concluded that the hybrid 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, presenting an innovative hybrid machine learning algorithm and making 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. The steps for this technique are the following: image acquisition, the preprocessing of images, the segmentation of images, the 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.

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

2.1. Data Acquisition

For the detection of malaria parasites, the data are collected from the 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 the NIH which has a malaria dataset containing 27,558 cell images with the same number of images of parasitized and uninfected cells. To train and test our model, this is an excellent dataset of labeled images. A training set of 24,802 images and a testing set of 2756 images have been created by dividing the malaria cell dataset in a ratio of 9:1 for the training and validation of the proposed hybrid. 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 cell smear samples as shown in Figure 1 and images of the infected red blood cell smear samples as shown in Figure 2.

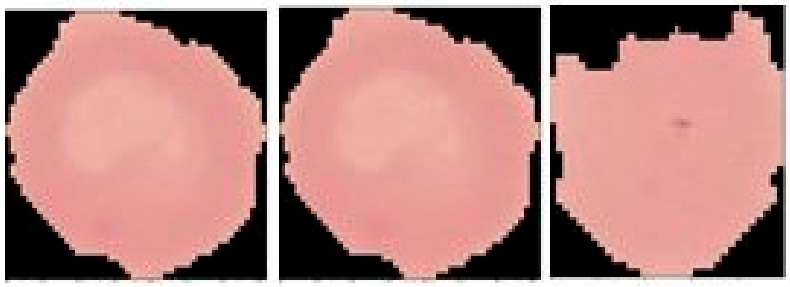

Figure 1. Microscopic images of normal red blood cells.

Figure 2. Microscopic images of malaria parasite-infected red blood cells.

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 methods for enhancement. Color normalization methods, including the common use of grayscale colors, have been applied for illumination and staining differences.

2.3. RGB to Grayscale 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 color model is an additional color model that adds different types of red, green and blue light.

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

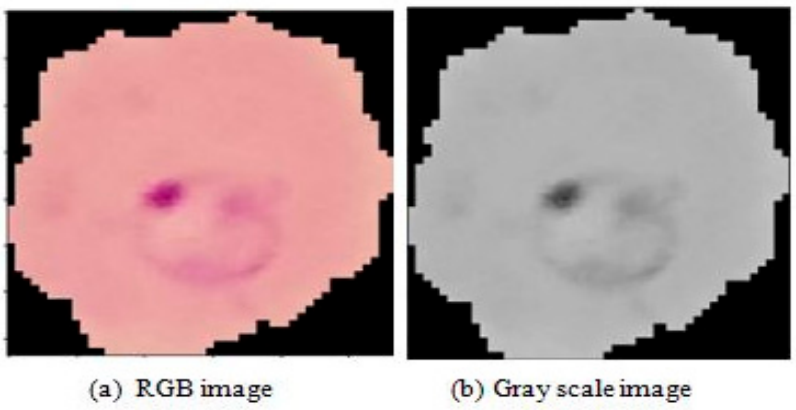

Figure 3. RGB and gray scale image.

2.4. Feature Extraction

Feature extraction is a method of identifying and classifying some of the specific points on an image [10].

2.4.1. Texture-Based Features

Texture-based characteristics help identify the distinct Haralick textures of the infected and uninfected red blood cells while determining the texture characteristics [11]. The matrix-based elements are determined using the following formulas:

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

$$\text{Correlation (C)} = \sum_{i,j} \frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma i \sigma j} \tag{2}$$

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

$$\text{Homogeneity} = \sum_i \sum_j \frac{1}{1+(i-j)^2} p_d(i,j) \tag{4}$$

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

2.4.2. Statistical Feature

Statistical values of the image are also known as second-order histograms. Statistical features of Haralick are determined using the following formulas:

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

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

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

2.4.3. Hu Moment Shape-Based Features

In 1962, Hu proposed moment invariants to be rotation, scaling, and translation (RTS) invariants with seven linked area features, 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, the moment of the image $f(x, y)$ is

$$\text{m}(i,j) = \sum_x x^i \sum_y y^j f(x,y) \tag{9}$$

2.5. Classification

This is the final and most 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 algorithms are used as a classifier, i.e., SGD, random forest, decision tree, logistic regression and Xgboost. The classification accuracies of decision tree, SGD, logistic regression, Xgboost, random forest and a hybrid of these three classifiers were compared. Compared to the other classifiers, the hybrid classifier's accuracy is high.

Based on the outcomes of the classification acquired in this task, the diagnostic accuracy is calculated using the following technique:

$$\text{Accuracy (A)} = \frac{(\text{TP})+(\text{TN})}{(\text{TP})+(\text{TN})+(\text{FP})+(\text{FN})} \tag{10}$$

$$\text{Recall (R)} = \frac{(\text{TP})}{(\text{TP}) + (\text{FN})} \tag{11}$$

$$\text{Precision (P)} = \frac{(\text{TP})}{(\text{TP}) + (\text{FP})} \tag{12}$$

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 of the blood. The suggested method's general block diagram is shown in Figure 4. The model consists of a microscopic image of six stages of thin blood smears.

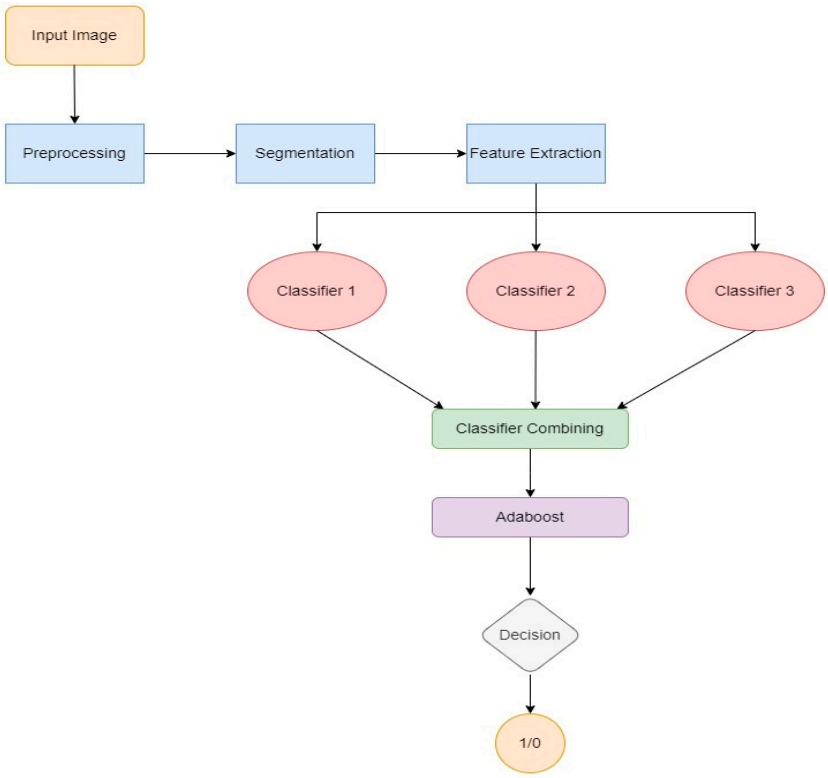

Figure 4. Pre-processing, segmentation, feature extraction, and final distinguishing of infected and uninfected erythrocyte using hybrid classifier.

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, and the one that is implemented is the one whose output will give less accuracy to predict the result of the ensemble of the classifiers. The hybrid classifier comprises 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 the separate classifiers. The misclassifications of every individual classifier are frequently overcome in the hybrid 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)

For the approach used for implementing the hybrid algorithm using SGD, Xgboost and random forest, one is implemented, but the accuracy to predict the result of the ensemble of classifiers is greater than the individual. For the final classification of infected and

uninfected red blood cells, the classifiers are paired using majority voting techniques. And then, the boosting method improves the accuracy of the hybrid machine learning algorithm.

2.9. Classification Based on Hybrid Classifier

The various hybrid classifier pairs are compared to achieve the final outcome of classification. The hybrid classifier's primary objective is to enhance classification accuracy. In the hybrid classifier, misclassifications of individual classifiers 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 on the 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 in Table 1 and Figure 5 is the SGD, logistic regression, and decision tree algorithm and the second in Table 2 and Figure 6 is the SGD, Xgboost and random forest algorithm. The following results shown in the Table 3 and Figure 7are the outcomes of both ways of implementing our proposed hybrid. Using the combination of SGD, Xgboost and random forest, an average of the probabilities for the hybrid is conducted, 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

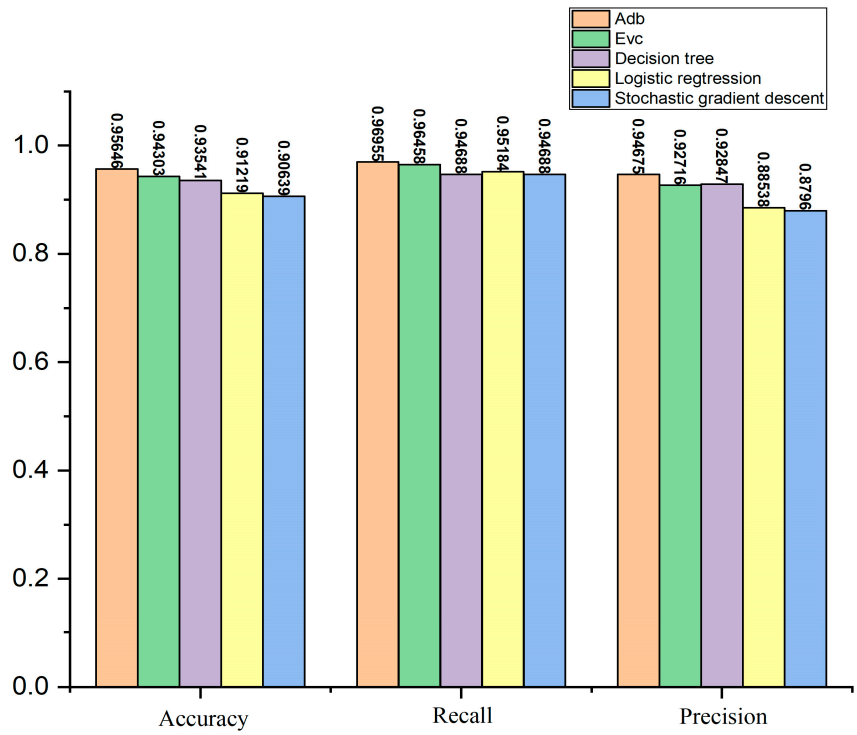

Figure 5. Comparison of individual classifiers in terms of accuracy, precision and recall.

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

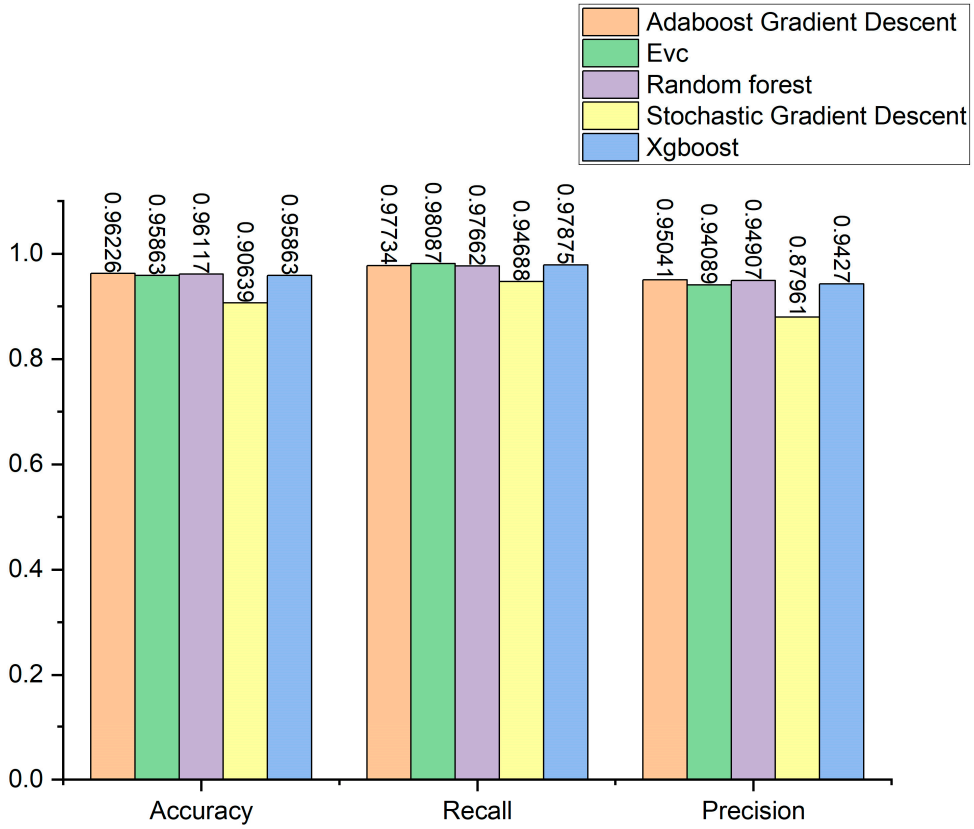

Figure 6. Comparison of individual classifiers in terms of accuracy, precision and recall.

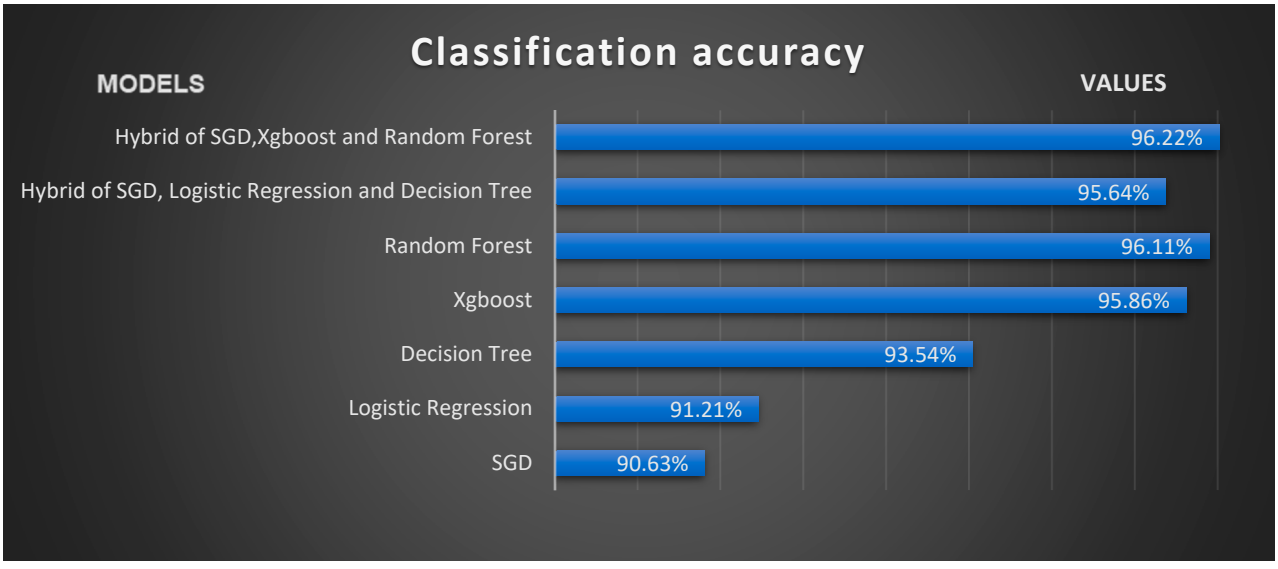

Figure 7. Performance metrics of models.

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%

Comparative Analysis

The classification accuracies of decision tree, SGD, logistic regression, Xgboost, random forest and the hybrid classifiers were compared. The accuracy of the hybrid classifier was the highest as compared to the other classifiers.

4. Conclusions

In this study, the proposed ensemble and hybrid algorithms demonstrate that hybrid machine learning techniques perform better than the individual algorithms on the selected medical dataset of malaria cells. 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 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 is applied to the same dataset, 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 the second hybrid algorithms is better than the combination of the first hybrid algorithms.

Author Contributions: Conceptualization, methodology and editing by S.A., and software validation, data curation, writing—original draft preparation by R.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Kaur, J. Malarial positive image retrieval using Content Based Retrieval Systems. *System* **2015**, *70*, 100.
2. Sarrafzadeh, O.; Dehnavi, A.M. Nucleus and cytoplasm segmentation in microscopic images using K-means clustering and region growing. *Adv. Biomed. Res.* **2015**, *4*, 174.

3. Rakshit, P.; Bhowmik, K. Detection of presence of parasites in human RBC in case of diagnosing malaria using image processing. In Proceedings of the 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), Shimla, India, 9–11 December 2013; pp. 329–334. [CrossRef]
4. Kaewkamnerd, S.; Uthaipibull, C.; Intarapanich, A.; Pannarut, M.; Chaotheing, S.; Tongsima, S. An automatic device for detection and classification of malaria parasite species in thick blood film. *BMC Bioinform.* **2012**, *13*, S18. [CrossRef] [PubMed]
5. Dave, I.R.; Upla, K.P. Computer aided diagnosis of Malaria disease for thin and thick blood smear microscopic images. In Proceedings of the 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2–3 February 2017; pp. 561–565. [CrossRef]
6. Chakrabortya, K. A Combined Algorithm for Malaria Detection from Thick Smear Blood Slides. *J. Health Med. Inform.* **2015**, *6*, 645–652. [CrossRef]
7. Imran Razzak, M.; Informatics, H. Automatic Detection and Classification of Malarial Parasite. *Int. J. Biom. Bioinforma.* **2015**, *9*, 1–12.
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*, 1325. [CrossRef]
9. Anitha Avula, V.; Asha, A. Improving Prediction Accuracy Using Hybrid Machine Learning Algorithm on Medical Datasets. *Int. J. Sci. Eng. Res.* **2018**, *9*, 1461–1467.
10. Charpe, K.; Bairagi, V.K.; Desarda, S.; Barshikar, S. A Novel Method for Automatic Detection of Malaria Parasite Stage in Microscopic Blood Image. *Int. J. Comput. Appl.* **2015**, *128*, 32–37. [CrossRef]
11. Bairagi, V.K.; Charpe, K.C. Comparison of texture features used for classification of life stages of malaria parasite. *Int. J. Biomed. Imaging* **2016**, *2016*, 7214156. [CrossRef] [PubMed]
12. Devi, S.S.; Roy, A.; Singha, J.; Sheikh, S.A.; Laskar, R.H. Malaria infected erythrocyte classification based on a hybrid classifier using microscopic images of thin blood smear. *Multimed. Tools Appl.* **2018**, *77*, 631–660. [CrossRef]