International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 15923-15934

DTBC: Decision Tree based binary Classification using with Feature Selection and Optimization for Malaria Infected Erythrocyte Detection

Md. Jaffar Sadiq[1]
Associate Professor, Department of Information Technology,
Sreenidhi Institute of Science & Technology
Ghatkeser, Hyderabad, Telangana State, India

Dr. V.V.S.S.S. Balaram[2]
Professor & HOD, Department of Information Technology,
Sreenidhi Institute of Science & Technology
Ghatkeser, Hyderabad, Telangana State, India.

Abstract

Malaria is one of the most rapidly spreading and contagious disease, mostly spread through microbes. Efficient treatment of the disease requires early and accurate estimation to ensure control from spreading and treatment in early phases. Accordingly, several studies have been put forward during the past decade. Analyzing the blood smear's images is one of the prominent works proposed in this context. This manuscript attempts to automate the process of diagnosis through machine learning techniques. The algorithm trains the model through different selected features of the input images and thereby uses the learning experience to classify the blood smears as disease prone or healthy. The CUCKOO search algorithm is used for designing a heuristic scale, which is further assessed through multiple experiments to evaluate its accuracy. Different performance evaluation measures like precision, sensitivity, specificity, and accuracy are used to assess the robustness of the model towards early identification of Malaria in the premature stage.

Keywords: Malaria disease, RBCs, ML approach, Blood Smear, case-specific analysis, premature detection

OVERVIEW

Diagnosing malaria disease at premature stage is a difficult task, which is a life-threatening disease mostly in Asia and Africa. Thus, detecting disease at premature stage improves chances of successful treatment. The researchers in [1] classified the malaria parasites that impact human beings into four categories-

Two protozoan parasite Species -

i. Plasmodium falciparum [2]

ii. Plasmodium vivax [2]

Two Parasitic protozoa species -

i. Plasmodium oval [3]

ii. Plasmodium malaria [3]

Of the four types mentioned above, Plasmodium vivax [2] is observed to be the most influencing in hot and humid atmospheres [4]. Detecting the presence of this parasite in blood smears in premature stage is important for effective functioning of the prescribed drugs. The WHO guidelines [5], [6] advise practitioners to perform microscopic diagnosis of blood smears of those patients with likely symptoms to initiate early treatment of Malaria. These tests enable identification of the parasite type and also estimate the count of their presence to determine the Malaria severity. Alternatively, rapid-tests are conducted in certain cases but these tests could not be used for premature detection and accordingly are not considered in this manuscript.

The diagnosis type adapted in this model is relatively easy and involves low costs. The process uses both thick and thin blood smears, which are used to determine the influence of parasite and the type of parasite respectively in the blood. In particular, to determine the scope of the disease in its premature stage requires analysis of the thick smears [6].

The diagnosis test results provide statistics of the parasite including its scope and type. Further, these outcomes are attested by a human expert with knowledge on the statistics. However, this exposes the Malaria identification accuracy to vary with respect to the experience and knowledge of the individual attesting the results. In particular, the chances of accurate detection in the premature stages are often low. Accordingly, computer-aided efficient detection approaches are highly preferred. This manuscript proposes implementing machine learning techniques for detection purpose due to its significance among all computer-aided models. The image attributes generated from the microscopic images of erythrocytes (Red Blood Cells) are vital for the training the

model to differentiate infected and healthy cells in the premature stage.

Digital Image Processing technique forms a crucial part of the model in optimal attribute selection stage. In specific, the edge based segmentation is employed for obtaining benchmark attributes from the microscopic images [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. Though the technique is highly efficient, it faces certain shortcomings including low contrast levels, unclear boundaries due to color similarity, irregular edges and noise involved.

Accordingly, this paper put forward a machine learning approach developed on the principles of CUCKOO search algorithm [20], which uses standard texture and morphological attributes. Further, the paper also presents solutions to handle certain challenges often faced in the process of implementing edge-based segmentation.

The next chapters in the paper present information on related literature (chapter II), suggested approach (chapter III), simulation results (chapter IV) and conclusion and future research scope (chapter V).

RELATED RESEARCH

Computer assisted diagnosis of Malaria by assessing microscopic images of the blood smears has been an important research area over the past decade, attracting several researchers to develop new suggestions and improvements. Contemporary literature presents several related research studies, relying on supervised learning [21], [22], [23], decision support approaches [24], digital image analysis [24], [25], [26] and pattern detection [27], [28]. Further, the researchers in [29] and [30] incorporated artificial NN algorithms for the diagnosis. Histogram-equalizer was incorporated in [31], where segmentation differentiates overlapping cells.

Further, unsupervised approaches also are observed in the contemporary literature including the studies in [32], [33], [34], [35]. ML techniques, image retrieval from content techniques [36], [37] and parasite prediction from the classified samples also are observed to be the prominent studies found in the literature.

Most of these standard approaches often face shortcomings of low contrast levels, same color intensity, irregular boundaries, noise levels and normal regions of the processed images. Excluding the study in [9], rest of the approaches obtain signatures of affected images and incorporate them to detect the extent of parasites in the new image. This can result in high false alarm rates even if small changes occur in the signatures. Accordingly, machine learning approaches are expected to be superior in performance over the aforementioned approaches. However, efficient functioning of ML techniques requires a large number of images to be fed during the learning phase and also requires selecting the optimal attributes of the image.

The ML techniques presented in [9] utilize 94 attributes and ML models SVM [38] and Bayesian approach [39] have been incorporated for learning and simulation stages. Optimal attributes have been chosen through the one-way-ANOVA [40] method. This approach showed 84% precision rates and was inconsistent for varying attribute count. Further, the attribute obtaining process in the approach involves image segmentation through Marker-controlled-watershed method [14], which is based on predicting gradients, which leads to under-segmentation. Accordingly, selection of optimal image attributes is not best.

To address these constraints, the model put forward in this manuscript suggests evolutionary computational ML techniques for selecting optimal features. The CUCKOO search algorithm is employed for feature selection. Further, edge based segmentation process is used for obtaining attributes from the considered blood smear images. This is due to the fact that the edge based segmentation is regarded as the most efficient approach in contemporary studies [35].

EMERGING CALCULATION DRIVEN SCALES FOR PREDICTING MALARIA SPAN

The suggested scale is developed in the sequence of different phases. The phases included in our research work include obtaining existing images of blood smears for learning, pre-processing of input images, identifying optimal attributes based on existing studies, implementing Decision Tree techniques over these attributes to design scale for disease-free and infected blood samples.

A. Representing Original Images as Grayscale Images

The techniques usable to represent a color (RGB) image to mono-color (Grayscale) image include microscopic image to single channel (Greyscale) image are perceptual luminance-preserving convertor [41], Luma convertor [42], Green-channel convertor [42] and PCA convertor [43]. Of these convertors, PCA convertor approach is strong and most desired fit for blood sample images [44], [45], [46]. Accordingly, this approach has been incorporated in our study to represent the color RGB images as grayscale images.

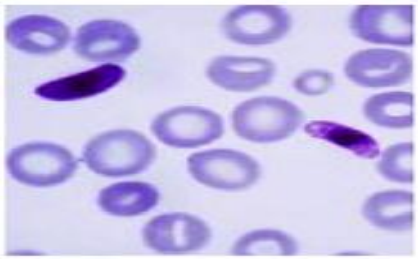
The highest contrast in post conversion image is obtained through linear least-square technique through PCA model. This is executed by evaluating the primary axis of RBC color utilizing RGB coding system. The optimal suitable regression line is generated through PCA regression, which lowers the distance from the data-point to axis on the image in the regression space. The pictorial depiction of this process is presented in Figure 1 where the RGB image is fed to the

model and required grayscale image is obtained as the outcome.

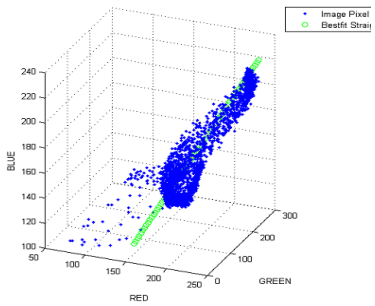
The angles between coordinates reflecting the Red, Green and blue colors are identified initially and later the cosine of these angles are computed to convert the image into grayscale image data. The computation is made on the basis of the below mathematical equation-1

$$\omega_{(x,y,z)} = \frac{\sum_{i=0}^{|P|}(r_i x + g_i y + b_i z)}{|x|+|y|+|z|} \quad \ldots \text{(Eq1)}$$

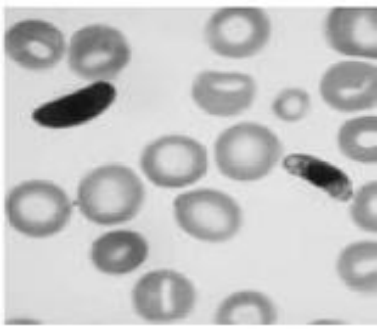
From the Eq1, the output value $\wedge(x,y,z)$ denotes the least the weights x, y, z applied to Red, Green, Blue and $|P|$ denotes the volume of the image as the number of pixels and r_i, g_i, b_i denotes the R, G, B values of the corresponding i^{th} pixel.



a) Colored Image Fed to the Model



b) PCA model based optimal fit coordinates graph



c) Resulting Grayscale image generated by the model

Figure 1: Depiction of converting colored image to grayscale image on the basis of PCA model

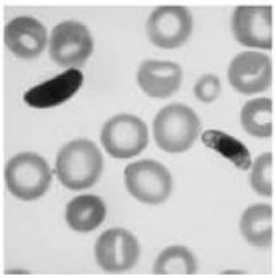
B. Adjusting the Image Contrast Levels

Multiple diagnosis factors can impact the interpretation of images having low contrast levels. Dull images pose an important challenge to categorize the blood sample into Malaria prone or non-prone categories. Accordingly, enhancing illumination of the image is the first phase of the classification procedure. Driven by the advantages of GE approach in enhancing image illumination levels [47], [14], this manuscript opts for this contrast boosting technology.

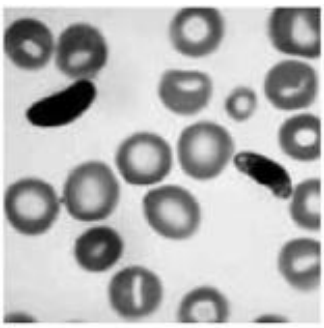
Mathematically, the Gamma equivalent to the given microscopic image $g_{(x,y)}$ is depicted through the below equation-2

$$\left.\begin{aligned} diff_{\min} &= \left(g_{(x,y)} - g_{(x,y)}^{\min}\right)^{\gamma} \\ diff_{\max\leftrightarrow\min} &= \left(g_{(x,y)}^{\max} - g_{(x,y)}^{\min}\right)^{\gamma} \\ f_{(x,y)} &= g_{(x,y)}^{\max} * \left|\frac{diff_{\min}}{diff_{\max\leftrightarrow\min}}\right| \end{aligned}\right\} \quad \ldots \text{(Eq2)}$$

The visualization of the fed low contrast image and the output image is achieved from GE ($\gamma = 0.5$) is presented in the below Figure 2.



a) Low contrast Image Fed to GE model



b) Output image of the GE for $(\gamma = 0.5)$

Figure 2: Depiction of the Functioning of the GE approach for contrast enhancement

C. Lowering Noise Levels

Resultant images from the above phases typically include salt & pepper noise. A typical median filter is applied for handling

this noise type [48]. Another important noise observed in the process is super-imposed image noise [49]. The same median filter in combination with Gaussian filter [49] is considered as ideal for removing this type of noise. Accordingly, the manuscript used the combination of both these filters to reduce noise in the samples.

1) Identification of spectral peaks in pattern noise

As discussed earlier, this research work uses spectral filtering to handle more pattern disturbances in the Fourier amplitude spectrum of the image including noises. The mathematical representation of the same is presented in the below equation-3

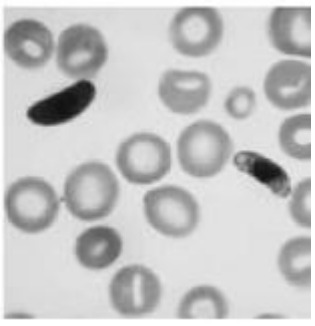
$$A(c_{(k,l)}) = \sum_{i=0}^{L-1}\sum_{j=0}^{W-1}\left(a(i,j)*e\right) \quad \ldots \text{(Eq3)}$$

In the above equation, $a(i,j)$ denotes the pixel of the given image of size $L \times M$, whose Fourier spectrum is computed. The notation $A(c_{(k,l)})$ denotes the amplitude of the coefficient $c_{(k,l)}$. The illuminant dots depicted on amplitude spectrum projects as peaks. The notation e in Eq3 refers the Fourier error coefficient. As the median filter is important filter [50], [51], and is also able to identify impulses as noise [52]. Accordingly, this peak identification approach utilizes median filter to identify peaks by presuming peaks as impulses. The same concept can be applied to our context in Fourier amplitude spectrum.

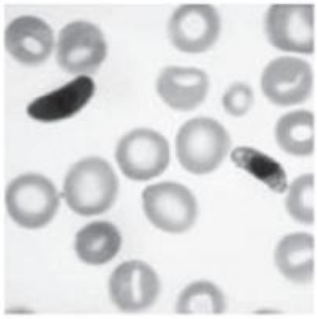
Further, the model adjusts both coefficients being concerned along with the useful amplitude coefficients in the window, in regard to this the model relies on Gaussian Filter.

The Gaussian filter is implemented to identify Gaussian surface, which is coated through two connected peaks. The procedure overhead is observed to utilize this surface for filtering. In case the signal consists of different sets of noise peaks, the procedure overhead is much large.

Accordingly, both the median and Gaussian filters of different volumes are applied to best possible reduction of noise. The median filters limit the area surrounding the peaks and later the Gaussian filter functions over this area to generate optimal image.



a) Grayscale image fed to the model

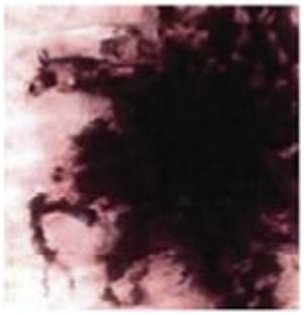


b) Outcome of the model

Figure 3: Blood sample Images fed into the proposed combined model and resultant images

D. Edge Identification

This phase of edge identification attempts to reform the borders of the RBCs through Canny-based filters [53]. These filters are predominantly implemented to save the image continuous edges. Establishment of boundaries of the input image is smoothed first using the median filter with $q \times q$ pixels as mask. This image is noise reduced image obtained from the previous stage. As malaria-prone section of the RBCs is depicted through the darkest part, the edges are the boundaries of the darkest parts as can be observed in Figure below. Accordingly, the filters are able to detect the thickness of the estimated RBC.



a) Image Fed into the Edge Identification model

b) Output image of the RBC with appropriate edge identification

Figure 4: Edge Identification over blood sample deploying Canny Edge model

E. RBCs Categorization through K-Means clustering approach

The preprocessing phases as discussed in subchapters 4.1, 4.2 and 4.3 results in grayscale images that are in the next levels fed as input to categorization phase using the K-Means clustering program with K value of 2. This model is believed to optimal due to the fact that pixels detected in the input image are fall into the category of Malaria prone RBCs and normal RBCs as depicted in Figure 5.

The conventional and easiest clustering approach termed as K-Means [54] has been implemented to group the considered image database. Let database Z be grouped in d dimension area as k groups. In the current scenario, the number of groups is considered as 2 (k=2). First, the healthy and affected RBCs are utilized to build prototypes so that healthy RBCs denote the corresponding group and affected ones denote the related group. Later, all the elements of database Z are grouped into the related group, on the basis of the closest prototype. In addition, the optimal prototype of every group is decided and if prototypes of either of the groups is found to be dissimilar from the previously decide prototype of the corresponding group, then the database Z will be regrouped based on prototypes in the group. This process is iterated until no modification in prototypes of all groups is found and upon successful determination of no changes, groups with corresponding elements are concluded. The distance function that used by k-means detects the distance between each entry of the given dataset D against the cluster centroids as depicted below: equation-4

$$\forall_{i=1}^{|D|}\left\{ f(x) = \max\left(\forall_{j=1}^{|C|}\left\{ \frac{\left| e_i \cap c_j \right|}{|c_j|} \right\} \right) \right\} \ldots \text{(Eq4)}$$

Intersecting of the elements exists in each entry e_i of the dataset D, and each centroid c_j listed in set C should be find first, further, the ratio of these common elements against the count of elements exists in centroid is considered as the similarity $\frac{\left| e_i \cap c_j \right|}{|c_j|}$.

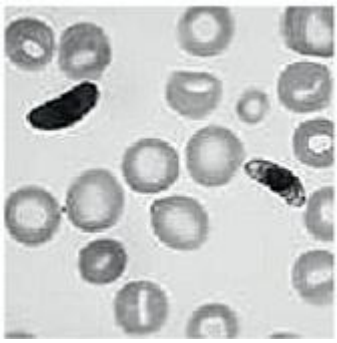
The max similarity of a dataset entry against all of the centroids is considered fittest, and moves that entry to the corresponding cluster represented by the respective centroid. Sequences involved in execution of the program are depicted below:

I. Choose k elements into the area denoted by points/ objects, which are to be grouped. These elements denote first cluster prototypes.

II. Allocate each point o the corresponding cluster, which has nearest matching prototype

III. Upon successful completion of allocation of all points, recomputed the locations of k-prototypes

IV. Iterate sequence II and sequence III till such point that the values of both the prototypes are unchanged. The outcome is the optimal classification of points into clusters, from which the parameter for minimization is computed



a) Image fed into K-Means cluster model

b) Outcome of K-cluster model which depicts two affected groups- white (infected) and black (healthy)

Figure 5: The images fed into K-means algorithm and outcome from the model

Theoretically, the iteration stops at some specific point of time but in practice, the k-Means clustering does not guarantee detection of best configuration related to determination of universal minimum of the function. The program is also highly variable to first chooses group centers. Accordingly, execution of K-means algorithm for grouping the selected images is as below:

a) Groups to be selected must be 2, due to the infected RBCs are the darkest spaces.

b) As the fed image is grayscale image, the pixels are separated based on the pixel intensity.

c) Accordingly, both the groups are built by evaluating all the pixels if they possess the darkest space or not.

To execute this, the primary centroids of the group 1 and 2 are pixels chosen from the most shaded portions of the image and the rest of the portions of the image in the corresponding order.

Further process augments the Malaria causing virus generated black and white image produced through FCM approach. Afterwards, a morphological process, erosion was implemented to nullify certain disturbances in the image. Finally, hole-filling technique was engaged to fill the holes to enable the data ready for classification process.

Most of the times, the output grayscale images of the Malaria evinced RBCs include extra spread areas around the infected RBCs. These extra spread areas must be removed and for accomplishing this, the morphological binary removal process [14] is employed. The structuring component $s_{(x,y)}$ employed to given input image $f_{(x,y)}$ yields resulting the image $g_{(x,y)}$.

In the current research work, a 3×3 square STREL is chosen for the removal function. Numerous structuring components of different volumes, a $3\times3, 5\times5$ and 7×7 have been evaluated. Following series of such evaluations, we found that 3×3 square SE recorded the highest performance among the three sizes.

2) Filling Holes identified in grouped RBC regions

Malaria influenced blood smears depicted through k-means clustering, often evinces the holes in the nucleus of the RBCs. These holes must be shaded to ensure that the model provides accurate classification of affected RBCs for moving on to next stages [55]. The filling image $h_{(x,y)}$ marks as dark in all regions of the given image $f_{(x,y)}$ that excludes the boundaries, where it will complement the image $f_{(x,y)}$ evincing the holes. The resultant image $g_{(x,y)}$ reflects as the original $f_{(x,y)}$ with no holes (see Figure 6).

a) Image fed to 'Connected-Component' Phase

b) Outcome image after deletion of misgrouped RBCs

c) Output image of 'Connected-Component' Phase

Figure 6: Images fed to the CCA model and the resultant outcomes

F. Obtaining Desired Attributes

The attributes in the scenario of classifying the texture along with morphological structure of the corresponding greyscale images are well researched and presented in the previous studies [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] are some of the prominent works focusing on the context.

3) Entropies

Entropy is defined as the extent of ambiguity observed in any context. With respect to sensitiveness of the unpredictability towards distinguishing malaria prone and healthy RBCs, the proposed approach takes into account all existing entropies published in existing literature [15], [16]. Further, to calculate the efficiency of these entropies, we constructed the normalized histograms. Five most prominent and suitable ones are used in this manuscript.

4) GLCM characteristics

The structure data exploited through such attributes is significant [8]. The total 19 attributes associated to power, entropy, variation and extent of correlation [27], [28] can be presented through GLCM characteristics [8]. The GLCMs are $n X n$ matrices, which denote n different dark colors observed in the input image. Moreover, the GLCM is utilized to understand the design parameters.

5) Gray level harmony matrix driven textural characteristics [8]

These attributes enable us to understand the coarse formation of the grayscale image. Consider the matrix $SM_{(i,j)}$ denotes the continuous occurrence of the grayscale value i of the given image $g_{(x,y)}$, which is in sequence of length j. Further this

matrix is in use to define all possible texture features [8], [10], [11].

6) Fractal dimension

The texture formation of any image is often analyzed through the fractal-dimension [18]. So as to detect the Malaria RBC through the texture formation, the grayscale image is regarded as the extra- d of the 2-d image. In addition, the fluctuations in this 3rd-D confirms the formation of the Malaria RBC. The model implemented to detect these fluctuations is termed as MDBC with sequential programming [9], [12].

7) LBP structure

The surface characteristics which depict the likeness of local nearby areas of the image are denoted as LBP [13], [17]. Consider P as the group of pixels present in the circle having radius R of the RBC image I and gp_c denotes the gray value of the pixel p_c at the nucleus of R, and the gray value of each near around pixel p_i of p_c denotes as gp_i. In regard to this, each pixel p_i projects as 0 or 1, which is in accordance of the gp_c.

8) Morphological characteristics

The characteristics put forward in [14] and characteristics of non-variant moments [7], [56] are considered as Morphological characteristics that are highly useful for demonstrating inconsistent RBCs detection. These characteristics are based on analyzing the variations in shape and volume of the healthy and Malaria prone RBCs.

G. Choosing Attributes

- Let the set N represents the features of normal erythrocytes depicted as a cluster during k-means algorithm, and the set M represents the features of malaria prone erythrocytes that also depicted as a cluster in k-means process.
- Apply OFS-Z [57] to select optimal features in regard to infected and healthy blood smears as follows
  - A feature said to be optimal if that feature coverage for both infected and normal erythrocytes are highly covariant. If feature challenges to this property such that coverage for both infected and normal erythrocytes is not diverse then that feature is discarded. The feature optimization is done as follows:
  - The values observed for all considered features in corresponding normal and infected erythrocytes are represented in matrix format of the respective order.
  - The corresponding matrix of the normal erythrocytes are represented as follows:
  - Each row is the values observed for all of the considered texture and morphological features corresponding to a normal erythrocyte.
  - Similarly, matrix will be built for infected erythrocytes, such that each row represents the values observed for all the features considered in respective of infected erythrocyte.
  - Further the significance of each feature is explored by estimating the z-score between the values observed for corresponding feature in respective to normal and infected erythrocytes. If Z-score found to be significant at given degree of probability threshold then the respective feature will be considered as optimal.

•Z-Score

The Z- score that denotes the given two vectors are distinct or not is assessed as follows:

- Let mo_f represents the frequency that observed for feature f in set M
- Let no_f represents the frequency that observed for feature f in set N
- Let $\overline{mo_f}$ represents the set of records, which does not contain feature f in set M
- Let $\overline{no_f}$ represents the set of records, which does not contain feature f in set N

Then the Z-score Z of feature f can be assessed as follows: equation-5

$$\left.\begin{aligned} cr_f &= mo_f - \left(mo_f + \overline{mo_f}\right) * \frac{(mo_f + no_f)}{mo_f + no_f + \overline{mo_f} + \overline{no_f}} \\ dv &= \sqrt{(mo_f + \overline{mo_f}) * (mo_f + no_f) * (1 - (mo_f + no_f))} \\ Zscore_{(f)} &= \frac{cr_f}{dv} \end{aligned}\right\} \ldots(\text{Eq5})$$

Further, find the degree of probability (p-value) of the Z-score $Z(f)$ from the Z-table. If degree of probability is less than the given degree of probability threshold then the feature f is optimal.

BINARY CLASSIFICATION BY DECISION TREES

A. Training Phase: Formatting Decision Tree

This section explores the training phase of the binary classification process that carried through Decision tree. The training phase builds the branches and links appropriate entries as nodes. The depicted branches are in hierarchy, where the feature patterns with max size n that selected as optimal feature will be connected to the branch as nodes that reside at 1^{st} level of the tree hierarchy. Similarly the feature sets of size $n-1$ that selected as the optimal features will be connected to branch exists at 2^{nd} level of the hierarchy, such that the feature sets of size $n-i$ will be connected to branch that exists at $(i+1)^{th}$ level of the hierarchy. The feature sets of size 2 (the minimum size) will be connected to branch that exists at the last level of the hierarchy.

The tree hierarchy is formed such that the first level of the hierarchy contains a branch that represents the nodes formed by the features of size n, second level of the hierarchy contains branches those represent the nodes formed by the features of size n-1, and m^{th} level of the hierarchy contains branches those represent the nodes formed by the features of size n-m-1. Such that. last level of the hierarchy contains branches those represent the nodes formed from the feature sets of size 2.

Hence, the depicted tree hierarchy contains of the branches, the m^{th} level of the hierarchy contains branches such that each branch represents feature patterns of size $n-(m-1)$, here n is the max length of the feature pattern depicted, such that first branch of this hierarchy represents the feature patterns of size $n-(m-1)$. According to the description given about the branch formation, in regard to represent the feature patterns of the normal blood smears, and infected erythrocytes, the depicted model builds two branch hierarchies IH, NH corresponding to the malicious, and benevolent feature patterns.

Upon building the tree hierarchies with appropriate feature patterns connected as nodes in the respective branches, the classification will be initiated to assess the fitness of the given record to predict the malaria scope that explored in in following section.

B. Classifying

The fitness of the given record estimates based on the number of compatible branches noticed in respective hierarchies IH, NH . Concerning this, for each branch, any egg of the respective branch is identical to the values observed in given record for the feature patterns in corresponding branch representative set, then the fitness of the given record in related to corresponding hierarchy will increment by the 1. This practice delivers the fitness related to malicious and benevolent state of the given record. Further the fitness ratio of the given record about to both hierarchies will measure, which is the average of the fitness related to number of branches in corresponding hierarchies. Then the root means square distance of the fitness values corresponding to both hierarchies should measure. Then these fitness ratios and root mean square distances corresponding to both hierarchies will use to confirm the state of the given record is prone to coronary vascular disease or not that explored in following section. The mathematical model to assess the fitness follows

step 1: Let R be the blood smear to be labeled as infected or normal

step 2: Let eR be the set of possible feature patterns projected from the given blood smear R .

step 3: $mf = \sum_{h=1}^{|IH|}\sum_{j=1}^{|IH_h|}\sum_{i=1}^{|eR|}\{1 \exists e_i \in br_{\{h,j\}} \wedge br_{\{h,j\}} \in IH\}$ //add 1 to fitness mf of the given record R related to malicious scope, if feature pattern e_i is found as node to the branch $br_{\{h,j\}}$ in hierarchy IH .

step 4: $ihnc = \sum_{h=1}^{|IH|} |IH_h|$ // number of branches in IH

step 5: $\langle mf \rangle = \frac{mf}{ihnc}$ //Finding the fitness ratio $\langle mf \rangle$ of the given record in related to malicious scope

step 6: $d_{mf} = \frac{\sum_{i=1}^{|mf|}\left\{\sqrt{(1-\langle mf \rangle)^2}\right\} + \langle mf \rangle * (ihnc - mf)}{ihnc}$ // The sum of absolute difference of the fitness ratio depicted, and max fitness in regard to each branch, (which is 1) for number of branches having nodes compatible to the feature patterns exist in eR and the fitness ratio multiplies by the number of incompatible branches, which is the difference between total number of branches and number of compatible branches that denoted as $ihnc - mf$

step 7: $nf = \sum_{h=1}^{|NH|}\sum_{j=1}^{|NH_h|}\sum_{i=1}^{|eR|}\{1 \exists e_i \in br_{\{h,j\}} \wedge br_{\{h,j\}} \in NH\}$ //add 1 to normal scope nf related to hierarchy NH if feature pattern e_i is compatible to any of the node connected to branch n_i in hierarchy NH .

step 8: $nhnc = \sum_{h=1}^{|NH|} |NH_h|$ // number of branches in NH

step 9: $\langle nf \rangle = \frac{nf}{nhnc}$ //Finding the fitness ratio $\langle nf \rangle$ of given record R , in related to normal scope

step 10: $d_{nf} = \frac{\sum_{i=1}^{nf}\left\{\sqrt{(1-\langle nf \rangle)^2}\right\} + \langle nf \rangle * (nhnc - nf)}{nhnc}$ // finding the root mean square distance of the fitness in regard to normal scope using the similar process defined in step 6

C. Discovering the record state

The fitness ratios $\langle mf \rangle, \langle nf \rangle$ and root mean square distances d_{mf}, d_{nf} obtained for given input record R should use to label the record R is prone to disease or normal. The label should define using the conditional flow that follows:

step 1: $if\left(\langle mf \rangle \cong \langle nf \rangle\right)$ Begin

step 2: $if\left(d_{mf} < d_{nf}\right)$ Begin

step 3: Label the record as infected

step 4: End //of step 2

step 5: Else $if\left(d_{mf} > d_{nf}\right)$ Begin

step 6: Label the record as normal

step 7: End // of step 5

step 8: Else //of condition in step 5

step 9: Record state is ambiguous// since the fitness ratios and root mean square distance obtained for both hierarchies is same

step 10: End //of step 1

step 11: Else Begin // of condition in step 1

step 12: $if\left(\langle mf \rangle > \langle nf \rangle\right)$ Begin

step 13: Label the record as infected

step 14: End //of step 11

step 15: Else $if\left(\langle mf \rangle < \langle nf \rangle\right)$ Begin

step 16: Label the record as normal

step 17: End //of step 14

step 18: Else Begin//of condition in step 15

step 19: Record state is ambiguous// since the fitness ratios and root mean square distance obtained for both hierarchies are not meeting the prescribed conditions

step 20: End // of step 18

step 21: End //step 11

SIMULATION PHASE AND EVALUATION OF OUTCOMES

A. The Corpus

In order to perform the experimental study, the labeled (malaria prone, normal) microscopic images from Ma Mic [58], and bio-Sig data [59] databases were collected under statistical guidelines [60]. The total samples are 1600, and among these only 1127 samples were used. The rest of the samples were ignored based on microscopic visible factors. The data statistics of the samples used in experiments were depicted in Table I.

Table I: Dataset Statistics

	Normal	Infected	Total
Training	213	576	789
Testing	91	247	338
Total	304	823	1127

B. The experimental setup

The experiments conducted on i5 generation Intel processor with 4 GB ram and windows operating system. The implementation of the proposed model carried out using FIJI [61], which is a medical image-processing tool. The overall process includes preprocessing, segmenting, feature extraction, feature optimization, and binary classification implemented using relevant Java APIs provided in FIJI.

C. Performance Analysis

The experiments were conducted for proposed model DTBC and another contemporary model called SEMPS [62], which is using the dataset and experimental setup stated in sec A, and sec B

The results obtained for SEMPS, and DTBC are explored in Table II

Table II: The comparison of performance metric values obtained from semps, and proposed model.

Test Statistics	DTBC Features count: 11	SEMPS features count: 17
Positives	247	247
negatives	91	91
TP	230	216
FP	8	13
TN	83	78
FN	17	31
Precision	0.966387	0.943231
Sensitivity	0.931174	0.874494
Specificity	0.912088	0.857143
Accuracy	0.926036	0.869822
Missing Rate	0.068826	0.125506

The experimental outcomes are presented in Table II. The precision that denotes the infected erythrocyte detection value is significantly high as it evinced 97% (0.966), whereas the

precision observed for other contemporary model SEMPS is 94% (0.943). The sensitivity that denotes the right detection rate of infected erythorcytes is observed for DTBC is 93% (0.931), whereas, the counterpart model SEMPS evinced 88% (0.875). The value depicted for metric specificity that denotes the right detection rate of normal erythrocytes are 91% (0.912), and 86% (0.857) from corresponding proposed model DTBC, and counterpart SEMPS. The missing rate that denotes the failure rate of detecting the infected erythrocytes evinced for DTBC, and SEMPS in respective order are 7% (0.07), and 13% (0.126). According to the metric values depicted from experimental study that carried on both proposed model DTBC and SEMPS, it is obvious to confirm the significance of the DTBC towards infected erythrocyte detection from microscopic images of the blood smears. Process complexity of found to be low in DTBC that compared to SEMPS. This is since, the proposed DTBC is using 11 features to build the decision tree, also performs ordered search to assess the class label of the given erythrocite. In contrast to this, the contemporary model is using 17 features to train the model and performs random search on nest hierarchy. The values depicted for statistical metrics from experiments carried on DTBC, and SEMPS are also visualized in a column chart the depicted below (see Figure 7)

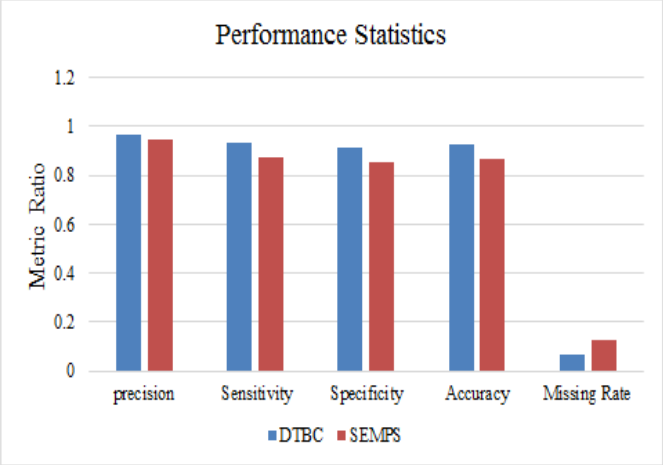


Figure 7: The comparison of the values depicted for statistical assessment metrics from DTBC, and SEMPS

CONCLUSION

This research paper proposed the ECHS scale for assessing the extent of parasites in the input blood smear images. The complete procedure is executed in different stages, which include pre-processing of images to obtain attributes, segmentation, attribute generation, choosing optimal ones and ECHS definition.

Contrary to standard ML approaches, which use SVM and Bayesian approaches for training and testing, the suggested approach determined identification accuracy in proportion to optimal attribute count falling within different hamming distance thresholds. Simulation results also confirmed that the proposed approach involved linear process overheads and significant, stable estimation accuracy rate.

The research findings of the study encourage researchers to conduct in depth research in different aspects like establishing the correlation among different attributes, possible impact on scale definition and incorporating emerging approaches like GA for optimal attribute selection.

REFERENCES

[1] Rougemont, Mathieu, et al. "Detection of Four Plasmodium Species in Blood from Humans by 18S rRNA Gene Subunit-Based and Species-Specific Real-Time PCR Assays." JOURNAL OF CLINICAL MICROBIOLOGY (2004): 5636-5643.

[2] Florens, L., et al. "A proteomic view of the Plasmodium falciparum life cycle." Nature 419.6906 (2002): 520-526.

[3] Pain, A., et al. "The genome of the simian and human malaria parasite Plasmodium knowlesi." Nature 455.7214 (2008): 799-803.

[4] Snow, Robert W., et al. "The global distribution of clinical episodes of Plasmodium falciparum malaria." Nature 434 (2005): 214-217.

[5] World Health Organization. "Guidelines for the treatment of malaria, World Health Organization." Geneva, Switzerland (2010): 9-12.

[6] Reyburn, Hugh. "New WHO guidelines for the treatment of malaria." (2010): c2637.

[7] Hu, Ming-Kuei. "Visual pattern recognition by moment invariants." Information Theory, IRE Transactions on 8.2 (1962): 179-187.

[8] Galloway, M. M. "Texture classification using gray level run length." Comput. Graph. Image Process 4.2 (1975): 172-179.

[9] Mandelbrot, Benoit B. "The fractal geometry of nature/Revised and enlarged edition." New York, WH Freeman and Co., 1983, 495 p. (1983).

[10] Chu, A., C. M. Sehgal, and J. F. Greenleaf. "Use of gray value distribution of run lengths for texture analysis." Pattern Recognition Letters 11.6 (1990): 415-419.

[11] Dasarathy, Belur V., and Edwin B. Holder. "Image characterizations based on joint gray level-run length distributions." Pattern Recognition Letters 12.8 (1991): 497-502.

[12] SARKAR, NIRUPAM, and BB CHAUDHURI. "An efficient differential box-counting approach to compute

fractal dimension of image." IEEE transactions on systems, man, and cybernetics 24.1 (1994): 115-120.

[13] Ojala, Timo, Matti Pietikaeinen, and Topi Maeenpaeae. "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns." IEEE Transactions on Pattern Analysis and Machine Intelligence 24.7 (2002): 971-987.

[14] Gonzalez, Rafael C., and Richard E. Woods. "Processing." (2002).

[15] Pharwaha, Amar Partap Singh, and Baljit Singh. "Shannon and Non-Shannon Measures of Entropy for Statistical Texture Feature Extraction in Digitized Mammograms." Proceedings of the World Congress on Engineering and Computer Science. Vol. 2. 2009.

[16] Ghosh, M., D. Das, and C. Chakraborty. "Entropy based divergence for leukocyte image segmentation." 2010 International Conference on Systems in Medicine and Biology.

[17] Krishnan, M. Muthu Rama, et al. "Textural characterization of histopathological images for oral sub-mucous fibrosis detection." Tissue and Cell 5.43 (2011): 318-330.

[18] Krishnan, M. Muthu Rama, et al. "Statistical Analysis of Textural Features for Improved Classification of Oral Histopathological Images." Journal of Medical Systems 2.36 (2012): 865-881.

[19] Celebi, M. Emre, et al. "An improved objective evaluation measure for border detection in dermoscopy images." Skin research and technology: official journal of International Society for Bioengineering and the Skin (ISBS)[and] International Society for Digital Imaging of Skin (ISDIS)[and] International Society for Skin Imaging (ISSI) 15.4 (2009): 444.

[20] Yang, Xin-She, Trumpinton Street, and Suash Deb. "Cuckoo Search via Lévy Flights." arXiv preprint arXiv:1003.1594 (2010).

[21] Ross, Nicholas E., et al. "Automated image processing method for the diagnosis and classification of malaria on thin blood smears." Medical and Biological Engineering and Computing 44.5 (2006): 427-436.

[22] Kaewkamnerd, Saowaluck, et al. "An automatic device for detection and classification of malaria parasite species in thick blood film." BMC Bioinformatics 13. Supple 17 (2012).

[23] Díaz, G., F. A. González, and E. Romero. "A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images." Journal of biomedical informatics 42.2 (2009): 296-307.

[24] Lai, C. H., et al. "A protozoan parasite extraction scheme for digital microscopic images." Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society 34.2 (2010): 122.

[25] Le, M. T. (2008). A novel semi-automatic image processing approach to determine Plasmodium falciparum parasitemia in giemsa-stained thin blood smears. BMC Cell Biology, 9(1), 15.

[26] Díaz, Gloria, Fabio Gonzalez, and Eduardo Romero. "Infected cell identification in thin blood images based on color pixel classification: comparison and analysis." Proceedings of the Congress on pattern recognition 12th Iberoamerican conference on Progress in pattern recognition, image analysis and applications. Springer-Verlag, 2007.

[27] Tek, F. Boray, Andrew G. Dempster, and Izzet Kale. "Computer vision for microscopy diagnosis of malaria." Malaria Journal 8 (2009): 153-153.

[28] Tek, F. Boray, Andrew G. Dempster, and İzzet Kale. "Parasite detection and identification for automated thin blood film malaria diagnosis." Computer Vision and Image Understanding 1.114 (2010): 21-32.

[29] Memeu, Daniel Maitethia, et al. "Detection of plasmodium parasites from images of thin blood smears." (2013).

[30] Yunda, Leonardo, Andrés Alarcón, and Jorge Millán. "Automated Image Analysis Method for p-vivax Malaria Parasite Detection in Thick Film Blood Images." Sistemas & Telemática 10.20 (2012): 9-25.

[31] Sio, S. W., et al. "MalariaCount: an image analysis-based program for the accurate determination of parasitemia." Journal of microbiological methods 68.1 (2007): 11-18.

[32] Tek, F. Boray, Andrew G. Dempster, and Izzet Kale. "Malaria parasite detection in peripheral blood images." in Proc. British Machine Vision Conference. 2006.

[33] Makkapati, V. V., and R. M. Rao. "Segmentation of malaria parasites in peripheral blood smear images." 2009 IEEE International Conference on Acoustics, Speech and Signal Processing.

[34] Purwar, Yashasvi, et al. "Automated and unsupervised detection of malarial parasites in microscopic images." Malaria Journal 10 (2011).

[35] Somasekar, J., and B. Eswara Reddy. "Segmentation of erythrocytes infected with malaria parasites for the diagnosis using microscopy imaging." Computers and Electrical Engineering 45.C (2015): 336-351.

[36] Das, D. K., et al. "Machine learning approach for automated screening of malaria parasite using light

microscopic images." Micron (Oxford, England: 1993) 45 (2013): 97-106.

[37] Khan, Mohammad Imroze, et al. "Content Based Image Retrieval Approaches for Detection of Malarial Parasite in Blood Images." International Journal of Biometrics and Bioinformatics (IJBB) 5.2 (2011): 97.

[38] Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

[39] Langley, Pat, and Stephanie Sage. "Induction of Selective Bayesian Classifiers." CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE. 1994.

[40] Tabachnick, B. G., and L. S. Fidell. "Using Multivariate Statistics. Boston: Ally and Bacon Pearson Education." (2001).

[41] Iwaki, Y. (2014). U.S Patent No. 8,861,878.

[42] Kanan, C., and G. W. Cottrell. "Color-to-Grayscale: Does the Method Matter in Image Recognition." PLoS ONE 7.1 (2012): e29740.

[43] Kovačević, Jelena, and Amina Chebira. "An introduction to frames." Foundations and Trends in Signal Processing 2.1 (2008): 1-94.

[44] A.S. Abdul-Nasir, M. M. (2013). Colour Image Segmentation Approach for Detection of Malaria Parasites. WSEAS Transactions on Biology and Biomedicine, 10, 41-55.

[45] Yeon, Jun, et al. "Effective Grayscale Conversion Method for Malaria Parasite Detection." (2014).

[46] Kim, Jong-Dae, et al. "Comparison of grayscale conversion methods for malaria classification." International Journal of Bio-Science and Bio-Technology 7.1 (2015): 141-150.

[47] Lai, C. H., et al. "A protozoan parasite extraction scheme for digital microscopic images." Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society 34.2 (2010): 122.

[48] Chokkalingam, Sp, K. Komathy, and M. Sowmya. "PERFORMANCE ANALYSIS OF VARIOUS LYMPHOCYTES IMAGES DE-NOISING FILTERS OVER A MICROSCOPIC BLOOD SMEAR IMAGE."

[49] Wei, Zhouping, et al. "median-Gaussian filtering framework for Moiré pattern noise removal from X-ray microscopy image." Micron (2012).

[50] Astola, Jaakko, and Pauli Kuosmanen. Fundamentals of nonlinear digital filtering. Vol. 8. CRC press, 1997.

[51] MathWorks. (2011). medfilt2. Retrieved from mathworks.com: http://www.mathworks.com/help/toolbox/images/ref/medfilt2.html.

[52] Aizenberg, I., T. Bregin, and D. Paliy. "New method for impulsive noise filtering using its preliminary detection." SPIE Proceedings. Vol. 4667. 2002.

[53] Gonzalez, R. C., R. E. Woods, and S. L. Eddins. "Digital image processing using MATLAB: Pearson Education India." (2004).

[54] Hartigan, J. A., and M. A. Wong. "Algorithm AS 136: A K-Means Clustering Algorithm." Applied Statistics 28.1 (1979): 100-108.

[55] Christ, MC Jobin, and R. M. S. Parvathi. "Segmentation of medical image using K-Means clustering and marker controlled watershed algorithm." European Journal of Scientific Research 71.2 (2012): 190-194.

[56] Das, D., et al. "Invariant moment based feature analysis for abnormal erythrocyte recognition." 2010 International Conference on Systems in Medicine and Biology.

[57] Sadiq, Jaffer MD, Balaram, V.V.S.S.S, "OFS-Z: Optimal Features Selection by Z-Score for Malaria Infected Erythrocyte Detection using Supervised Learning." Proceedings of the First International Conference on Computational Intelligence and Informatics. Springer Singapore, 2018.

[58] http://fimm.webmicroscope.net/Research/Momic/mamic

[59] http://www.biosigdata.com/?download=malaria-image

[60] Altman, D. G., et al. "Statistical guidelines for contributors to medical journals." British medical journal (Clinical research ed.) 287.6385 (1983): 132-132.

[61] Schindelin, Johannes, et al. "Fiji: an open-source platform for biological-image analysis." Nature Methods 9.7 (2012): 676-682.

[62] Jagtap, Chaya D., and N. Usha Rani. "Heuristic Scale to Estimate Premature Malaria Parasites: Scope in Microscopic Blood Smear Images." Indian Journal of Science and Technology 10.8 (2017).