Sādhanā (2018) 43:125
https://doi.org/10.1007/s12046-018-0878-y

FHDT: Fuzzy and Hyco-entropy-based Decision Tree Classifier for Tuberculosis Diagnosis from Sputum Images

K S MITHRA* and W R SAM EMMANUEL

Department of Computer Science, NM Christian College, Manonmaniam Sundaranar University,
Abishekapatti, Tirunelveli 627 012, India
e-mail: ksmithra1@gmail.com; sam_emmanuel@nmcc.ac.in

MS received 4 July 2017; revised 4 October 2017; accepted 12 October 2017; published online 30 June 2018

Abstract. Tuberculosis (TB) is one of the infectious diseases spread by the infectious agent *Mycobacterium tuberculosis*. Sputum smear microscopy is the primary tool used for the diagnosis of pulmonary TB, but has its limitations such as low sensitivity and large observation time. Hence, an automated technique is preferred for the diagnosis of TB. This paper develops a technique for TB diagnosis based on the bacilli count by proposing Fuzzy and Hyco-entropy-based Decision Tree (FHDT) classifier using sputum smear microscopic images. The proposed technique involves three steps: segmentation, feature extraction and classification. Initially, the input sputum smear microscopic image is subjected to a colour space transformation, for which a thresholding is applied to obtain the segmented result. Important features such as length, density, area and few histogram features are extracted for FHDT-based classification that classifies the segments into few-bacilli, non-bacilli and overlapping bacilli. An entropy function, called hyco-entropy, is designed for the optimal selection of feature. For further analysis of classification, that is, to count the number in the overlapping bacilli, the fuzzy classifier is adopted. FHDT classifier is evaluated in terms of Segmentation Accuracy (SA), Mean Squared Error (MSE) and Missing Count (MC) using microscopic images taken from ZNSM-iDB, where it can attain maximum mean SA of 0.954 and mean MC of 2.4.

Keywords. Adaptive thresholding; sputum smear microscopy; fuzzy entropy; hyco-entropy; hyperbolic weighted entropy.

1. Introduction

One of the main health concerns in recent times is tuberculosis (TB) [1, 2]. *Mycobacterium tuberculosis* [3], a nonspore, rod-shaped, aerobic bacterium, is the agent of TB. As TB is an airborne infection, it mainly affects the lungs, leading to pulmonary TB [4, 5]. Based on the information obtained from World Health Organization (WHO), 8.6 million new TB cases were reported worldwide in 2011 that resulted in 1.3 million deaths [6, 7]. In the microscope, typically, bacilli appear in the form of straight or slightly curved rods. It takes beaded and branching form, occasionally occurring in single, pairs or small clumps. The sizes of the bacilli vary between 1 and 10 μm in length and 0.2 and 0.6 μm in width [8]. The WHO suggested that about 300 viewpoints must be inspected per sample of sputum and the number of bacilli that indicates the severity of the disease must be calculated in microscope inspection of the TB sputum smear. Hence, good-quality images are required to be captured automatically before processing from a large number of viewpoints [9]. The two widely used tests to check the infectious subject are sputum smear microscopy and biological culture. For decision-making based on a quick inspection, sputum smear microscopy is used, whereas biological culture test takes about four weeks, even though it is commonly accepted as the gold standard [10].

Among the various TB detection tools available, sputum smear microscopy is commonly used. However, it is performed manually and is time consuming. A laboratory technician must spend at least 15 min per slide by leaving the number of slides that can be screened. For microscopic inspection, the images of sputum smear are stained by either auramine or Ziehl–Neelsen (ZN) staining. Auramine staining plays a major role in fluorescence microscopy, whereas ZN staining is employed in bright field microscopy [11]. An acid-fast fluorochrome having an intense light source, such as a mercury vapour lamp with high pressure or a halogen lamp, is utilised in fluorescence microscopy, while in conventional microscopy ZH acid-fast stains with a conventional artificial light source or sunlight are used. Compared to bright field microscopy, fluorescence microscopy and its stains are expensive [12, 13]. The TB-smear test can be stated positive only after analysing non-bacilli

*For correspondence

objects if the bacilli are scarce. In this case, the WHO suggests that at least 100 non-overlapped images taken from the sample have to be examined so that even one bacillus is enough to find positive sputum smear [10]. Yet, manual screening for TB-smear microscopy may provide false negative results due to sparse bacilli and inspection of very few fields. Therefore, it is necessary to develop an automatic diagnostic process to enhance the sensitivity and the accuracy of the test. Pattern-recognition [14, 15] and image-processing methods are the promising tools employed for automatic screening of TB sputum smear images [16].

Automated TB diagnosis techniques can handle a large number of TB cases easily by maintaining the same accuracy. Presently, in microscopic images, computer-aided diagnosis (CAD) plays a key role. Some of the advantages of automatic screening include a considerable reduction in the labour workload of clinicians, improvement in the sensitivity of the screening and high diagnostic accuracy by increasing the number of images, which can be analysed by the computer. Even though it is advantageous, bacteria segmentation of certain species is a complex task. The shape of the bacillus cannot be considered as a discriminant feature, as other bacteria species may have the same morphology. Hence, in addition to bacilli shape, bacilli colour should also be considered to enhance the discrimination accuracy [1, 17]. An automatic system of TB detection can inspect the existence of TB bacteria easily and automatically from the focused images with or without human intervention. The steps involved in the automation [18] are pre-processing, segmentation, feature extraction and classification.

This paper presents a technique for the diagnosis of TB using sputum smear microscopic images by detecting and counting the number of bacilli. The proposed technique is a three-step process that involves segmentation, feature extraction and classification. The technique performs a colour space transformation, followed by thresholding to segment the image. Statistical features, together with colour, length, density and area, are extracted in the feature extraction process. Finally, a classifier, called FHDT, is designed by hyco-entropy that classifies the segments into bacilli, non-bacilli or overlapped bacilli. By counting the number of bacilli in the given image with overlapped bacilli, the fuzzy classifier provides the result that helps in the diagnosis of TB.

The main contributions of the proposed technique used for the diagnosis of TB are:

- Introduction of FHDT classifier by modifying the Decision Tree (DT) using new entropy function, for the classification of image segments into few-bacilli, non-bacilli or overlapped bacilli, based on the feature vector.
- Designing a novel entropy function, hyco-entropy—by blending fuzzy entropy and hyperbolic weighted entropy—for the best selection and splitting of features that makes the classification effective.

The organisation of the paper is as follows: Section 2 presents the literature survey, where different techniques used for the diagnosis of TB are deliberated. Section 3 explains the proposed technique of TB diagnosis with the proposed FHDT classifier. Section 4 demonstrates the results of the proposed technique and discusses its performance in a comparative analysis and section 5 concludes the paper.

2. Motivation

Due to the urgent need of automated monitoring of TB diseases, various techniques of TB diagnosis have been developed and used extensively in the literature. Along with a brief description of the techniques, their limitations are also stated.

Ebenezer Priya and Subramanian Srinivasan [4] performed object- and image-level classification by considering digital TB images using Multi-Layer Perceptron (MLP) neural network that utilises Support Vector Machine (SVM). For the sputum smear images that were recorded under image acquisition protocol, the TB objects with bacilli and outliers in the images were segmented based on active contour method. The approach has higher accuracy and sensitivity, but lower specificity due to non-uniform morphology.

To separate the overlapping bacilli in the sputum smear images, Priya Ebenezer and Srinivasan Subramanian [16] developed the method of concavity (MOC). The performance of MOC was compared with that of Multi-phase Active Contour (MAC) and Marker-Controlled Watershed (MCW) based on the evaluation using statistical mean quality score. Even though the mean quality score is better, the performance is low when the bacilli density is high.

Ricardo Santiago-Mozos *et al* [10] presented a Bayesian methodology to take decisions in the screening system that considered the false alarm rate. Moreover, a complete screening system was developed in TB diagnosis for sputum smears. Fewer computations were required for the overall process. However, the initial performance of the approach was not acceptable.

Douglas [11] addressed the location of Regions-Of-Interest (ROIs) in the scanned sputum smear slides for detection of TB by focusing on microscope auto-positioning to find the point of reference, orientation and position on the slides. Virtual slide maps and geometric hashing were used to localise a query image that indicates the point of reference. It is robust, tolerating illumination changes, but has low hit rate.

An approach based on Random Forest (RF) was developed by SelenAyas and Murat Ekinci [8] for automated classification of Mycobacterium TB in microscopic images of ZN-stained sputum smears. The RF supervised learning approach was improved for the classification of pixels based on local colour distributions as candidate bacilli regions. Hence, a pixel was labelled as either a candidate TB bacilli pixel or not, and the candidate pixels were grouped using connected component analysis. The accuracy is high, whereas the sensitivity of the approach is low.

Rethabile Khutlang *et al* [19] presented two-class pixel classifier methods for the automated identification of Mycobacterium TB in images of ZN-stained sputum smears that were obtained using a bright-field microscope by segmenting the candidate bacillus objects. Even though the matching score is high, it considers the touching bacilli as non-bacilli.

Sadaphal *et al* [20] established an innovative computational algorithm to recognise ZN-stained acid-fast bacilli (AFB) in the digital images. The colour-based Bayesian segmentation approach could identify the 'TB objects' with artefacts removed. This method could overcome the challenges, such as low depth and extreme stain variation, facilitating electronic diagnosis of TB. However, it did not provide a proper validation.

Costa Filho *et al* [21] presented a technique for detection of TB bacillus in sputum smear microscopy. The technique was composed of two steps, such as bacillus segmentation that selected the input variables for the segmentation using scalar selection approach, and post-processing, where three filters, such as size filter, geometric filter and rule-based filter, were used for the separation of bacilli. The sensitivity and the hit rate are high with reduced error rate, yet, it requires an improved database.

2.1 *Challenges*

Following are the challenges observed in the existing techniques of TB diagnosis studied in the literature survey:

- One of the major challenges in processing the sputum smear images that performs the separation of overlapping bacilli is the higher density of bacilli. Estimating the size and the shape of the image without the separation of touching and overlapping TB objects may result in gross errors during the identification and classification of images [16].
- The human eye has substantial sharpness to discriminate the fine details that consist of high-frequency information, whereas it is not sensitive to images of low frequency or that is which is slowly varying. Hence, computerised automated techniques are required to achieve the challenges with regard to the identification of TB objects, that is, bacilli and outliers, from the sputum smear images [4].
- Another challenge deals with the consideration of bacilli shape as a discriminant feature, as other species and particles of bacteria have the same morphology. Hence, it is necessary to consider bacilli colour, in addition to bacilli shape, in order to enhance the accuracy in discrimination [1, 17].
- The techniques discussed in the literature survey may increase the speed or the sensitivity of microscopy images. However, the quality of the diagnosis is determined based on several conditions, such as heavy workload, poor equipment, and inexpert or unmotivated staff [22]. Moreover, the techniques in ref. [4, 8, 10, 11, 16, 19] were based on traditional classifiers such as SVM, neural network and RF classifier, which are not suitable for the overlapping bacilli.
- The decision trees utilised in ref. [23, 24, and 25] had disadvantages, such as the pre-partitioning, requires previous knowledge of the data, classification is difficult to understand.

3. Proposed work

This section presents the proposed technique of detecting TB by counting the number of bacilli in the sputum smear microscopic images. The proposed technique involves three steps, such as segmentation, feature extraction and classification. Colour space-based segmentation is done over the input image, wherein thresholding is performed to detect the bacilli in the transformed greyscale image.

Based on the segmented image, the features, such as colour, mean, variance, length, density and area, are extracted. Then, an entropy function, hyco-entropy is designed to develop a new classifier, named FHDT. The proposed classifier classifies the image segments into few-bacilli, non-bacilli or overlapping bacilli, where the number of overlapping bacilli is counted using fuzzy classifier. The proposed technique used for TB detection is illustrated using a block diagram shown in figure 1.

3.1 *Colour space-based bacilli segmentation*

The segmentation process utilised in the proposed technique is based on colour space model. The input sputum smear microscopic image is an RGB image, which is transformed into L*u*v space using CIELuv colour space model. Taking 'u' space from the image, Otsu thresholding, which is a kind of global thresholding that depends on greyscale image, is done. The steps involved in the segmentation process are given below:

Step 1: Read the input image
Step 2: Apply CIELuv colour space model
Step 3: Take 'u' space from the image
Step 4: Apply Otsu thresholding for the binarisation of the image

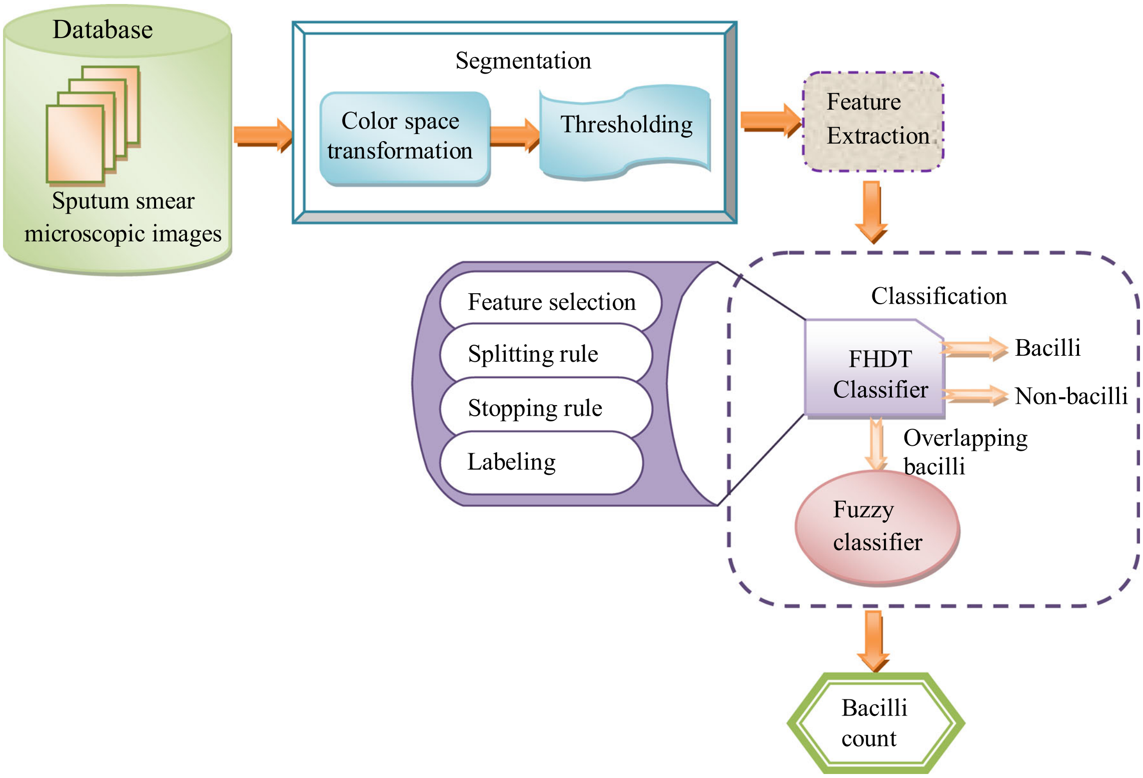


Figure 1. Block diagram of the proposed technique of TB diagnosis using FHDT classifier.

Step 5: Apply morphological opening in the resulting image to remove the noise

The segmentation process is illustrated using figure 2. The input image shown in figure 2(a) is a few-bacilli image. Figure 2(b) shows the colour space result, and the resulting image after applying CIELuv colour space is shown in figure 2(c). Figure 2(d) depicts the segmented output of the input few-bacilli image.

3.2 *Feature extraction for the classification of bacilli*

The proposed process of feature extraction performed over the segmentation output is explained in this part. The important features extracted from the image are colour, mean, variance, length, density and area. Extraction of these six features from the image makes the further process of classification easy for the diagnosis of TB. Colour, mean and variance are the histogram features [26] that are computed based on a statistical model, where the histogram is considered as the probability distribution model of the intensity levels. The histogram is a plot made with the grey-level values of a colour channel against the number of pixels. Based on the shape of the histogram, the information regarding the bacilli in the image can be obtained. The probability of the first-order histogram p(b) is given in Eq. (1),

$$p(b) = \frac{N_b}{P \times Q} \tag{1}$$

where b represents the grey level, N_b is the number of pixels at b and $P \times Q$ is the size of the image.

3.2a *Colour*: For classification with respect to visual features, colour is the widely used. Robustness, effectiveness and simplicity in computation are the important factors that make colour the commonly adopted feature. To extract the colour features, a colour histogram of the image in each colour channel, that is, Red (R), Green (G) and Blue (B), is taken separately. The colour histogram, which represents the percentage of pixel in the image, is computed based on the number of pixels having the same colour as in Eq. (2).

$$\left.\begin{aligned} p^R &= N_b^R(b) \\ p^G &= N_b^G(b) \\ p^B &= N_b^B(b) \end{aligned}\right\} ;\ 1 \le b \le 255 \tag{2}$$

where $N_b^R(b)$, $N_b^G(b)$ and $N_b^B(b)$ are the number of pixels at grey level b in R, G and B bands, respectively. Hence, the colour histogram H is obtained from Eq. (3):

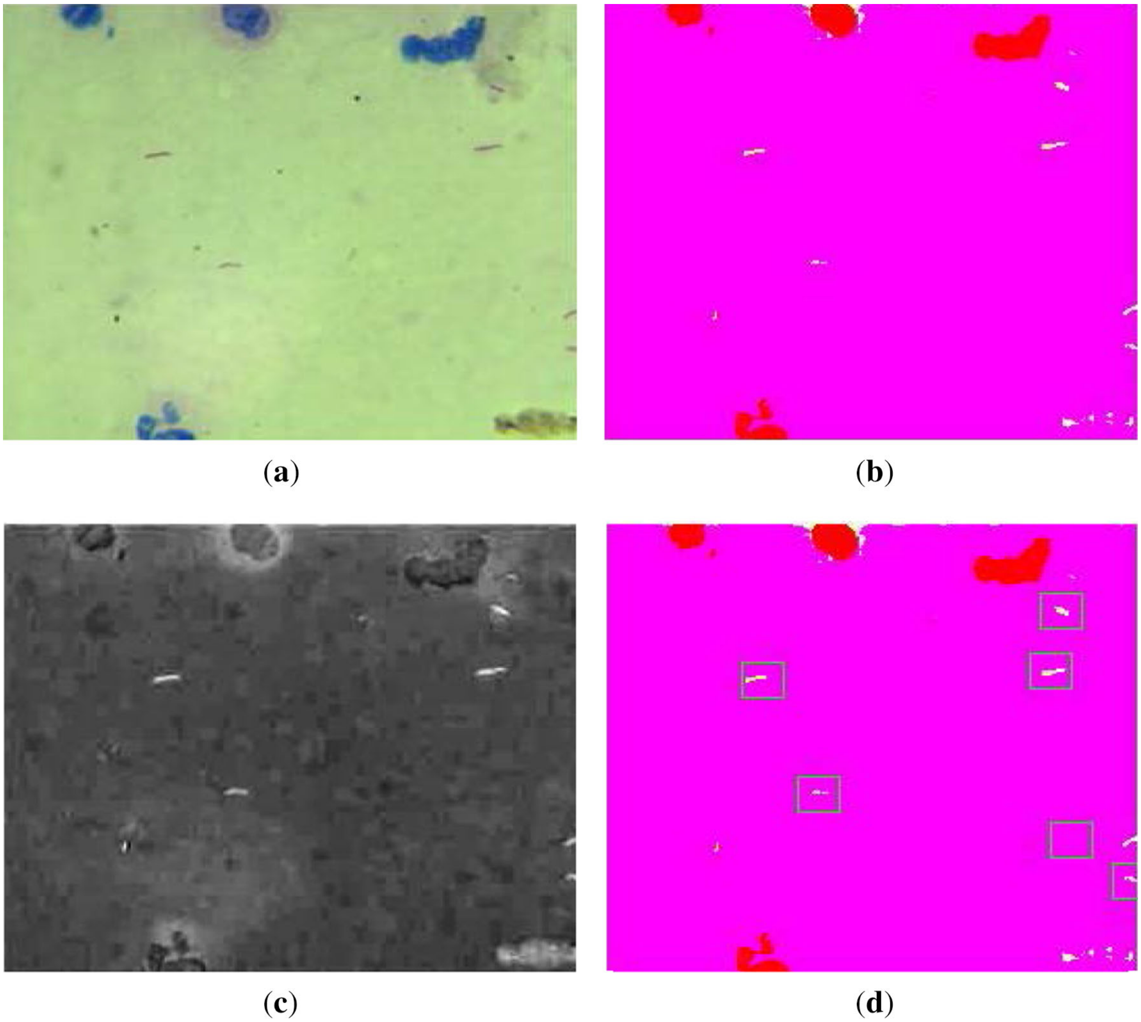


Figure 2. Segmentation process. (**a**) Input image, (**b**) CIELuv color space, (**c**) U matrix and (**d**) Segmented bacilli.

$$H = \begin{bmatrix} p^R & p^G & p^B \end{bmatrix} \tag{3}$$

where p^R, p^G and p^B are the histogram probabilities in R, G and B bands.

3.2b *Mean*: The mean, also known as the average value, indicates the brightness of the image. A bright image has a high mean, whereas a dark image has a low mean. The number of grey levels in the segmented region varies from 0 to 255. The mean of an image $\bar{b}$ is defined in Eq. (4) by

$$\bar{b} = \sum_{b=0}^{G-1} bp(b) = \sum_{m=1}^{P}\sum_{n=1}^{P} \frac{I(m,n)}{P \times Q} \tag{4}$$

where $p(b)$ is the histogram probability, G is the total number of intensity levels and $I(m,n)$ is the input segmented image.

3.2c *Variance*: The variance, which refers to the contrast of the image, is high for the high-contrast image and is low for the low-contrast image, as it illustrates the dispersion or the spread in the data. It is computed as the average squared deviation of each value from the mean, as expressed in Eq. (5)

$$\sigma^2 = \sum_{b=0}^{G-1} (b - \bar{b})^2 p(b) \tag{5}$$

where $\bar{b}$ is the mean and $p(b)$ is the probability.

3.2d *Length*: The length is measured along the horizontal and the vertical axes of the image by measuring the difference between the starting and the ending positions of the object, as given Eqs. (6) and (7).

$$H_L = |X_{\min} - X_{\max}| \tag{6}$$

$$V_L = |Y_{\min} - Y_{\max}| \tag{7}$$

where $X_{\min}$ and $X_{\max}$ represent the positions on the horizontal axis and $Y_{\min}$ and $Y_{\max}$ represent the positions along the vertical axis.

3.2e *Density*: To extract the pixel density features, a method called zoning [27] is employed. Here, the images are partitioned into zones of particular sizes that are predefined such that the features for each zone are measured. Let $Z = \{Z_1, Z_2, \ldots, Z_d\}$; $1 < l \leq d$ represent the zones in an image segment, and d is the total number of zones. Zoning, which gives the local characteristics of the image, computes the average pixel density based on a ratio using Eq. (8).

$$D(l) = \frac{N_l^F}{K_l} \tag{8}$$

where N_l^F is the number of foreground pixels in the l^{th} zone and K_l is the total number of pixels in the l^{th} zone.

3.2f *Area*: For the extraction of features based on area, the area is calculated as the total number of pixels in the segmented portion of the image. The number of black pixels in each segment of the image is computed and the computation results give the features that are used for the classification of bacilli in diagnosing TB.

Thus, in Eq. (9), a feature vector f of dimension 1 × 6 is obtained for each segment of the image.

$$f = \{f_1, f_2, \ldots, f_h\} \tag{9}$$

where h is the total number of features extracted from the image and its value is 6 in this work. The feature vector obtained is the input used for the construction of DT.

3.3 *Proposed fuzzy and hyco-entropy-based decision tree for the classification of bacilli*

In this section, the proposed classifier, FHDT, developed for the classification of bacilli in the image that helps in the diagnosis of TB, is explained. The DT [28] is a tool used for classification taking a tree structure form through the rules obtained from the input feature vector and results in a tree with decision and leaf nodes. The proposed FHDT classifier segregates into few-bacilli, non-bacilli and overlapping bacilli. The proposed classifier is developed by modifying the DT with a new entropy function, hyco-entropy that utilises fuzzy entropy and hyperbolic weighted entropy, for the selection of features. The hyperbolic function used here is the hyperbolic cosine function for the feasible selection. The fuzzy entropy can partition the input features into decision regions selecting the suitable features for the classification. With the combined effects of fuzzy entropy and hyperbolic weighted entropy, an effective classification can be performed. The four stages included in the proposed classifier are discussed subsequently.

3.3a *Feature selection using hyco-entropy*: To select the best feature in FHDT classifier, hyco-entropy is designed. The proposed entropy function is computed for the j^{th} feature as given in Eq. (10) such that the feature having maximum value forms the best feature.

$$HC(f_j) = w.E(f_j) \tag{10}$$

where w is the weight function obtained from Eq. (11) and $E(f_i)$ is the entropy.

$$w = 2\left[1 - \frac{1}{1 + \cosh(-E(f_j))}\right] \tag{11}$$

where cosh(.) is the hyperbolic cosine function employed rather than the exponential function used in paper [28]. The hyperbolic cosine function is usually used for defining complex calculations to simplify the results. Moreover, this can bring out the best feature, which is better than one obtained using the exponential function, to perform the classification together with the entropy, given in Eq. (12).

$$E(f_j) = \sum_{y=1}^{v(f_j)} f_y^E(F) \tag{12}$$

where $f_y^E(F)$ is the fuzzy entropy, F is the fuzzy set and $v(f_j)$ is the unique value. The fuzzy decision regions that are generated can reduce the computational load and the complexity and, thereby, the speed of classification will be enhanced. The selection procedure with the utilisation of fuzzy entropy [29] improves the classification rate by neglecting insignificant features.

Let $c = \{c_1, c_2, \ldots, c_k\}$; $1 < i \leq k$ be the class representation with k number of classes, where $k = 3$ represents the three classes: bacilli, non-bacilli and overlapping bacilli. Then, the fuzzy entropy is obtained using Eq. (13).

$$f_y^E(F) = -M_y \log_2 M_y \tag{13}$$

where M_y is the match degree with the fuzzy set F and is formulated with respect to a membership degree as in Eq. (14).

$$M_y = \frac{\sum_{S_i^y \in c_i} \mu(v(f_j))}{\sum_{y=1}^{v(f_j)} \mu(v(f_j))} \tag{14}$$

where $\mu(v(f_j))$ is the mapped membership degree and S_i^y is the set of elements that belongs to the i^{th} class. Then, the mapped membership degree of the unique value of the j^{th} feature is given by a ratio in Eq. (15).

$$\mu(v(f_j)) = \frac{p(v(f_j))}{g} \tag{15}$$

where $p(v(f_j))$ is the probability of occurrence of the unique values and g is the number of data. The element set belonging to the particular class is given by the probability that belongs to the class i, as given in Eq. (16).

$$S_i^y = \frac{p(v(f_j)) \in c_i}{g}\; ; 1 \le v(f_j) \le z \tag{16}$$

Thus, the proposed hyco-entropy obtained by combining fuzzy entropy and hyperbolic weighted entropy, which will provide the best feature selection. FHDT generates the root node initially, and with the features selected using hyco-entropy, the branch is formed.

3.3b *Hyco-entropy gain-based splitting rule*: After the selection of features, the splitting process is carried out by determining the optimal split value. The splitting rule provides the split value based on the unique values of the features. This requires information gain, called hyco-entropy gain, calculated from Eq. (17) based on the difference between hyco-entropy and conditional hyco-entropy.

$$HCG(f_j, f_q) = HC(f_j) - CHC(f_j, f_q) \tag{17}$$

where $HC(f_j)$ is the hyco-entropy and $CHC(f_j, f_q)$ is the conditional hyco-entropy, which are defined in Eqs. (18) and (19).

$$CHC(f_j, f_q) = \sum_{j=0}^{v(f_j, f_q)} p_j HC(f_j, f_q) \tag{18}$$

where p_j is the probability and $HC(f_j, f_q)$ is the hyco-entropy with features f_j and f_q. The function $HC(f_j, f_q)$ is represented similar to Eq. (10), but with the two features, as defined in Eq. (19).

$$HC(f_j, f_q) = w_h.E(f_j, f_q) \tag{19}$$

where the weight function w_h is defined in Eq. (20)

$$w_h = 2\left[1 - \frac{1}{1 + \cosh(-E(f_j, f_q))}\right] \tag{20}$$

where cosh(.) is the hyperbolic cosine function. Equation (21) defines the entropy with the features f_j and f_q.

$$E(f_j, f_q) = \sum_{y=1}^{v(f_j, f_q)} f_y^E(F) \tag{21}$$

where $f_y^E(F)$ is the fuzzy entropy, F is the fuzzy set and $v(f_j, f_q)$ is the unique values of feature attributes f_j and f_q. The fuzzy entropy considering the attributes f_j and f_q is given in Eq. (22).

$$f_y^E(F) = -M(f_j = j, f_q = j) \log M(f_j = j, f_q = j) \tag{22}$$

where $M(f_j = j, f_q = j)$ is the match degree with the fuzzy set F and is formulated with respect to a membership degree as given in Eq. (23).

$$M(f_j = j, f_q = j) = \frac{\sum_{S_i^y \in c_i} \mu(v(f_j, f_q))}{\sum_{y=1}^{v(f_j, f_q)} \mu(v(f_j, f_q))} \tag{23}$$

where $\mu(v(f_j, f_q))$ is the membership degree with respect to the unique values of the features and S_i^y is the element set that belongs to the i^{th} class. Equation (24) gives the mapped membership degree of the features f_j and f_q.

$$\mu(v(f_j, f_q)) = \frac{p(v(f_j, f_q))}{g} \tag{24}$$

where $p(v(f_j, f_q))$ is the probability of occurrence of the unique values of the features and g is the number of data. The set of elements belonging to the i^{th} class is obtained from Eq. (25).

$$S_i^y = \frac{p(v(f_j, f_q)) \in c_i}{g}\; ; 1 \le v(f_j, f_q) \le z \tag{25}$$

Based on the splitting rule, the node is plotted and the branch is grown. FHDT classifier processes the splitting rule until the stopping criterion is met.

3.3c *Stopping rule*: The stopping criterion is reached when the node considered has the label distribution covering most of the data.

Algorithm 1. Pseudocode of FHDT.

```
Algorithm 1 FHDT Construction
procedure FHDT
    Create the root node
    for (each attribute $f_j$)
        Calculate$HC(f_j)$ using Eq. (10)
    end for
    Find the best feature using Eq. (16) to place in the node
    Fix up the split value of the best feature
    for (each split value)
        Divide the samples in accordance with the class$c$
        Calculate $HCG(f_j, f_q)$using Eq. (17)
    end for
    Construct the branch based on the best split value
    Build up the tree by forming a new node recursively and
      splitting the samples till the stopping rule is met
end procedure
```

3.3d *Labeling*: For the leaf node, the label is determined by identifying the class having the maximum number of data satisfied. Thus, the FHDT classifier performs classification providing the result based on the three classes, bacilli, non-bacilli and overlapping bacilli. The proposed technique is developed for counting the number of bacilli in the sputum smear images if the result is overlapping bacilli. The procedure of the proposed FHD classifier is illustrated using

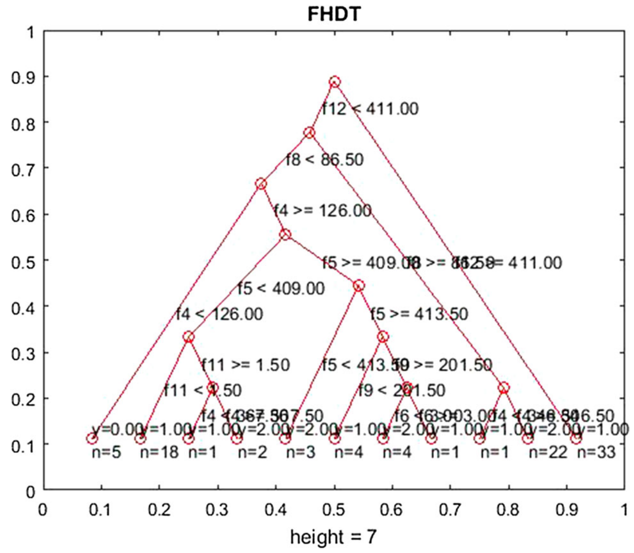


Figure 3. Construction of FHDT.

the pseudo code given in Algorithm 2 and the constructed FHDT is shown in figure 3.

3.4 *Adapting fuzzy classifier for counting the overlapped bacilli*

For further analysis of classification, that is, to count the number of bacilli in the overlapping bacilli, a fuzzy logic system [30] is employed. The robustness, simplicity in implementation, among others, make the fuzzy logic system an effective decision-making system. It analyses the information based on the fuzzy sets, which are represented by three linguistic terms, namely "small," "medium" or "large."The fuzzy logic system is composed of three components, such as fuzzification, inference, and defuzzification. In fuzzification, the input value is converted into linguistic terms, whereas in defuzzification, the fuzzy facts are converted into a crisp value. Meanwhile, the fuzzy interface involves the rule base, where the if-then rules are obtained. Based on the bacilli count in the overlapping bacilli provided by the fuzzy logic system, the diagnosis of TB can be made. The decision is made depending on the input factors, which are the features extracted, such as length, density and area. Figure 4 shows the fuzzy logic system used for counting the number of bacilli.

3.4a *Input parameters*: The input factors to the fuzzy logic system are horizontal length (H_L), vertical length (V_L), density ($D(l)$) and area, which is denoted as$A$. These parameters, which are applied to the fuzzification part of the system, determine the number of bacilli in the image, providing the number of counts, denoted as 'a', based on the fuzzy score. Larger values of these factors indicate that the bacilli are overlapping and, hence, these are utilised for the determination of bacilli count.

3.4b *Fuzzy membership functions*: Following are the labels of the fuzzy variables used in the fuzzy system:

- H_L = {"small", "medium", "small"}
- V_L = {"small", "medium", "large"}
- $D(l)$ = {"small", "medium", "large"}
- A = {"small", "medium", "large"}

Depending on these labels, the count a produces the output using the fuzzy scores. These are fed to the inference part of the system, where a number of fuzzy rules are obtained. Here, the triangular membership function is utilised.

3.4c *Fuzzy if-then rules*: The rule base in the fuzzy inference generates a number of fuzzy rules based on the input parameters and the labels. Some of the sample rules generated are given as follows:

i) if H_L is "large", then a is "large".
ii) if $D(l)$ is "medium", then a is "*medium*".
iii) if A is "*small*", then a is "*small*".

As shown above, under varying conditions based on the input variables and the labels, the rules are formed. In the defuzzification part, based on the rules obtained with the linguistic terms "*small*", "*medium*" and "large", the fuzzy scores 2, 3 and 4, ar e generated, which gives the corresponding output. If the term associated with the input variable is "*small*", the fuzzy system provides the count $a$as 2. If the term is "*medium*", the value of $a$is 3 and 4 for "large". Thus, the fuzzy logic system counts the number of bacilli in the overlapping bacilli.

4. Results and discussion

This section demonstrates the results of the proposed FHDT classifier developed for the diagnosis of TB using sputum smear microscopic images. The experimental set-up and the comparative analysis performed are explained in the following subsections.

4.1 *Experimental set-up*

The experiment is performed using a PC of the following configuration: Windows 10 OS, Intel core processor with CPU 2.14 GHz and 2GB installed memory. The software tool used for the implementation of the proposed technique is MATLAB.

4.1a *Dataset description*: The database used for the experimentation is Ziehl–Neelsen Sputum smear Microscopy image Database (ZNSM-iDB) [31]. The database has a collection of digital images of diverse smear microscopy taken from three microscopes. The datasets include three categories of images, such as non-bacilli, few-bacilli, and overlapping bacilli. Ten images from few-bacilli and non-bacilli category of images and five images from overlapping bacilli category are selected in the experiment.

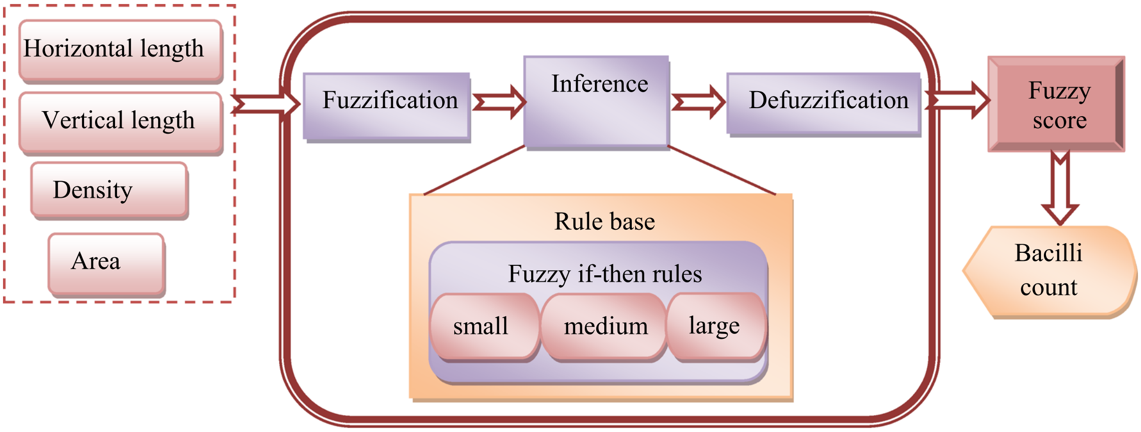


Figure 4. Fuzzy logic system to determine the count of bacilli.

4.2 *Comparative techniques*

In the comparative analysis, three methods are used to evaluate the performance of the proposed technique. The methods utilised for the comparison are (i) SVM [32]; (ii) Train Br [33]; and (iii) Train LM (Applied Levenberg–Marquardt (LM) algorithm instead of Back propagation (BP) in [33]). In ref. [32], a method was followed for bacillus identification, where the segmentation was carried out using SVM and neural network classifiers, whereas in ref. [33], TB was diagnosed using multilayer neural networks (MLNNs) that used BP for training. These techniques are compared with the proposed FHDT classifier to evaluate its performance.

4.3 *Evaluation metrics*

Comparing the techniques on performance criteria, three metrics such as segmentation accuracy (SA), mean squared error (MSE), and missing count (MC) are utilised.

SA measures the degree of segmentation accuracy obtained by the comparative techniques, as represented in Eq. (26).

$$SA = \frac{B \cap C}{B \cup C} \tag{26}$$

where B is the bacillus pixel output of an image after applying the classification technique and C is the ground truth bacilli pixel of the image.

MSE refers to the average of the squares of values that deviate from the original and is derived from Eq. (27).

$$MSE = \frac{1}{|B|}\sum_{i=1}^{|B|}(r_i - s_i)^2 \tag{27}$$

where r_i is the number of bacilli in the i^{th} bacillus segment of an image,s_i is the original bacilli count in the i^{th} bacillus segment of the image and $|B|$ is the total number of bacilli.

Missing count is measured from Eq. (28) based on the difference of actual bacilli count and the count measured.

$$MC = \sum_{i=1}^{|B|}|r_i - s_i| \tag{28}$$

where $|.|$ denotes the absolute value.

4.4 *Experimental results*

This section describes the experimental results of the proposed FHDT classifier using sputum smear microscopic images. Figure 5 presents the segmentation results of the proposed technique. In figure 5a, the input few-bacilli image considered for the segmentation is given. The colour space obtained using CIELuv for the input few-bacilli image is shown in figure 5b and the segmented result is depicted in figure 5c. The second category, that is, no-bacilli image, taken as input, is presented in figure 5d, and its colour space and the segmented output are illustrated in figure 5e and f, respectively. Similarly, for the input overlapping image given in figure 5g, the CIELuv colour space result is shown in figure 5h and the segmented result in figure 5i. As shown in the results, a clear segmentation output can be obtained using the colour space-based segmentation approach.

4.5 *Performance analysis*

The performance of the proposed technique is compared with that of three existing techniques, namely, SVM, Train Br, and Train LM, based on the evaluation metrics by varying the percentage of training data from 50 to 80.

4.5a *Analysis based on SA*: Figure 6 demonstrates the performance analysis based on SA for the training data varied as 50%, 60%, 70% and 80%, respectively, in the comparative techniques. The first ten images represent the

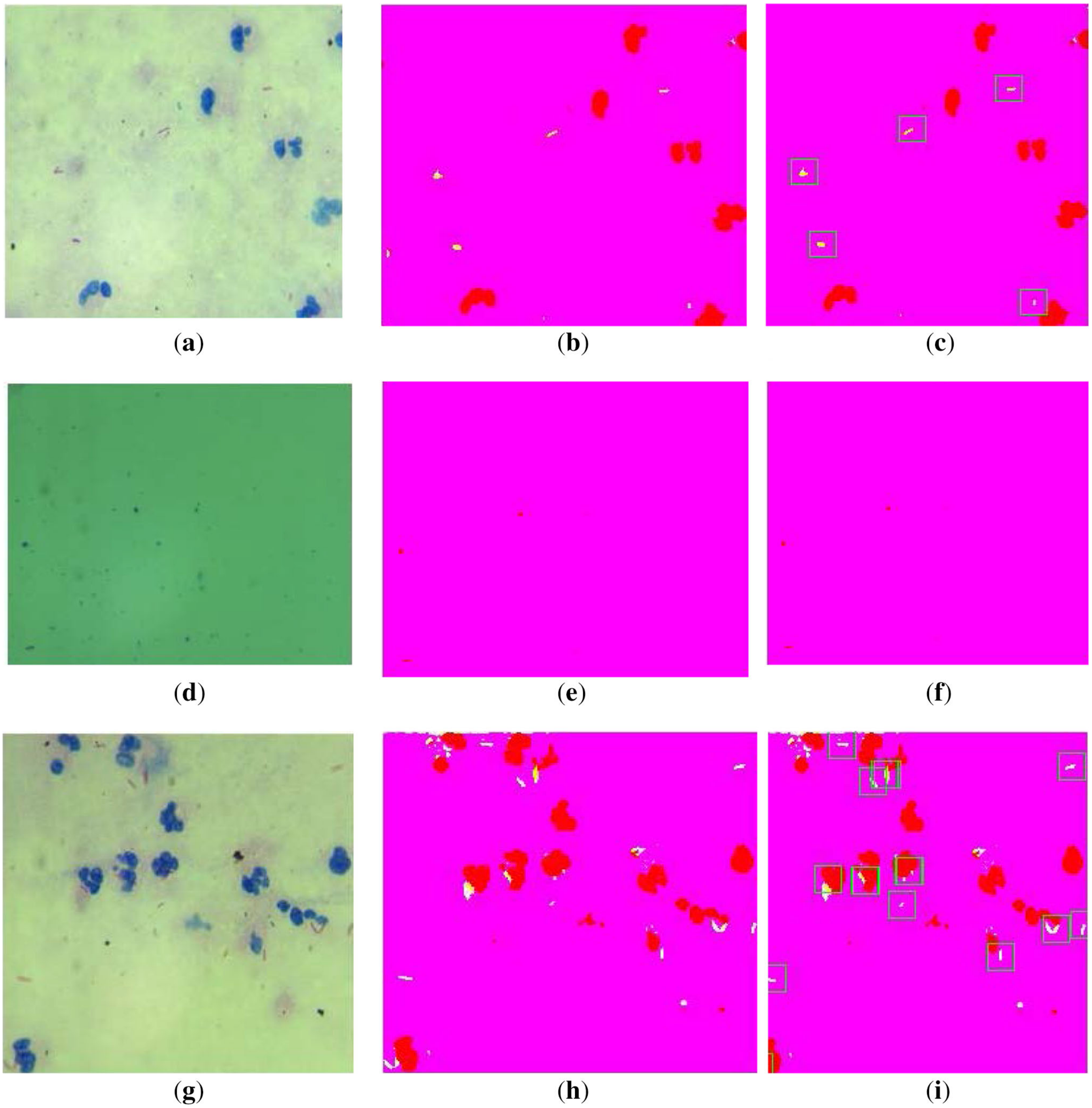
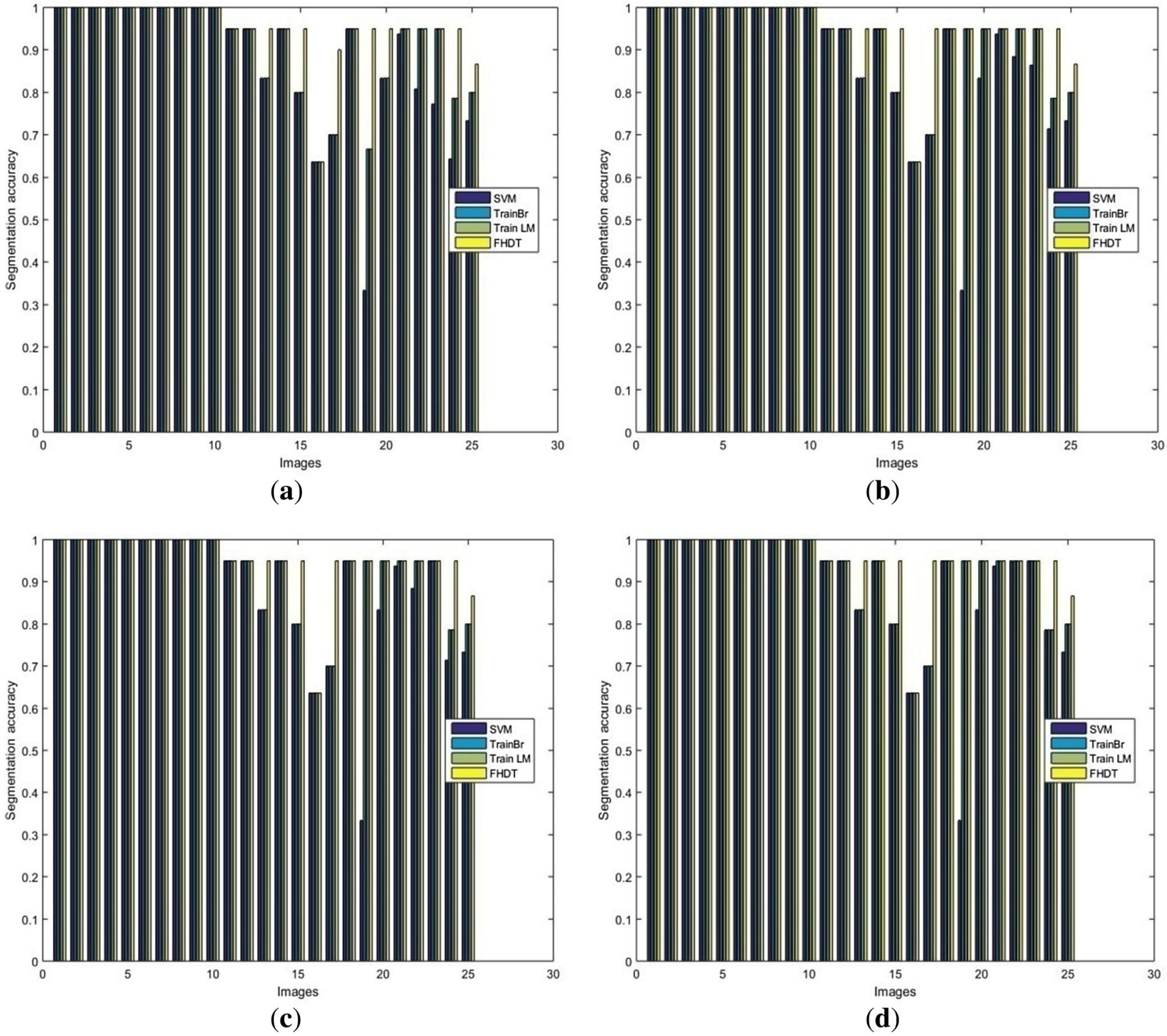


Figure 5. Segmentation results. **(a)** Input few-bacilli image, **(b)** CIELuvcolor space of figure 5a, **(c)** Segmented output of figure 5a, **(d)** Input no-bacilli image, **(e)** CIELuvcolor space of figure 5d, **(f)** Segmented output of figure 5d, **(g)** Input overlapping image, **(h)** CIELuvcolor space of figure 5g, and **(i)** Segmented output of figure 5g.

no-bacilli category; the next ten represents the few-bacilli category; and the remaining five images represent the overlapping bacilli category of images. In figure 6a, the analysis based on SA for 50% training data is shown. SA obtained for the non-bacilli images by all the four techniques are 1, that is,100%. When analysing the few-bacilli category, SA of 0.95 is achieved by all the techniques for the first two few-bacilli images. For the third image, that is, image 13 in the graph, SA attained by the existing techniques is 0.833, whereas that in the proposed FHDT is 0.95. For the other seven few-bacilli images, too, FHDT has higher SA values than the existing techniques. Analysing overlapping-bacilli images, that is, images from 21 to 25, maximum SA of 0.95 is attained by the proposed technique,

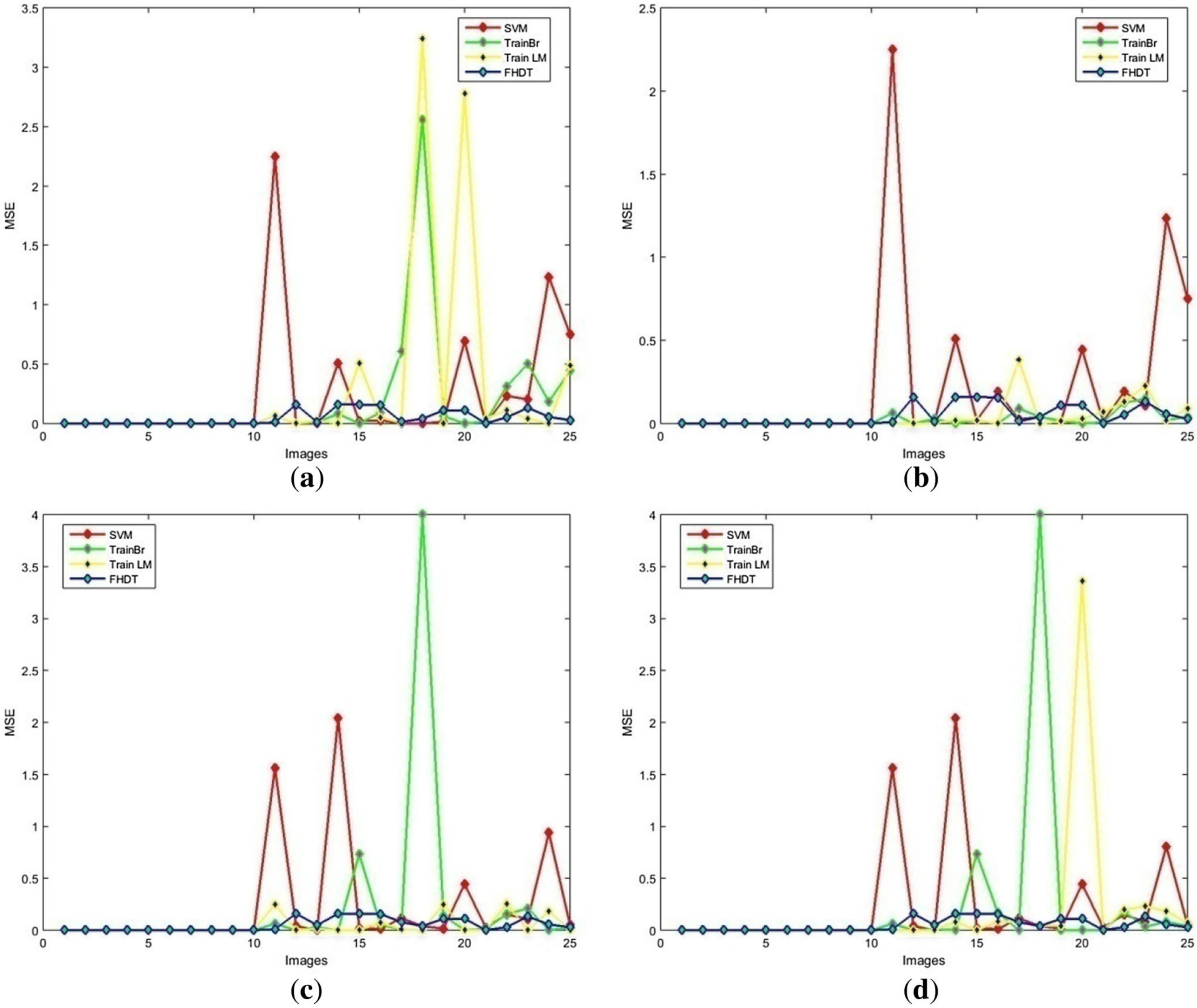


(**a**) (**b**) (**c**) (**d**)

Figure 6. Performance analysis based on SA for **(a)** 50% training data, **(b)** 60% training data, **(c)** 70% training data and **(d)** 80% training data.

while SVM has only 0.938 as the maximum SA value. Figure 6(b) depicts the SA analysis for 60% training data, where the maximum accuracy of 1 is achieved by all the techniques compared for no-bacilli images. For the seventh few-bacilli image, that is, image 17, FHDT can attain the maximum SA of 0.95, while the existing SVM, Train Br, and Train LM, can obtain only 0.7 SA. The SA analysis with 70% training data is illistrated in figure 6(c), where SA obtained by the existing techniques is 0.8 for image 15, while the proposed FHDT can achieve 0.95 SA. The graph showing the SA analysis result for 80% training data is illustrated in figure 6(d). As shown in the figure, the proposed FHDT technique can attain SA of 0.95 for most of the cases, whereas the existing techniques can attain the same value rarely. The SA value for image 17, which is a few-bacilli image, produced by the existing techniques is 0.7, when FHDT produced 0.95.

4.5b *Analysis based on MSE*: The performance analysis based on MSE obtained in the comparative techniques is illustrated using figure 7, for various training data percentages. In figure 7(a), the result of analysis based on MSE for 50% training data is presented. For no-bacilli images, the MSE value attained by all the techniques is 0, since the image has no bacilli for segmentation. When the few-bacilli images are analysed, SVM has an MSE of 2.25 for the first few-bacilli image, while Train Br and Train LM are having an MSE of 0.0625. Meanwhile, the proposed FHDT has only 0.01 as the error measured. For the last

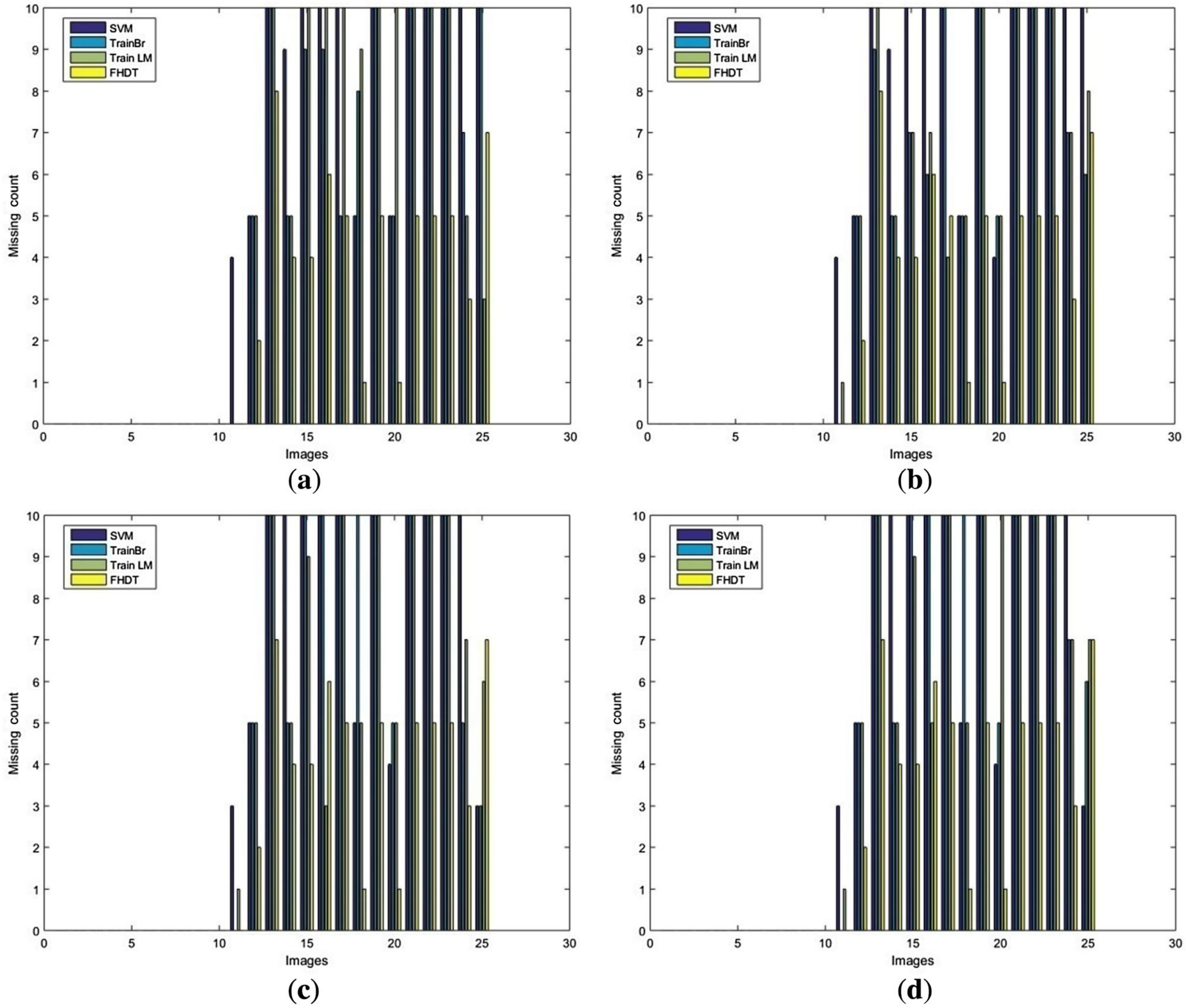


(a) (b) (c) (d)

Figure 7. Performance analysis based on MSE for **(a)** 50% training data**, (b)** 60% training data, **(c)** 70% training data and **(d)** 80% training data.

category, that is, overlapping image, the minimum MSE attained in FHDT is 0.001 for image 21, while in SVM, Train Br and Train LM, it is 0.0013, 0.017 and 0.424, respectively. Figure 7(b) demonstrates the MSE analysis graph for 60% training data is shown. The minimum MSE observed in FHDT is 0.01 for the first few-bacilli image, whereas it is 2.25 and 0.0625 in SVM and Train Br. The MSE analysis for 70% training data is presented in figure 7(c), where a minimum MSE of 0.01 is attained in FHDT for image 11. Meanwhile, MSE value of 1.563, 0.063 and 0.25 are produced in SVM, Train Br and Train LM. The MSE analysed in the comparative techniques for 80% training data is shown in figure 7(d). Here, the MSE value obtained for image 21 is 0.027 by SVM, while the proposed FHDT can obtain 0.001.

4.5c *Analysis based on MC*: In figure 8, the performance analysis based on MC is presented by varying the training data as 50%, 60%, 70% and 80%. Lower the missing count, greater is the performance. Figure 8(a) sketches out the MC analysis graph for 50% training data. The MC value for no-bacilli images is 0 in all the techniques compared. In image 11, which is a few-bacilli image, the MC measured by SVM is 4, while FHDT has 0 MC. For image 15, when SVM, Train Br and Train LM, have MC of 10, 9 and 10, the proposed FHDT technique has only 4 MC. The MC analysis graph for 60% training data is presented in figure 8(b), where the least MC attained by FHDT is 0 for image 11 and the maximum value obtained is 8 for image 13. Meanwhile, the existing SVM approach has MC of 4 and 10. This clearly shows the effectiveness of the proposed technique. In figure 8(c), MSE analysis plot for 70% training data is shown. For image 11, when SVM and Train LM have MC of 3 and 1, FHDT has 0 MC. The value is just 5 for image 21 in the proposed FHDT, while the existing techniques have MC value of 10. Figure 8(d) demonstrates the



(a) **(b)** **(c)** **(d)**

Figure 8. Performance analysis based on MC for **(a)** 50% training data, **(b)** 60% training data, **(c)** 70% training data and **(d)** 80% training data.

analysis result for 80% training data based on MSE in all the comparative techniques. When the overlapping images are analysed, in image 24, SVM, Train Br and Train LM have MC of 10, 7 and 7, whereas that in FHDT is just 3.

4.6 *Discussion*

Table 1 presents the performance analysis of the proposed technique compared with that of the existing techniques, as discussed in section 4.5.

Table 1 presents the performance comparison obtained by the average values of SA and ME in all the techniques compared. As shown in the table, the values marked in bold shows the performance of the proposed FHDT classifier, indicating that the performance is better than that of existing techniques compared. The mean SA attained by

Table 1. Mean performance comparison.

Training data		SA	MC
50%	SVM	0.873	5.12
	Train Br	0.908	4.52
	Train LM	0.908	4.68
	FHDT	**0.952**	**2.44**
60%	SVM	0.882	5.08
	Train Br	0.924	4.2
	Train LM	0.924	4.16
	FHDT	**0.954**	**2.44**
70%	SVM	0.886	4.8
	Train Br	0.924	4.52
	Train LM	0.924	4.24
	FHDT	**0.954**	**2.4**
80%	SVM	0.892	4.8
	Train Br	0.924	4.72
	Train LM	0.924	4.56
	FHDT	**0.954**	**2.4**

SVM for 50% training data is 0.873, whereas, in FHDT, it is 0.954. MC obtained by the proposed classifier is 2.44, which is also better than that of the existing techniques, as SVM, Train Br and Train LM have 5.12, 4.52 and 4.68, respectively. Thus, from the mean performance comparison, it is observed that the proposed approach has a maximum average SA of 0.954 and 2.4 as the average MC.

5. Conclusion

In this paper, a technique for counting the bacilli is developed using sputum smear microscopic images by proposing FHDT classifier for the diagnosis of TB. The technique is composed of three steps, such as segmentation, feature extraction and classification. Segmentation is done in the input microscopic image based on colour space-based bacilli segmentation followed by Otsu thresholding. In the feature extraction process, features like colour, mean, variance, length, density and area are extracted. The proposed FHDT classifier with the utilisation of the newly designed hyco-entropy function classified the segmented image into one of the three classes, such as bacilli, non-bacilli and overlapping bacilli. The number of bacilli in the overlapping bacilli is counted using the fuzzy logic system for the diagnosis of TB. The performance of FHDT classifier is evaluated using three metrics, namely SA, MSE and MC, using sputum smear microscopic images obtained from ZNSM-iDB and is compared with existing techniques, such as SVM, Train Br and Train LM. From the comparative analysis, the proposed FHDT classifier could attain maximum performance with mean SA of 0.954 and mean MC of 2.4. In future, we will utilise other classifiers for further improving the performance of the TB diagnosis.

References

[1] Davies PDO 2017 Tuberculosis diagnosis, reference module in biomedical sciences. In: *International Encyclopedia of Public Health* (Second Edition), Elsevier, pp. 219–228

[2] Priya E and Srinivasan S 2016 Validation of non-uniform illumination correction techniques in microscopic digital TB images using image sharpness measures. *Int. J. Infect. Dis.* 45(Supplement 1): 406

[3] Bhutia R, Narain K, Devi K R, Singh T S K, Mahanta J 2013 Direct and early detection of Mycobacterium tuberculosis complex and rifampicin resistance from sputum smears. *Int. J. Tuberc. Lung Dis*. 17(2): 258–261

[4] Priya E and Srinivasan S 2016 Automated object and image level classification of TB images using support vector neural network classifier. *Biocybern. Biomed. Eng*. 36(4): 670–678

[5] Sotaquirá M, Rueda L and Narvaez R 2009 Detection and quantification of bacilli and clusters present in sputum smear samples: A novel algorithm for pulmonary tuberculosis diagnosis. In: *Proceedings of International Conference on Digital Image Processing*, pp. 117–121

[6] Makkapati V, Agrawal R and Acharya R 2009 Segmentation and classification of Tuberculosis Bacilli from ZN-stained Sputum Smear Images. In: *Proceedings of 5th Annual IEEE Conference on Automation Science and Engineering*, Bangalore, India, pp. 22–25

[7] World Health Organization 2013 *Global Tuberculosis Report* 2013 Geneva, Switzerland

[8] Selen A and Ekinci M 2014 Random forest-based tuberculosis bacteria classification in images of ZN-stained sputum smear samples. *Signal Image Video Process.* 8(1): 49–61

[9] Zhai Y, Liu Y, Zhou D and Liu S 2010 Automatic identification of Mycobacterium Tuberculosis from ZN-stained sputum smear: Algorithm and system design. In: *Proceedings of the 2010 IEEE International Conference on Robotics and Biometrics*, Tianjin, China

[10] Santiago-Mozos R, Pérez-Cruz F, Madden M G and Artés-Rodríguez A 2014 An automated screening system for tuberculosis. *IEEE J. Biomed. Health Inform*. 18(3): 855–862

[11] Patel B and Douglas T S 2012 Creating a virtual slide map from sputum smear images for region-of-interest localisation in automated microscopy. *Comput. Methods Programs. Biomed.* 108(1): 38–52

[12] Panicker R O, Soman B, Saini G and Rajan J 2016 A review of automatic methods based on image processing techniques for tuberculosis detection from microscopic sputum smear images. *J. Med. Syst.* 40(1): 1–13

[13] Khutlang R, Krishnan S, Whitelaw A and Douglas T S 2009 Detection of tuberculosis in sputum smear images using two one-class classifiers. In: *IEEE International Symposium on Biomedical Imaging*, *From Nano to Macro*, Boston, MA, pp. 1007–1010

[14] Barman S, Chattopadhyay S and Samanta D, Bag S and Show G An efficient fingerprint matching approach based on minutiae to minutiae distance using indexing with effectively lower time complexity. In: *Proceedings of the International Conference on Information Technology* (*ICIT*), pp. 179–183

[15] Singh S, Bag S and Jenamani M 2015 Relative similarity based approach for improving aggregate recommendation diversity. In: *Proceedings of the India Conference* (*INDICON*), pp. 1–6

[16] Priya E and Subramanian S 2015 Separation of overlapping bacilli in microscopic digital TB images. *Biocybern. Biomed. Eng.* 35(2): 87–99

[17] Forero-Vargasa M, Sroubekd F, Alvarez-Borregoc J, Malpicab N, Cristobale G, Santosb A, Alcalaf L, Descof M and Coheng L 2002 Segmentation, autofocusing and signature extraction of tuberculosis sputum images. In: *Proceedings Volume 4788, Photonic Devices and Algorithms for Computing IV.* p. 171

[18] Nayak R, Shenoy V P and Galigekere R R 2010 A new algorithm for automatic assessment of the degree of tb-infection using images of ZN-stained sputum smear. In: *Proceedings of International Conference on Systems in Medicine and Biology*, Kharagpur, India, pp. 16–18

[19] Khutlang R, Krishnan S, Dendere R, Whitelaw A, Veropoulos K, Learmonth G, and Douglas TS 2010 Classification of Mycobacterium tuberculosis in images of ZN-stained sputum smears. *IEEE Trans. Inf. Technol. Biomed.* 14(4)

[20] Sadaphal P, Rao J, Comstock G W and Beg M F 2008 Image processing techniques for identifying Mycobacterium tuberculosis in Ziehl-Neelsen stains. *Int. J. Tuberc. Lung Dis.* 12(5): 579–582

[21] Costa Filho C F F, Levy P C, Xavier C D M, Fujimoto L B M and Costa M G F 2015 Automatic identification of tuberculosis mycobacterium. *Biomed. Eng.* 31(1): 33–43

[22] Makkapati V, Agrawal R and Acharya R 2009 Segmentation and classification of tuberculosis bacilli from ZN-stained sputum smear images. In: *Proceedings of 5th Annual IEEE Conference on Automation Science and Engineering*, pp. 217–220

[23] Bockstaller C, Beauchet S, Manneville V, Amiaud B and Botreau R 2017 A tool to design fuzzy decision trees for sustainability assessment. *Environ. Modell. Softw*. 97: 130–144

[24] Ebtehaj I, Bonakdari H and Zaji A H 2017 A new hybrid decision tree method based on two artificial neural networks for predicting sediment transport in clean pipes. *Alex. Eng. J.*

[25] Kumar A, Hanmandlu M and Gupta H M 2013 Ant colony optimization based fuzzy binary decision tree for bimodal hand knuckle verification system. *Expert Syst. Appl.* 40(2): 439–449

[26] Sergyan S 2008 Color histogram features based image classification in content-based image retrieval systems. In: *Proceedings of 6th International Symposium on Applied Machine Intelligence and Informatics*, Herlany, pp. 221–224

[27] Chacko A M M O and Dhanya P M 2015 A comparative study of different feature extraction techniques for offline Malayalam character recognition. *Comput. Intell. Data Min.* 2: 9–18

[28] Mane V M and Jadhav D V 2016 Holoentropy enabled-decision tree for automatic classification of diabetic retinopathy using retinal fundus images. *Biomed. Eng. Biomed. Tech.* 62(3): 321–332

[29] Lee H M, Chen C M, Chen J M and Jou Y L 2001 An efficient fuzzy classifier with feature selection based on fuzzy entropy. *IEEE Trans. Syst. Man Cybern. Part B* (*Cybernetics*). 31(3): 426–432

[30] Nghiem T P and Cho T H 2009 A fuzzy-based interleaved multi-hop authentication scheme in wireless sensor networks. *J. Parallel Distrib. Comput.* 69(5): 441–450

[31] Ziehl–Neelsen Sputum smear Microscopy image DataBase (ZNSM-iDB) from http://14.139.240.55/znsm/

[32] Costa Filho C F F et al 2015 Automatic identification of tuberculosis mycobacterium, *Res. Biomed. Eng.* 31(1): 33–43

[33] Elveren E and Yumuşak N 2001 Tuberculosis disease diagnosis using artificial neural network trained with genetic algorithm. *J. Med. Syst.* 35(3): 329–332