www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Volume-8, Issue-1, February 2020
Research Paper

Int. J. Sc. Res. in Network Security and Communication

E-ISSN:2321-3256

Using Knowledge Discovery to Enhance Classification Techniques for Detect Malaria-Infected Red Blood Cells

J. A. Alkrimi[1]*, Sh. A. Toma[2], R. S. Mohammed[3], L. E. George[4]

[1]College of Dentistry, University of Babylon, Babylon, Iraq
[2]College of Medicine, Baghdad University, Baghdad, Iraq
[3]Al Mansur Institute of Medical Technology, Middle Technical University, Baghdad
[4]Department of Computer Science, College of Sciences, Baghdad University, Iraq

**Corresponding Author: jameela_ali65@yahoo.com, 009647823838400*

Received: 09/Jan/2020, Accepted: 1/Feb/2020, Published: 28/Feb/2020

Abstract— Malaria is one of the three most serious diseases worldwide, affecting millions each year, mainly in the tropics where the most serious illnesses are caused by Plasmodium falciparum. The aim of this research paper is to enhance the main machine-learning classification algorithms that used for malaria-infected red blood cells (MRBCs) and evaluation the classification model accuracy. This study uses knowledge discovery technique to analyses the blood smear images. The system that determines the computerized methods of image analysis generally involves three main phases. Firstly, data collection, pre-processing and feature extraction are conducted based on the characteristics of normal and MRBCs. Secondly, application knowledge discovery process to extracts high quality information of normal and MRBCs. Thirdly, using prediction model of classification machine learning algorithms to classify 1000 RBCs sample. After that, use ten-fold cross-validation to evaluation overfitting model and the confusion matrix to evaluate the performance of a classification model. The results indicate that the algorithms achieve high accuracy more than 92.3%. Also, obtain high prediction 90.8%, reliability 92% and ability to distinguish positive and negative classification model 93%. In addition, the reduction in time build the model was very clearly, 13.6 second and 5.8 times faster respectively.

Keywords *— knowledge discovery, machine learning classification algorithms, feature extraction and feature redaction, red blood cells*

I. INTRODUCTION

Knowledge discovery is type of data mining. Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. It is the process of extracting useful knowledge from collection data [1]. The data perhaps including the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. In real life application, data mining can be used to predict potential diseases that an individual may have based on patient's health records [2,3,4]. Nowadays, data mining has increasing applications in medical science [5]. Data mining lets doctors to provide.

Malaria is a serious public health problem cause of death in many developing countries. It caused by a blood parasite named Plasmodium. In 2015, an estimated 214 million cases of malaria [6,7]. Malaria progresses when the protozoan parasites of Plasmodium are transmitted through the bites of infected female Anopheles mosquitoes, which infect the red blood cells [8]. Malaria parasites attacks RBCs when they enter the bloodstream. In malaria, parasitemia is main step to segment RBCs from blood images and classify them as either parasite-infected or normal. Additionally, the morphology of cells can be clearly observed in thin blood images [3,9,10]. Moreover, Plasmodium has five species that can cause human infection namely, P. falciparum, P. vivax, P. ovale, P. malaria and P. Knowles [8,11].

Classification is one of the data mining strategies used to predict and classify the predetermined data for specific classes. In the past few years, several study found that used machine learning algorithms for automatic detection of malaria infection from microscopic images of stained blood cells to avoid human interpretation error [12]. Different algorithms for classification have been proposed by researchers [13,14]. The most common algorithms are artificial neural network (ANN), Naive Bayes and support vector machine (SVM) [15]. Sajana et al. have measured and reported that numerous conventional machine learning and data mining algorithms can be applied to classify malaria-infected RBCs. [16,17,18] provide high classification accuracy using ANN and SVM. [19] have used ANN technology to detect the infection of malaria RBCs after analyzing digital holographic interferometric microscopic images.

[20] have proposed a method for the detection of malaria parasites based on KNN with ANN techniques. Jan et al. have reviewed on automated diagnosis of

malaria parasite in thick and thin microscopic blood smears images using ANN and SVM [21]. Diaz et al. have applied multilayer perceptron and SVM techniques to classify malaria parasites in RBCs [20,22]. Sharma et al. have realistic SVM and ANN to predict malaria outbreaks, which is the key to controlling malaria morbidity [23]. Kapor and Rani have engaged an efficient decision three algorithm using J48 to reduce error pruning, Bayesian and SVM use the best set of their discriminating features to provide high classification accuracy [24]. Narayanan et al. have development of the computer-assisted malaria parasite characterization and classification through light microscopic images of peripheral blood smears and machine learning approach [25]. Singla et al. have used convolutional neural networks (CNNs) classification algorithms for malaria-infected stages with limited labelled data size [26].

The organization of the paper is as follows: section II introduces the malaria red blood cells (MRBCs) classification scheme, that includes image processing, knowledge discovery techniques, and classification and evaluation performance, while their enhancements and the results are discussed in sections III. Finally, the conclusions of these discussions are shown in the end of this paper.

II. THE METHODOLGY

This study, we suggested MRBCs classification scheme consists of three main stages as shown in Figure 1. The first stage implies that all image processing steps require determining the background/target colors and isolating the cell area (target) from the surrounding. Then, the external and internal boundaries of central pallor area pixels of the cell cut-out were traced. After that, the trace points were used to determine some adopted geometrical features, such as; Fourier descriptors, aspect ratio and moments, which have been used to describe the shapes of RBCs. Also, some textural features were computed to evaluate the spatial color variation within the RBC. In the second stage application knowledge discovery techniques in order to extracting useful knowledge from malaria data collection. The third stage includes machine learning classification algorithms and evaluation performance results. Figure 1shpw the MRBCs classification scheme.

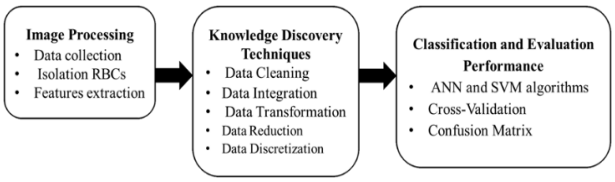


Figure 1. The MRBCs classification scheme

A. Image Processing

Data Collection-

Data collection stage includes capturing of digital blood smear images, image processing and feature extraction. In capturing of digital image, the peripheral blood smear slides of anemic cases were obtained from the hematology unit of the Pathology Department, Faculty of Medicine in Serdang Hospital from March 2012 to August 2012. All peripheral blood slides related to hematological cases were stained by experts with May Grünwald–Giemsa under a light microscope with 10×100 magnification. No loss of parasite was observed during the staining of thick blood smear; the artifacts and parasites were observed in their natural location. To diagnose malaria under a microscope, a drop of the patient's blood is applied to a glass slide, which is then immersed in a staining solution to make parasites more visible under a conventional light microscope, usually with a 100× oil objective. Two different types of blood smears, namely, thick and thin smears, are typically prepared for malaria diagnosis. A thick smear is used to detect the presence of parasites in a drop of blood. Thick smears allow more efficient detection of parasites than thin smears. However, thin smears, which is a result of spreading a drop of blood across a glass slide, still have advantages. They allow the examiner to identify malaria species and recognize parasite stages more easily. Figure 2 shows the blood smear images of patients infected with malaria.

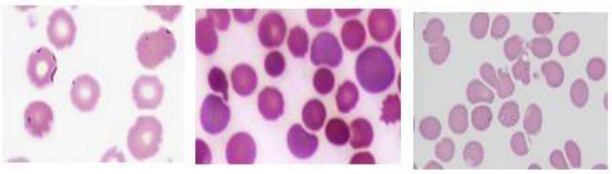


Figure 2. Blood smear images infected malaria

Isolation RBCs-

The capture raw images require several preprocessing steps to isolate each individual normal and abnormal RBC. Each step includes the algorithm, methods and results. The result of each process is input into the second process, and the algorithms improve the quality of the preprocessing. The first step is the conversion of the image from colored to grey. The image uses spectral analysis to obtain the grey stretched image then apply the new image using gamma mapping method to address the distortion in the image edges. To smooth out the MRBCs, the image uses mean filter algorithm. Thereafter, banalization process uses the morphological tiling operation algorithm to convert the grey image to a binary image. The isolation of an individual RBC image is achieved using threshold algorithm as show in Figure 3.

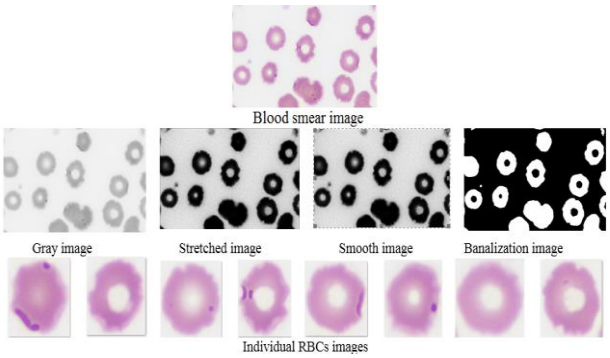


Figure 3. The preprocessing blood smear images

Feature Extraction-

Features extraction is the process of defining a set of features, or image characteristics, which will represent important information for analysis and classification. Ideally, healthy RBCs are doughnut-shaped, whereas infected RBCs become spherical due to the growth of parasites within the RBC. In our work, the different classification techniques of healthy and infected RBC are compared using textures and geometric and statistical features. The set of features are based on the characteristics of normal and MRBCs. The geometric features include the irregularity, size and circularity of RBCs. The textures' features are based on statistical measures which include variance, contrast and moment.

B. Knowledge Discovery Technique

Knowledge discover technique is a sequential process that extracts information from a MRBCs data set. The data collected from the 1000 normal and malaria RBCs image includes 72 features. The data features FMRBCs saved in excel file. The data collected perhaps incomplete, inconsistent, inadequate and it consisting of noise, redundant groups. In order to improve the quality of mining before training data, sequential process includes data cleaning, data integration, data transformation, data reduction and data discretization. The main aim of data cleaning step, is to identify, remove errors and duplicate data, in order to create a reliable dataset. This improves the quality of the training data for analytics and enables accurate decision-making. In this research principal component analysis (PCA) technique was applied to reduce the redundancies features.

The motivations of using PCA: first, to improve the prediction performance of the predictors; Second, to provide rapid and cost-effective predictors; and finally, to provide an improved analysis of the underlying process that generated the data in the ML algorithms. After application PCA technique, the anew data set includes 5 components for 1000 samples named CMRBCs. After that, the FMRBCs and CMRBCs data saved in excel file which allows data to be saved in a tabular format. The data transformation process, convert excel file to the comma-separated values CSV file. CSVs look like a garden-variety spreadsheet but with a .csv extension. After it easy to convert csv file to ARFF data file using Weka soft wear. ARFF data files used as input data in statistical machine learning classifier algorithms. Weka software tool has been used as a platform to perform the classification tests.

C. Classification Machine Learning Algorithms

The classification task uses MRBCs and CMRBCs with two machine learning algorithms: ANN and SVM. These two algorithms will be done using Weka software. Weka software tool has been used as a platform to perform the classification tests. Each algorithm has its own properties, the first algorithm is generative and the second is discriminative.

Artificial Neural Networks (ANN)-

ANN is a type of generative modelling in which machine learning occurs in an unsupervised learning context. It is also a broad area of machine learning where models learn probability distribution P(X) and generate samples from that distribution. We can use these models to create new unseen images by training them on a dataset of images. The ANN structure comprises three layers: input, hidden and output. The first layer is a set of inputs that present 72 features or 5 components. An input vector is used for all radial basis functions, each with different parameters. The second layer presents the hidden layer with a non-linear RBF activation function and a linear output layer. The hidden units and linear activation functions on input and output units will generate a network that is equivalent to a linear classifier. Then, larger numbers of hidden units will be necessary to obtain good classification results. WEKA system that uses a k‐means clustering algorithm to determine the centers and widths of the radial basis functions. The weights are determined by logistic regression. The adjustable parameters include the number of clusters and the ridge parameters for the linear regression. These parameters were experimentally determined; the number of clusters tested, clustering seed and the ridge parameter, which is tested and has a value of $1\times10-8$. Table1 show the values for number of cluster and clustering seed for both data sets. The selection was done after several after several attempts to get high accuracy as show in Table 1.

TABLE 1. NUMBER OF CLUSTERS AND NUMBER OF CLUSTERING SEED

Dataset	Number of Clusters	Clustering Seed		
		1	4	6
FMRBCs	100	82.1%	84. 8%	87.70%
CMRBCs	100	94%	94%	94.1%

Support Vector Machines (SVM)-

SVM algorithm is based on a learning system which uses statistical learning methodology. These algorithms are widely used for classification. In SVM, the optimal boundary, known as the hyperplane, of two sets in a vector space is obtained independently on the probabilistic distribution of training vectors in the set. This hyperplane locates the most distant boundary from the vectors near the boundary in both sets. The vectors that are placed near the hyperplane are called supporting vectors. If the space is not linearly separable, no separating hyperplane is present. Sequential minimal optimization (SMO) is used for training a support vector classifier through polynomial or RBF kernels. SMO replaces all the missing values and transforms nominal attributes into binary ones. A single layer of a hidden neural network uses exactly the same form of model as an SVM [27]. The parameters that are used with RBF kernel and polynomial kernels of SMO-SVM; C the complexity constant, E the exponent for the polynomial kernel, S the seed for the random number generator and

T the tolerance parameter. The two fixed parameters are E and S with values of 1.0 x 10-12 and 1, respectively. We selected value of C = 20, T = 0.9 to be used in the model for classifying both data sets after several attempts until the best results (accuracy) are achieved.

D. Evaluation Performance of Classification Algorithms

Performance evaluation is an important aspect of the machine learning process. The performance of a machine learning classification model can be evaluated in various ways as classification accuracy alone cannot be trusted to select a well-performing classification model [28,29]. Thus, evaluation classification can be conducted in two ways. Firstly, the features used by the machine learning models are their own measurements. Secondly, learned models use performance evaluation [30,31]. At first, cross-validation and accuracy are applied. Then, confusion matrix is applied in order to accurately evaluate the machine learning classifiers.

Accuracy-

The classification of FMRBCs and CMRBCs datasets results must be accurate. Classification accuracy analysis is based on the main parameters' mean squared error (MSE) and mean absolute error (MAE) as well as time. Where the MAS is a measure of how close a fitted line is to data points and it is observed variation in measurements of a typical point. MSE is used for the validation of models. Model validation is an important part of the machine learning process. Validation methods consist of creating models and testing them on a dataset [32,36]. Table shows the results of comparative ANN and SVM algorithm in the classification of both data sets. The comparative results of ANN and SVM algorithms in Table 2.

TABLE 2. THE COMPARATIVE RESULTS OF ANN AND SVM ALGORITHM IN BOTH DATA SET

Algorithms	Dataset	MAE	RMSE	Accuracy	Time sec.
ANN	FMRBCs	0.125	0.3536	87.5%	59.44
	CMRBCs	0.064	0.206	94.1%	4.34
SVM	FMRBCs	0.211	0.331	82.70%	6.04
	CMRBCs	0.137	0.117	92.3%	1.03

Cross-Validation(CV)-

In machine learning, cross-validation methods considered as the most recommended evaluation methods. In a cross-validation test, all the data are used as training and testing datasets [32,33]. In an n-fold CV test, the entire dataset is randomly partitioned into equal-sized n parts. In this study, we have used the most common partitions: (3,5and 10-fold). The 10-fold cross validation technique was used in training and testing both the data sets. Where, 10-fold CV selected after several attempts to achieve higher classification accuracy as show in Table 3. Both datasets are split into 30% training and 70% testing samples. In the training phase, the pre-determined data and its associated class label are used for classification. In the testing step, the test data are used to estimate the accuracy of the classifier rule. If the accuracy of the classifier rule on test data is acceptable, the rule can be applied for further classification of unseen data. The most important factor in medical diagnosis is the accuracy of the classifier.

TABLE 3. COMPARES THE ALGORITHMS ACCURACY WITH K- FOLD VALUE FOR BOTH DATA SETS

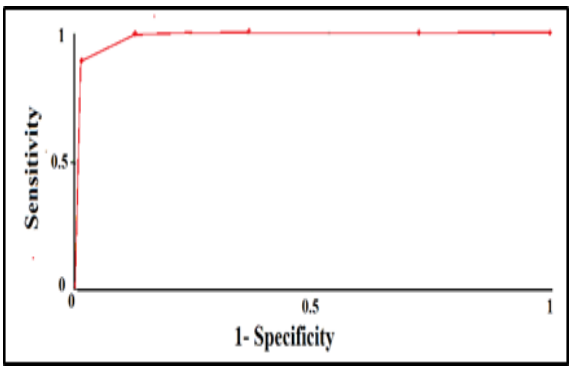
Algorithms	Data set	k-fold		
		3-fold	5-fold	10-fold
ANN	FMRBCs	84%	88.70%	87.50%
	CMRBCs	82.20%	84.50%	94%
SVM	FMRBCs	78.40%	81.59%	89.70%
	CMRBCs	84.20%	86.30%	92%

Confusion Matrix-

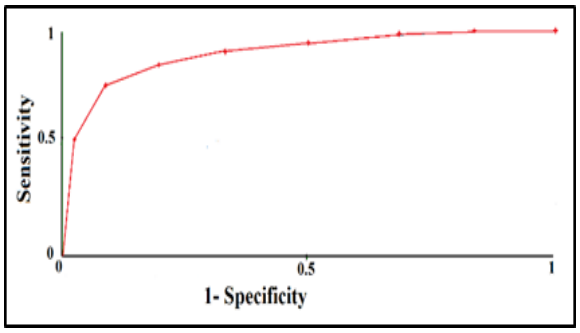
The metrics selected to evaluate the machine learning model is important as the choice influences how the performance of machine learning algorithms is measured and compared [34]. A confusion matrix is a summary of prediction results on a classification problem. It shows the ways in which a classification model is misinterpreted when making predictions [35]. In this research, the numbers and graphs of the main parameters are used to compare the evaluation performance between classifiers ANN and SVM algorithms for both datasets. Sensitivity, specificity, f-measure and kappa statistic are used as numeric parameters. Also, receiver operating characteristics (ROC) are used as graphical parameters, it used to performance of a diagnostic test. Table 4 shows the evaluation performance of SVM and ANN algorithms, while Figure 4 shows the ROC curves of ANN and SVM.

TABLE 4 NUMERICAL EVALUATION PERFORMANCE OF SVM AND ANN ALGORITHMS

Algorithms	ANN		SVM	
Dataset	**CMRBCs**	**FMRBCs**	**CMRBCs**	**FMRBCs**
Sensitivity	92.0%	0.89	0.92	87.20
Specificity	89.0%	86.67	0.82	80.10
F-measure	92.9%	89.80	0.91	88.40
ROC	98.8%	93.20	0.80	78.90
Kappa statistic	93.0%	89.80	0.90	0.88
Accuracy	94.0%	90.40	92.30	89.60



A. ANN ROC curve



B. SVM ROC CURVES

Figure 4. ROC Curves of ANN and SVM algorithm

III. RESULTS

In this section, we report our comparison and analysis of the accuracy results on machine learning classifiers for two data set before and after applying knowledge discover technique. The results showed that the improved ANN classification algorithm had performed better than SVM. The classification task gives the best performance of prediction models as show in Table 2. Also, it gives the best accuracy with 10-fold as show in Table 3. The ANN algorithms achieves the best reliability than SVM. Where it is gives significant prediction agreement with the actual class label as show in Table 4. for Cohen's kappa coefficient value.

IV. CONCLUSION

In summary, we have analyzed and compared the most command classification algorithms that are applied in malaria red blood cells images ANN and SVM. Anew two data sets are created namely FMRBCs and CMRBCs. The performance of classification accuracy is compared to choose the best model. The experimental results show that ANN is more accurate than SVM, having 94.1% accuracy compared with SVM's 92.3% and reduction in time build the model was very clearly, 13.6 second and 5.8 times faster respectively after applying knowledge discover technique. The classifier's accuracy is evaluated using two points. Firstly, the features used by the machine learning models are their own measurements. Secondly, the performance of learned models is evaluated. In numeric parameters, the experimental results show that ANN achieves 92% sensitivity, 89% specificity, 92% f-measure and 93.08% kappa statistic. In graphical parameters, the ROC curves ANN's superior ability to classify the four types of MRBCs.

References

[1] Orenga-Roglá, S. &. (2019). Methodology for the implementation of knowledge management systems. Business & Information Systems Engineering, 61(2), 195-213Abouelmehdi, K. B.-H. (2018). Big healthcare data: preserving security and privacy. . Journal of Big Data, 5(1), 1.

[2] Akter, F. H. (2018). Classification of Hematological Data Using Data Mining Technique to Predict Diseases. Journal of Computer and Communications, 6, 76-83. https://doi.org/10.4236/jcc.2018.64007 Received.

[3] Olayinka, T. C. (2019).). Predicting Paediatric Malaria Occurrence Using Classification Algorithm in Data Mining. Journal of Advances in Mathematics and Computer Science, 1-10.

[4] Witten, I. H. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[5] Pandey, S. C. (2016). Data mining techniques for medical data: a review. International Conference on Signal Processing. Communication, Power and Embedded System (SCOPES), (pp. 972-982). IEEE.

[6] Fernando, M. B. (2018). Plasmodium Falciparum Malaria and Severe Dengue Coinfection in a Pulmonary Tuberculosis Patient: Case Report and Literature Review. Archives of Clinical and Medical Case Reports, 2(3), 82-89.

[7] Report, W. H. (2015). URL: www.who.int/malaria/publications/world-malaria-report-2015.

[8] Poostchi, M. S. (2018). Image analysis and machine learning for detecting malaria. Translational Research, 194, 36-55.

[9] Varma, S. L. (2019). Detection of Malaria Parasite Based on Thick and Thin Blood Smear Images Using Local Binary Pattern. In Computing Communication and Signal Processing, Springer, Singapore., pp. 967-975).

[10] Savkare, S. S. (2011). Automatic detection of malaria parasites for estimating parasitemia. International Journal of Computer Science and Security (IJCSS), 5(3), 310.

[11] Mahdieh Poostchi, K. S. (2019). Image analysis and machine learning for detecting malaria. Translational Research, Volume 194, Pages 36-55.

[12] Saiprasath, G. B. (2019). Performance comparison of machine learning algorithms for malaria detection using microscopic images. IJRAR February 2019, Volume 6, Issue 1.

[13] Chakraborty, D. D. (2015). Computational microscopic imaging for malaria parasite detection: a systematic review. Journal of microscopy, 260(1), 1-19.

[14] Sajana, T. &. (2018). Classification of Imbalanced Malaria Disease Using Naïve Bayesian Algorithm. International Journal of Engineering & Technology, 7(2.7), 786-790.

[15] Adam, A., Chew, L. C., Shapiai, M. I., Jau, L. W., Ibrahim, Z., & Khalid, M. (2011, December). A Hybrid Artificial Neural Network-Naive Bayes for solving imbalanced dataset problems in semiconductor manufacturing test process. In 2011 11th International conference on hybrid intelligent systems (HIS) (pp. 133-138). IEEE.

[16] Vanaja, S. &. (2015). Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey. Journal of Computer Science (JCS), 11(1), 30-52.

[17] Rajaraman, S. J. (2019). Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. PeerJ, 7, e6977.

[18] Sajana, T. &. (2018,). Majority Voting Algorithm for Diagnosing of Imbalanced Malaria Disease. I. n International Conference on ISMAC in Computational Vision and Bio-Engineering (pp. 31-40). Springer, Cham.

[19] Olugboja, (2017). July). Malaria parasite detection using different machine learning classifier. In 2017 International Conference on Machine Learning and Cybernetics (ICMLC), (Vol. 1, pp. 246-250). IEEE.

[20] S.T. Khot, R. (2014). Prasad. Optimal Computer Based Analysisfor Detecting Malarial Parasites. Advances in Intelligent systems and computing, :69-80.

[21] Jan, Z. K. (2018). A review on automated diagnosis of malaria parasite in microscopic blood smears images. Multimedia Tools and Applications, 77(8), 9801-9826.

[22] Gloria Diaz, F. A. (2009). A semi-automatic method for quantification and classification of erythrocytes infected with

malaria parasites in microscopic images. Elsevier - Journal of Biomedical Informatics., (42): 296–307.
[23] Sharma V, A. K. (2015). Malaria outbreak prediction model using machine learning. International. Journal of Advanced Research in Computer Engineering & Technology (IJARCET)., 4(12).
[24] Kapor P, R. R. (2015). Efficient decision tree algorithm using j48 and reduced error pruning. International. Journal of Engineering Research and General Science, 3(3): 2091-2730.
[25] Narayanan, B. N. (2019). Performance analysis of machine learning and deep learning architectures for malaria detection on cell images. Applications of Machine Learning International Societ, (Vol. 11139, p. 111390W).
[26] Singla, N. &. (2019). Deep learning enabled multi-wavelength spatial coherence microscope for the classification of malaria-infected stages with limited labelled data size. arXiv preprint arXiv:1903.06056.
[27] S.S. Keerthi, S. S. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation, 13(3), pp 637-649,
[28] Das, D. K. (2013). Machine learning approach for automated screening of malaria parasite using light microscopic images. Micron, 45, 97-106.
[29] Sokolova, M. N. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation.

Authors Profile

Dr. J.A.Alkrimi, get the Bachelor of Statistics from Administration and economy Baghdad University, Iraq, In 1987 and Master of Information Technology From Iraqi Commission for computers& Informatics. Informatics Institute for Postgraduate Studies, Iraq in year 2006 and Doctoral in Information and Communication Technology, University Tenaga Nasional (UNITEN), Malaysia in year 2015. and currently working as Assistant Professor in Department of Basic Sciences, College of Dentistry, University of Babylon, Iraq since 1989.

