Water Quality Classification Approach based on Bio-inspired Gray Wolf Optimization

Asmaa Hashem Sweidan1,*, Nashwa El-Bendary2,*, Aboul Ella Hassanien3,5,*, Osman Mohammed Hegazy3, Abd El-karim Mohamed4

1Faculty of Computer Sciences and Information, Fayoum University, Fayoum, Egypt
2Arab Academy for Science, Technology, and Maritime Transport, Cairo, Egypt
3Faculty of Computers and Information, Cairo University, Cairo, Egypt
4Faculty of Science, Cairo University, Fayoum University, Fayoum, Egypt
5Faculty of Computers and Information, BeniSuef University, Egypt
*Scientific Research Group in Egypt, (SRGE), http://www.egyptscience.net

Abstract—This paper presents a bio-inspired optimized classification approach for assessing water quality. As fish liver histopathology is a good biomarker for detecting water pollution, the proposed classification approach uses fish liver microscopic images in order to detect water pollution and determine water quality. The proposed approach includes three phases; preprocessing, feature extraction, and classification phases. Color histogram and Gabor wavelet transform have been utilized for feature extraction phase. The Machine Learning (ML) Support Vector Machines (SVMs) classification algorithm has been employed, along with the bio-inspired Gray Wolf Optimization (GWO) algorithm for optimizing SVMs parameters, in order to classify water pollution degree. Experimental results showed that the average accuracy achieved by the proposed GWO-SVMs classification approach exceeded 95% considering a variety of water pollutants.

Keywords—classification, features extraction, color histogram, Gabor wavelet, fish liver, histopathology, water pollution, support vector machines (SVMs), gray wolf optimization (GWO), microscopic images

I. INTRODUCTION

Water pollution is a challenging issue for water quality management. There are two main methods for assessing and detecting water quality: (1) physical and chemical analysis, and (2) biological monitoring methods. Continuous chemical analysis is a complex and expensive process, and also it provides limited data about the chemical compounds [1], [2]. On the other hand, biological method is to expose changes in water quality. The most familiar criteria used to assess pollution of water related to the health of ecosystems. Biological methods are valuable for detecting pollution. So, fish has been used as biomarker for water quality in order to present kind of detection of biological changes due to exposure to chemical pollutants [3], [4].

In research, a limited number of water quality assessment systems have been proposed based on support vector machines algorithm with optimized parameters. For example, in [5], authors proposed an approach based on employing video images analysis biomonitoring and support vector machines (SVMs) algorithm in water quality assessment. The proposed approach depends on behavioral parameters of fish during an acute toxicity test and the hue, saturation, intensity (HSI) color model as feature extraction. The approach proposed in [5] combined both SVMs and genetic algorithm (GA) to enhance the efficiency and accuracy of water quality assessment, and it achieved an accuracy of 80% . Also, in [6], authors proposed hybrid model of Genetic Algorithm and SVMs for prediction of Chlorophyll in reservoirs. The proposed system extracted features using the feature selection method by GA. Authors used hybrid model of GA-SVMs algorithm to optimize the parameters for enhancing the efficiency of results.

The aim of this paper is to use fish liver as biomarker to detect and classify water pollution. It presents a multi-class content-based image classification system to assess water pollution via investigating and classifying the different fish liver microscopic images. The proposed approach combined color and texture for feature extraction and used SVMs classification along with gray wolf optimizer for classifying water quality.

The rest of this article is organized as follows. Section II describes the different phases of the proposed system. Section III introduces the tested fish liver microscopic images dataset and discusses the obtained experimental results. Finally, Section IV presents conclusions and discusses future work.

II. PRELIMINARIES

A. Color and texture features

Color as a basic feature for image representation is a widely used feature in image classification and retrieval problems. Color histogram is a color descriptor that shows representation of colors distribution in an image. Hence, for digital images, it represents the number of pixels that have colors in each range of color, which extends the color space of each image [7]. The
three color moments (mean, variance, and skewness) [7], [9] are extracted from each region of all the color channels. In this paper, we combined a color histogram and color moment feature vectors. By this combination we have efficient way for representing color distribution in any image.

The three moments (mean, variance and skewness) for color images are calculated as shown in equations (1), (2), and (3) [8]–[10].

\[
x_j = \frac{\sum_{i=1}^{M \times N} x_{i,j}}{M \times N} \quad (1)
\]

\[
\gamma_j = \sqrt{\frac{1}{N \times M} \sum_{i=1}^{N \times M} (x_{i,j} - \bar{x})^2} \quad (2)
\]

\[
S_j = \frac{1}{N \times M} \sum_{i=1}^{N \times M} (x_{i,j} - \bar{x})^3 \quad (3)
\]

Where the size of the colored images \(N \times M\) pixels, \(x_{i,j}\) is the value of image pixel \(i\) of color channel \(j\) (e.g RGB, HSV), \(\bar{x}\) of observations is the mean for each channel \(i = (H, S, a, b)\), \(\gamma_j\) is the standard deviation, and \(S_j\) is the skewness for each channel.

On the other hand, texture is one of the important features to determine objects or regions of interest in the image, and describes the visual patterns that contain important information about the structural arrangements of the surface and its relationship to the surrounding environment [8], [10].

In this paper, we used Gabor filter transform texture feature extraction approach. Gabor filters are a group of wavelets. For a given image \(I(x, y)\) with size \(MN\), with Gabor filters \(g_{m,n}\) the corresponding discrete Gabor wavelet transform is calculated as shown in equation (4) [9], [12].

\[
W_{m,n}(x, y) = \int I(x_1, y_1)g_{m,n}(x-x_1, y-y_1) \, dx_1 \, dy_1 \quad (4)
\]

After applying Gabor transformation on the image with different orientation at different scale, the following array of magnitudes is resulted as shown in equation (5).

\[
E(m, n) = \sum_{x,y} g_{m,n}(x, y) \quad (5)
\]

These magnitudes represent the energy content at different scale and orientation of the image. In order to identify the homogeneous texture, the mean and standard deviation of the magnitude of the transformed coefficients are calculated as shown in equations (6) and (7) [11], [12].

\[
\mu_{mn} = \frac{E(m, n)}{PQ} \quad (6)
\]

\[
\rho_{mn} = \sqrt{\frac{1}{N \times M} (E(m, n) - \mu_{mn})^2} \quad (7)
\]

Where \(P\) and \(Q\) are the image sizes. Feature vector \(f\) (texture representation) is created using \(\mu_{mn}\) and standard deviation \(\rho_{mn}\) as feature components. Four scales and six orientations are used in common implementation and the feature vector is given by equation (8) [12]

\[
f = (\mu_{00}, \rho_{00}, \mu_{01}, \rho_{01}, \ldots, \mu_{35}, \rho_{35}) \quad (8)
\]

Our goal from using texture and color features is to assign the pixels in the original image to a relatively small number of groups, where each group represents a set of pixels that are coherent in their color and local texture properties.

B. Support Vector Machines (SVMs)

Support Vector Machines algorithm is a group of supervised learning models used for classification and regression analysis of high dimensional datasets as well as associated learning algorithms that analyze data and recognize patterns [13], [14]. Moreover, SVMs is a binary class classification method that solves problems by attempts to find the optimal hyperplane separation between classes. It depends on the training cases that are placed on the edge of descriptor class, so-called support vectors, and ignores any other cases. The vectors near the hyperplane are the support vectors [15], [16].

SVMs algorithm is based on finding the hyperplane that gives the largest minimum distance to the training. This distance receives the important name of margin within SVMs. Therefore, the optimal separating hyperplane maximizes the margin of the training data that separates a positive class from a negative class [13], [17].

Given a set of \(n\) input vectors \(x_i\) and outputs \(y_i \in \{-1, +1\}\), one tries to find a weight vector \(w\) and offset \(b\) defining a hyperplane that maximally separates the examples. This can be formalized as the maximize problem in equation (9)

\[
\text{maximize } W(\lambda) = \sum_{j=1}^{n} \lambda_j - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \quad (9)
\]

Subject-to \(\sum_{j=1}^{n} \lambda_j y_j, C \geq \lambda \geq 0\).

Where the coefficients \(\lambda_i\) are non-negative. The \(x_j\) with \(\lambda_i > 0\) are called support vectors. \(C\) is a parameter used to trade off the training accuracy and the model complexity so that a superior generalization capability can be achieved. \(K\) is a kernel function transforms the data into a higher dimensional feature space to make it possible to perform the linear separation.

C. Grey Wolf Optimization

Grey wolf optimizer (GWO) is a new meta-heuristic technique. It can be applied for solving optimized problems and
achieves excellent results [20], [21]. In fact, The GWO mimics the grey wolves’ leadership hierarchy and hunting mechanism.

To simulate the leadership hierarchy, there are four types of grey wolves which are alpha (α), beta (β), delta (δ) and omega (ω). Those four types can be used for simulating the leadership hierarchy. The hunting (optimization) is guided by three wolves (α, β and δ). The ω wolves follow them [21], [22].

During the hunt process, it is known that grey wolves surround their prey. Mathematically, this is modeled by equations (10) and (11) [20], [21]:

\[
\vec{D} = |\vec{\bar{C}}.\vec{X}_p(t) - \vec{X}(t)| \tag{10}
\]

\[
\vec{X}(t + 1) = \vec{X}_p(t) - \vec{A}.\vec{D} \tag{11}
\]

Where \( t \) is the current iteration, \( \vec{A} \) and \( \vec{\bar{C}} \) are coefficient vectors, \( \vec{X}_p \) is the vector of the prey position, and \( \vec{X} \) indicates the vector of the grey wolf position.

\( \vec{A} \) and \( \vec{\bar{C}} \) vectors can be calculated as shown in equations (12) and (13)

\[
\vec{A} = 2\bar{\alpha}.r_1 - \bar{a} \tag{12}
\]

\[
\vec{D} = 2.r_2 \tag{13}
\]

where components of \( \bar{a} \) are linearly decreased from 2 to 0, over the course of iterations and \( r_1, r_2 \) are random vectors in [0, 1].

To mimic the hunting process of grey wolves, assume that the α (the best candidate solution), β and δ have a superior knowledge about the possible position of prey. So, the best obtained three solutions are saved so far and force other search agents (including ω) to update their positions according to the position of the best search agents. To update the grey wolves positions, equations (14), (15), and (16) are being used [21], [22]:

\[
\vec{D}_\alpha = |\vec{\bar{C}}_1.\vec{X}_\alpha - \vec{X}|, \vec{D}_\beta = |\vec{\bar{C}}_2.\vec{X}_\beta - \vec{X}|, \vec{D}_\delta = |\vec{\bar{C}}_3.\vec{X}_\delta - \vec{X}| \tag{14}
\]

\[
\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1.(\vec{D}_\alpha), \vec{X}_2 = \vec{X}_\beta - \vec{A}_2.(\vec{D}_\beta), \vec{X}_3 = \vec{X}_\delta - \vec{A}_3.(\vec{D}_\delta) \tag{15}
\]

\[
\vec{X}(t + 1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \tag{16}
\]

### III. THE PROPOSED GWO-SVMs OPTIMIZED CLASSIFICATION APPROACH

As Tilipia is pollution resistant species, they are superlative to be used as bio-indicator for water pollution. The datasets used for experiments were constructed based on real sample images for fish liver, in different histopathological stages, exposed to copper and water pH. The datasets contain colored JPEG images as 125 images and 45 images were used as training and datasets, respectively. Training dataset is divided into 4 classes representing the different histopathological changes and water quality degrees. The proposed approach utilizes texture and color feature extraction methods in addition to SVMs machine learning algorithm, combined with grey wolf optimizer (GWO-SVMs) for parameter optimization. The proposed approach includes three main phases; namely, preprocessing phase, feature extraction phase, and classification phase, as depicted in figure 1.

#### A. Pre-processing phase

During this phase, the proposed approach prepares images for the features extraction phase. So, input microscopic images are resized, image background are removed to get region of interest (RoI), and also images are converted from RGB to HSV color space, and Gabor filters are applied to gray level images.

#### B. Feature extraction phase

In this phase, the proposed approach applies color features (color histogram and color moments) and texture features (Gabor transform) for fish liver microscopic images. The PCA algorithm is used to transform the input space into sub-spaces for dimensionality reduction. After feature extraction, the input space is transformed into sub-spaces for dimensionality reduction and 1D 8x2x2 HSV histogram is calculated, 8 levels for hue and 2 level for each of saturation and value. In addition, nine color moments, three for each channel (H, S and V channels), are calculated. Texture representation is produced using Gabor filters on each image with 4 scales and 6 orientations and an array of magnitudes is obtained. The mean (\( \mu_{mn} \)) and standard deviation (\( \rho_{mn} \)) are calculated as the feature components. Scales and orientations are used in common implementation and the feature vectors are calculated by equation (17) [23].

\[
f = (\mu_{00}, \mu_{03}, \mu_{01}, \rho_{01}, \ldots, \mu_{35}, \rho_{35}) \tag{17}
\]

Then, feature vector will be a combination of HSV 1D histogram, the nine color moments, and Gabor texture.

#### C. Classification phase

In this phase, the SVMs classifier is used for classification of extracted feature vectors via classifying input images using a trained model. This phase employs retrieving all the best matching images from the matching class of the input image using SVMs, whereas the outputs are the corresponding water pollution degree equivalent to each image in the testing dataset. The proposed approach developed SVMs model by using the GWO to optimize the previous input parameters. The SVMs most common kernel functions are, Linear, Polynomial, Multilayer Perception (MLP), and Radial Basis Function (RBF). For optimizing their parameters, the following givens should be considered:

- Linear kernel function does not have any parameters to be optimized
The proposed GWO-SVMs approach developed SVMs model by using the gray wolf optimization (GWO) in order to optimize the previously stated input parameters. Eventually, they should converge on good, possibly optimal positions. In this research one-against-all approach has been applied to solve multi-class problem. Finally, the system determines the water quality based on the images database. Algorithm (1) shows the details of the SVMs classification approach based on the GWO optimizer. The GWO algorithm makes iterations of investigation of new regions and exploiting solutions until reaching near-optimal solution. The goal of this optimization is to find the best areas of space complex search through the interaction between individuals in the population. Furthermore, the gray wolf optimization proves too much durability against configuration compared to Practical Swarm and Genetic Algorithm optimizers.

Algorithm 1 GWO-SVMs classification approach
1: Initialize population
2: Evaluate the fitness function
   2.1 Train SVMs classifiers using cross validation for each kernel function with its parameters
   2.2 Calculate the accuracy rate of cross validation
   2.3 Put the best accuracy in C population as “accuracy”
   2.4 If accuracy satisfies the termination condition, go to step3 otherwise update population and go to step 2.1
3: Update the SVMs parameter and Go to step 2
4: Select the highest accuracy of SVMs with cross validation
5: Classify the water pollution degree

IV. EXPERIMENTAL ANALYSIS AND DISCUSSION

Nile Tilapia “Oreochromis niloticus” is superlative to be used as bio-indicator for water pollution. The datasets used for experiments were constructed based on real sample microscopic images for fish liver in different histopathological change stages exposed to copper and water pH. Fish images were collected from Abbassa farm, Abo-Hammad, Sharkia Governorate, Egypt. Training dataset is divided into 4 classes representing the different water quality degrees; namely excel-
Excellent water quality: where fish hepatocytes are rich in glycogen content, hence the cytoplasmic area of the liver cells displayed a strong Periodic Acid Schiff (PAS) stain of glycogen reactivity reflected by deep magenta coloration, as shown in figure 2.a.

Good water quality: where fish exposed to copper at pH 9 and rich in glycogen content, hence the cytoplasmic area of the liver cells displayed a strong PAS reactivity reflected less than control, as shown in figure 2.b.

Moderate water quality: where fish exposed to copper at pH 7 and rich in glycogen content, hence the cytoplasmic area of the liver cells displayed a strong PAS reactivity reflected by moderate magenta coloration, as shown in figure 2.c.

Bad water quality: where fish exposed to copper at pH 5 and rich in glycogen content, hence the cytoplasmic area of the liver cells displayed a strong PAS reactivity reflected by evident lowering in the intensity of the magenta color, as shown in figure 2.d.

The proposed approach was tested using different initial-ization number. As previously stated, the GWO has been employed to optimize the input parameters in the SVMs and extract feature using combination of color feature (color histogram and color moments) and texture feature (Gabor texture). Moreover, SVMs algorithm was employed with different kernel functions that are: linear, radial basis function (RBF), and multilayer perceptron (MLP) kernel functions for water pollution degree classification. Figure 3 depicts experimental results that show accuracy obtained via applying each kernel function with optimizing input parameters for different number of initialization training images per class. The results of the proposed GWO-SVMs classification approach were evaluated against human expert assessment for measuring obtained accuracy. As shown in figure 3, the proposed GWO-SVMs classification approach achieving accuracy of 95.41 %, 94.43 %, and 88.18 % for linear, RBF, and MLP kernel functions, respectively.

Accordingly, the best accuracy achieved was via applying the proposed GWO-SVMs classification approach with the linear kernel function. Also, as shown in figure 3, the classification accuracy increased for all kernel functions as the number of initialization training images per class increases.

Figure 4 shows comparative analysis for Accuracy of SVMs against GWO-SVMs. The depicted results showed that the proposed GWO-SVMs approach outperformed the typical SVMs algorithm considering all tested kernel functions, via optimizing the input parameters of the SVMs classifier.

V. CONCLUSIONS AND FUTURE WORK

This paper proposes an approach for classifying water pollution degree based on fish liver microscopic images as biomarker. The proposed optimized classification approach used the machine learning SVMs classifier combined with the bio-inspired GWO for parameter optimization. Based on the
obtained results, GWO-SVMs parameter optimization classification approach has achieved an accuracy of 95.41% that outperforms the accuracy achieved by the SVMs classification algorithm. Also, it has been observed that the classification accuracy increased for all kernel functions as the number of initialization training images per class increases. For future work, it is planned to enhance the approach proposed in this paper via considering other bio-inspiring optimization algorithms and different fitness functions to assess water pollution.

REFERENCES


