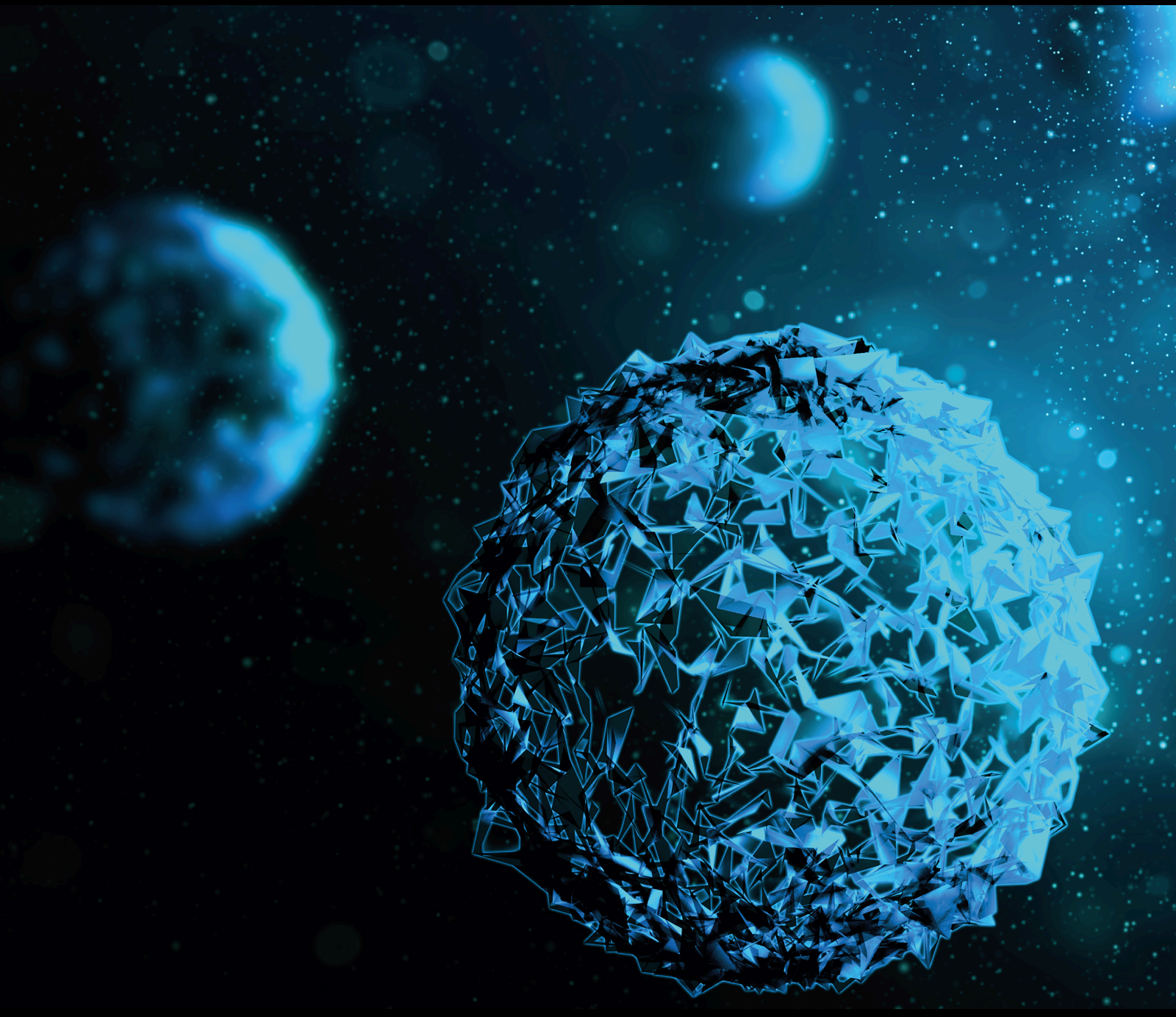
BioMed Research International

Application of Intelligence Methods in Biosciences

Lead Guest Editor: Alireza Baghban
Guest Editors: Fathollah Pourfayaz and Ravinder Kumar



Application of Intelligence Methods in Biosciences

BioMed Research International

Application of Intelligence Methods in Biosciences

Lead Guest Editor: Alireza Baghban
Guest Editors: Fathollah Pourfayaz and Ravinder Kumar

Section Editors

Academic Editors

Bioinformatics

Contents

Hindawi
BioMed Research International
Volume 2022, Article ID 9760864, 1 page
https://doi.org/10.1155/2022/9760864

Corrigendum

Corrigendum to "Estimation of Isentropic Compressibility of Biodiesel Using ELM Strategy: Application in Biofuel Production Processes"

Marischa Elveny,[1] Meysam Hosseini,[2] Tzu-Chia Chen,[3] and S. M. Alizadeh[4]

[1]*Data Science & Computational Intelligence Research Group, Universitas Sumatera Utara, Medan, Indonesia*
[2]*Department of Mathematics, Campus of Bijar, University of Kurdistan, Sanandaj, Kurdistan, Iran*
[3]*CAIC, DPU, Bangkok, Thailand*
[4]*Petroleum Engineering Department, Australian College of Kuwait, West Mishref, Kuwait*

Correspondence should be addressed to Marischa Elveny; marischaelveny@usu.ac.id

Received 4 January 2022; Accepted 4 January 2022; Published 20 May 2022

In the article titled "Estimation of Isentropic Compressibility of Biodiesel Using ELM Strategy: Application in Biofuel Production Processes" [1], Adedoyin Isola Lawal was added to the author list in error. With the agreement of all authors, Adedoyin Isola Lawal is now removed from the author list due to the lack of their contribution to the article. The corrected author list is shown above.

References

[1] M. Elveny, M. Hosseini, T.-C. Chen, A. I. Lawal, and S. M. Alizadeh, "Estimation of Isentropic Compressibility of Biodiesel Using ELM Strategy: Application in Biofuel Production Processes," *BioMed Research International*, vol. 2021, Article ID 7332776, 7 pages, 2021.

Hindawi
BioMed Research International
Volume 2022, Article ID 9761279, 2 pages
https://doi.org/10.1155/2022/9761279

Corrigendum

Corrigendum to "3D-QSAR-Based Pharmacophore Modeling, Virtual Screening, and Molecular Docking Studies for Identification of Tubulin Inhibitors with Potential Anticancer Activity"

Salimeh Mirzaei,[1] Razieh Ghodsi,[2,3] Farzin Hadizadeh,[2,3] and Amirhossein Sahebkar[3,4,5]

[1]*Department of Medicinal Chemistry, Faculty of Pharmacy, Hormozgan University of Medical Sciences, Bandar Abbas, Iran*
[2]*Department of Medicinal Chemistry, School of Pharmacy, Mashhad University of Medical Sciences, Mashhad, Iran*
[3]*Biotechnology Research Center, Pharmaceutical Technology Institute, Mashhad University of Medical Sciences, Mashhad, Iran*
[4]*Applied Biomedical Research Center, Mashhad University of Medical Sciences, Mashhad, Iran*
[5]*School of Pharmacy, Mashhad University of Medical Sciences, Mashhad, Iran*

Correspondence should be addressed to Amirhossein Sahebkar; amir_saheb2000@yahoo.com

Received 10 December 2021; Accepted 10 December 2021; Published 5 January 2022

In the article titled "3D-QSAR-Based Pharmacophore Modeling, Virtual Screening, and Molecular Docking Studies for Identification of Tubulin Inhibitors with Potential Anticancer Activity" [1], the authors identified the errors under Section 3.3 3D-QSAR Contour Map Analysis. The corrected content is shown below.

3.3. 3D-QSAR Contour Map Analysis

To analyze 3D-QSAR results, the model was superimposed on the most active ligand (compound 22 (Figure 6(a)) and the least active ligand (compound 62 (Figure 6(b)). The generated contour plots (Figures 6(c)–6(h)) showed the correlation of structural properties of compounds including electron-withdrawing, hydrophobic, and H-bond donor properties concerning their activity. Blue cubes indicated favorable regions while red cubes indicated unfavorable regions for biological activity [42, 43].

The hydrogen-bond donor nature was compared for the most active compound 22 (Figure 6(c)) and the least active compound 62 (Figure 6(d)). In Figure 6(c), blue cubes were observed at regions lied over the amine group present at position 4 of the quinoline ring. On the other hand, in the least active compound 62 without an amino group at the same steric position (Figure 6(d)), no blue cube was observed in the same region. Therefore, the presence of N-aryl with the hydrogen donor amine group was vital for the cytotoxicity and tubulin inhibitory activity. This assumption was further supported by the low activity of compounds 65-71, which do not have N-aryl at position 4 of the quinoline ring.

Figures 6(e) and 6(f) show the favorable and unfavorable hydrophobic features for the most active compound and least active compound. Figure 6(e) reveals that the blue cubes were generated around the hydrophobic arylstyryl at position 2 and N-aryl at position 4 of the quinoline core were essential for anticancer activity. In Figure 6(f), red cubes were generated at position 4 of the quinoline core of the least active compound. In this compound, a chloro substituent was present at this region instead of the hydrophobic N-aryl group. Thus, the results revealed that red-colored unfavorable regions at these positions could be responsible for the decrease of activity. This was also confirmed by less activity of compounds 65-71 possessing the chloro group at position 4 of the quinoline ring.

In Figure 6(g), blue cubes were observed at the para position of N-aryl indicating the preference of electron-withdrawing groups at this position (the presence of an electronegative atom, such as oxygen or nitrogen, was desirable

because of the inductive electron-withdrawing effect). Also, blue cubes were observed at the para position of the styryl group at position 2 of the quinoline core of the most active compound possessing the electron-withdrawing group (NO_2). It seems that the presence of an electron-withdrawing group at styryl moiety increased the anticancer activity. The high activity of compounds 23-30 and 45-49 possessing NO_2 and F groups at styryl moiety supports this finding.

References

[1] S. Mirzaei, R. Ghodsi, F. Hadizadeh, and A. Sahebkar, "3D-QSAR-based pharmacophore modeling, virtual screening, and molecular docking studies for identification of tubulin inhibitors with potential anticancer activity," *BioMed Research International*, vol. 2021, Article ID 6480804, 20 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 2295920, 11 pages
https://doi.org/10.1155/2021/2295920

Research Article

COVID-19 Diagnosis Using Capsule Network and Fuzzy *C*-Means and Mayfly Optimization Algorithm

Ali Farki,[1] Zahra Salekshahrezaee,[2] Arash Mohammadi Tofigh,[3] Reza Ghanavati,[4] Behdad Arandian,[5] and Amirahmad Chapnevis[6]

[1]*Department of Information Technology Engineering, Industrial and Systems Engineering Faculty, Tarbiat Modares University, Tehran, Iran*
[2]*Florida Atlantic University, College of Engineering and Computer Science, Boca Raton, Florida 33431, USA*
[3]*Department of General Surgery, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
[4]*Department of Chemical and Petroleum Engineering, Sharif University of Technology, Tehran, Iran*
[5]*Department of Electrical Engineering, Dolatabad Branch, Islamic Azad University, Isfahan, Iran*
[6]*Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran*

Correspondence should be addressed to Behdad Arandian; b.arandian@iauda.ac.ir

Received 23 July 2021; Accepted 22 September 2021; Published 19 October 2021

Academic Editor: Paul Harrison

The COVID-19 epidemic is spreading day by day. Early diagnosis of this disease is essential to provide effective preventive and therapeutic measures. This process can be used by a computer-aided methodology to improve accuracy. In this study, a new and optimal method has been utilized for the diagnosis of COVID-19. Here, a method based on fuzzy *C*-ordered means (FCOM) along with an improved version of the enhanced capsule network (ECN) has been proposed for this purpose. The proposed ECN method is improved based on mayfly optimization (MFO) algorithm. The suggested technique is then implemented on the chest X-ray COVID-19 images from publicly available datasets. Simulation results are assessed by considering a comparison with some state-of-the-art methods, including FOMPA, MID, and 4S-DT. The results show that the proposed method with 97.08% accuracy and 97.29% precision provides the highest accuracy and reliability compared with the other studied methods. Moreover, the results show that the proposed method with a 97.1% sensitivity rate has the highest ratio. And finally, the proposed method with a 97.47% *F*1-score rate gives the uppermost value compared to the others.

1. Introduction

In recent decades, several new diseases have emerged in different geographical areas with pathogens including the Ebola virus, Zika virus, NIPA virus, and coronaviruses. Recently, a new type of pathological infection has emerged in Wuhan, China. The new strain is severe acute respiratory syndrome 2 (SARS-CoV-2), which causes Coronavirus Disease 2019 (COVID-19).

Following the increase in the number of patients, the Chinese public clinical and scientific associations reacted quickly to allow the new virus to be identified promptly, and the viral gene sequence to be identified and distributed to other countries around the world. Following extensive research on January 30, 2020, the World Health Organization (WHO) declared the prevalence of public health emergencies to be an international concern [1].

By increasing the extension of this disease, researchers have worked on different methods for early detection of this case at least for minimizing the outbreak. People with suspected COVID-19 should determine as soon as possible if they are infected [2]. Therefore, they should quarantine themselves, receive medical treatment, and inform and warn their relatives. One of the most popular and less harmful imaging methods for diagnosis of this area is chest X-ray imaging. Chest X-ray images are images that use small doses of ionizing radiation to take pictures of the inside of the body called radiographs [3]. The X-rays can help physicians

in different cases such as bone fractures, dislocations, or joint inflammation, abdominal pain, and also some cancer cases [4]. Based on this fact, lots of researchers go through work on using chest X-ray imaging for the diagnosis of COVID-19.

Pereira et al. [5] proposed a method for the identification of COVID-19 in chest X-ray images using hierarchical and flat classification scenarios. They used hierarchical and multiclass learners for disease identification. COVID-19 texture was also explored from the chest X-ray images of pneumonia.

Minaee et al. [6] proposed another method for COVID-19 detection from radiology images. The method was performed on chest X-ray images from publicly available datasets. The images were injected to train four general convolutional neural networks, containing ResNet50, ResNet18, SqueezeNet, and DenseNet-121 for COVID-19 diagnosis in chest X-ray images. The method assessed the models on the residual images, and most of the networks presented high sensitivity and specificity ratios. Even though the efficiency was so hopeful, they presented that more examinations were needed to provide a more consistent estimation.

Rasheed et al. [7] proposed a different methodology for the diagnosis of COVID-19 from chest X-ray images. They used two widely used classifiers including logistic regression (LR) and convolutional neural networks (CNN). They also used the principal component analysis (PCA) to decrease the complexity of the system and to increase the speed of the system. They utilized an online available dataset incorporating GAN to have 500 X-ray images. The final results showed high accuracy for the proposed system.

Elaziz et al. [8] proposed another diagnosis system for the detection of COVID-19 from chest X-ray images. They extracted features from the input images based on fractional multichannel exponent moments (FrMEMs). Then, parallel computing was used for speeding up the system. Then, a modified manta-ray foraging optimization algorithm was used to select the main features. The method was assessed based on the COVID-19 X-ray dataset. The proposed method provided high accuracy for the datasets.

Rehman et al. [9] proposed another computer-aided method for fast detection of COVID-19 based on a convolutional neural network (CNN). The method used the residual neural network (ResNet50) for the purpose. For validating the proposed method, it is performed on a dataset containing X-ray images. Simulation results showed about 98% accuracy for the proposed method.

As can be observed from the literature, different techniques based on chest X-ray images have been proposed for the proper diagnosis of COVID-19. The main objective of this paper is to present another optimal methodology to provide a diagnosis system with higher accuracy. The key purpose is to deliver a new precise computer-aided diagnostic system for COVID-19 diagnosis to help physicians for the detection of COVID-19. Here, we utilized a new optimal machine learning-based approach for the computer-aided diagnosis of COVID-19. This study employs an optimal configuration for proper diagnosis of COVID-19 based on an optimized fuzzy C-ordered means (FCOM) and an improved version of the enhanced capsule network (ECN). The ECN is modified by using the mayfly optimization (MFO) algorithm. The final results show well results for the proposed method.

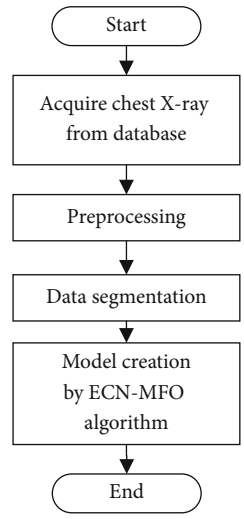


Figure 1: The graphical abstract of the proposed method.

2. Database Preprocessing

The present study proposed a new method based on a clustering approach to provide a proper automatic segmentation system for COVID-19 diagnosis. More specifically, a new optimized fuzzy C-means method (FCM) has been proposed based on a newly developed version of a new metaheuristic algorithm to offer a system with higher efficiency.

However, the traditional clustering methods such as K-means clustering make partitions, wherein each cluster includes just one pattern (a set including accurate and crisp values), fuzzy clustering spreads this technique more to give or associate the present patterns in an image with the data clusters of the image based on a membership function. The fuzzy C-means contains a "soft clustering" methodology and delivers a precise calculation for the cluster membership and is utilized successfully for image clustering applications, especially in medical imaging.

The proposed developed FCM is then used as a new method for segmentation of the chest X-ray of COVID-19. The system classification has been accomplished according to the enhanced capsule network (ECN). The method is based on using deep learning including multiple stages of receipt of raw data supposed as a beginning stage, and the model classification presentation has achieved at the final stage. The graphical abstract of the system is provided in Figure 1.

2.1. Dataset Description. The method of authentication is proposed based on a standard test case of the COVID-19 dataset. Several datasets have been proposed for the diagnosis of COVID-19. The presented study uses three datasets including a popular resource collected by the Renmin Hospital of Wuhan University and two affiliated hospitals, Sun Yat-sen Memorial Hospital and the Third Affiliated Hospital of the Sun Yat-sen University in Guangzhou with 12 and 76

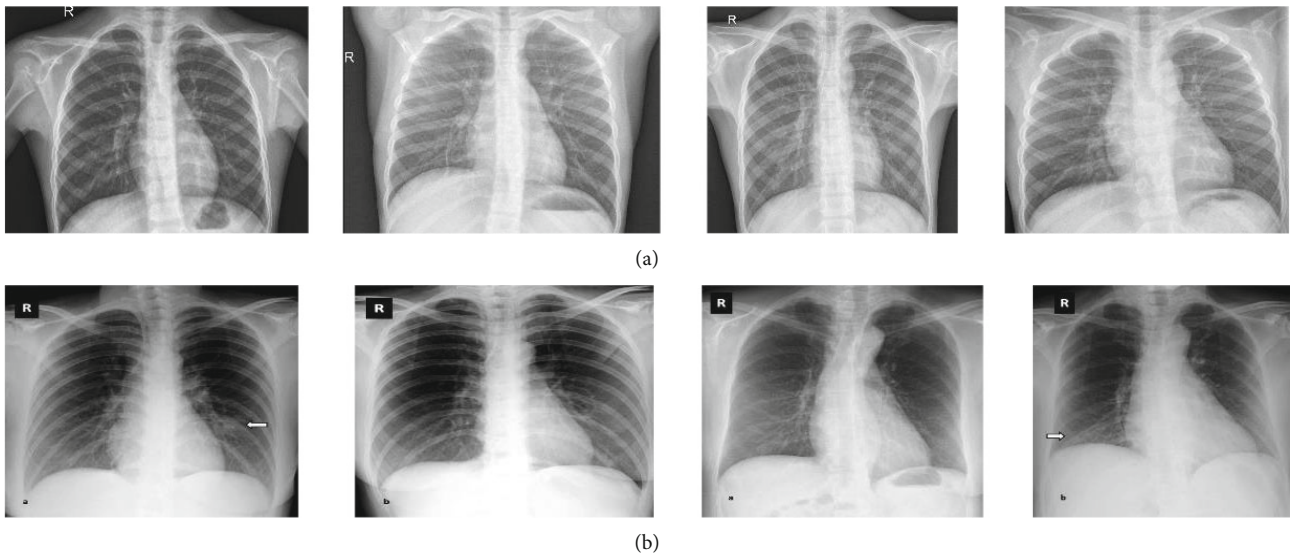


Figure 2: Some examples of the chest X-ray images collected from these datasets: (a) normal case and (b) COVID-19 case.

patients [10]. Figure 2 shows some examples of the chest X-ray images collected from these datasets.

2.2. Data Normalization. Sometimes, the raw data we have for analysis is not suitable for a group of statistical tests and to be able to use this category of statistical tests and to increase the accuracy of the analysis, we have to make changes in the raw data. One of these changes is called data conversion. Data conversion is a mathematical method used to modify variables that do not follow the statistical assumptions of normality, linearity, and uniform scattering, or have patterns with unusual outliers.

Among data conversion methods, data normalization has high efficiency. Normalization has different meanings in statistics, the simplest use of which is to normalize data or normalize variables, and is a method that puts data in the same domain when they are not [11]. In other words, a data miner may encounter situations where the properties of the data include values that are in different ranges or domains. These large-value features may have a much greater effect on the cost function than low-value features. This problem will be solved by normalizing the properties so that their values are in the same range [12]. In constructing a metamodel from the data, before model training begins, the data is subdivided into its largest corresponding values to be normalized to values between zero and one scales, to minimize the effect of the absolute scale and have almost all inputs in the same range.

The min-max method is one of the popular and simplest normalization methods in medical imaging [13]. Based on this method, over and above the unifying data scale, the data changing edges will be distributed in the range between 0 and 1. By assuming attribute X, such that it has a mapping from the data set between $X_{\min}$ and $X_{\max}$, the min-max normalization (X_{norm}) will be achieved by the following [14]:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}. \tag{1}$$

2.3. Contrast Enhancement. Another preprocessing step to improve the image quality is contrast enhancement. The image contrast enhancement, especially in medical images, is an important issue that can be improved the accuracy of the system in a sensible term. This process can increase the contrast between different considered objects to simplify the next segmentation steps. In the present study, contrast enhancement has been used on COVID-19 chest X-ray images to highlight the significant areas with keeping the other areas fixed. The study uses a 16-bit lookup table for this purpose that is then stored on a disc. This technique can be mathematically formulated as follows [15]:

$$B_{\text{hist}} = \frac{A_{\text{hist}} - \text{Min}_{\text{hist}}}{\text{Max}_{\text{hist}} - \text{Min}_{\text{hist}}}, \tag{2}$$

where Min_{hist} and Max_{hist} represent the lowest and the highest levels of the gray magnitudes of the main image histogram, respectively, and A_{hist} and B_{hist} represent the input and the output images before and after contrast enhancement, respectively. Figure 3 shows some examples of the image processing applied to the input images.

3. Mayfly Optimization (MFO) Algorithm

Mayfly is a tiny, fragile, and soft-body insect that has more than 3100 types around the world. However, this insect needs almost one year to birth; it dies after a maximum of 1 day of living. The main target of their birthing is mating. Most of them even do not bother themselves with feeding. The mayfly swarms for mating usually include several males from a few to hundreds of individuals, which are about 1 m to 4 m above the ground for about 1.5 hours to 2 hours in the early morning. Formatting and grabbing the females, males do some nuptial dance over a characteristic up-and-down pattern of movement. Afterward, the couples were released to the vegetation for mating. The mayfly optimization algorithm uses this conception for optimization [16]. This

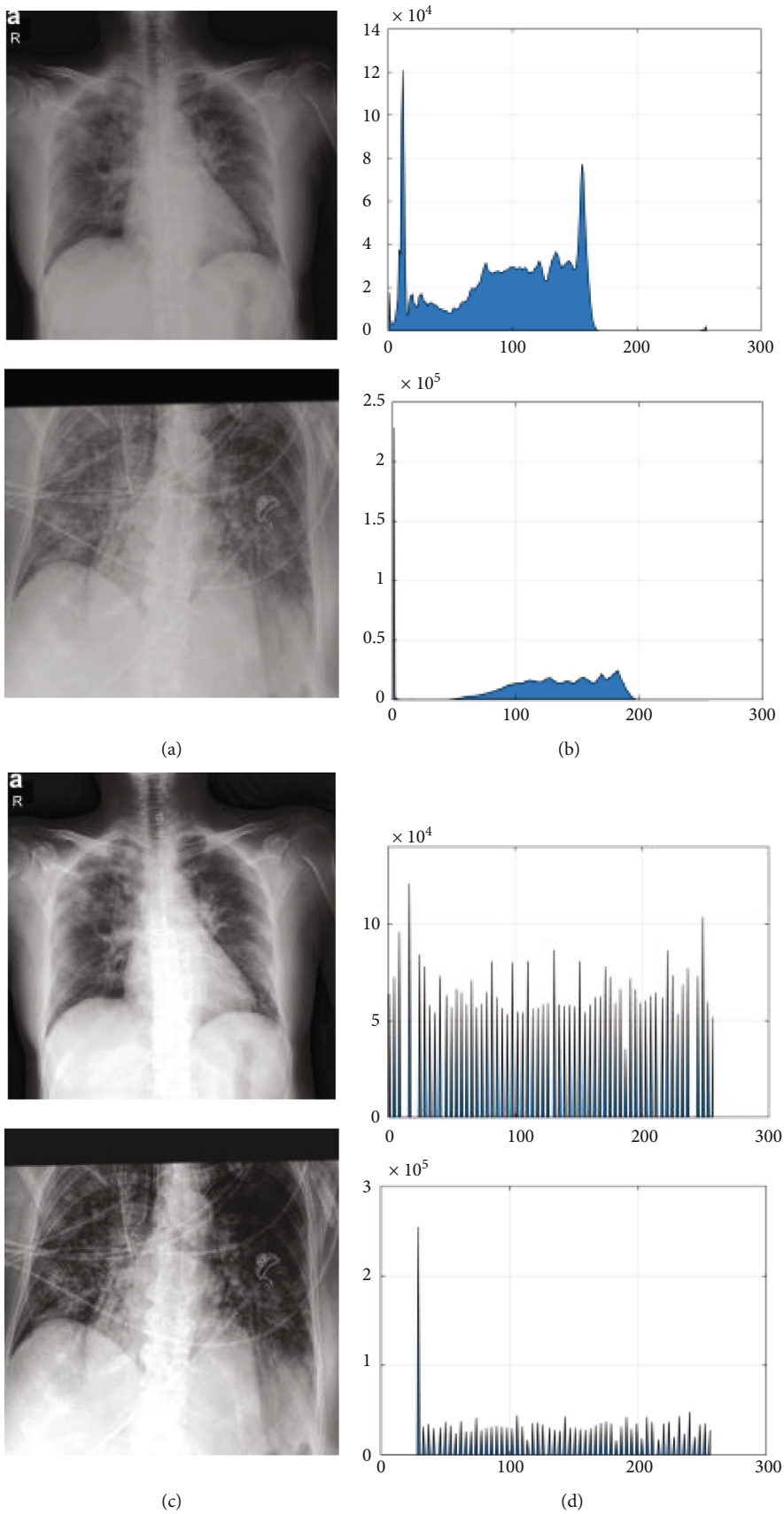


Figure 3: Some examples of the image processing applied to the input images: (a) input image, (b) histogram of (a), (c) image after preprocessing, and (d) histogram of (c).

algorithm uses a hybrid conception of the particle swarm optimization (PSO) algorithm, the genetic algorithm (GA), and the firefly algorithm (FA). Based on the MFO algorithm, the initial population is divided into two classes of male and female mayflies that are generated randomly. The initial population (candidates) is considered a d-dimensional vector $X = [x_1, x_2, \cdots, x_d]^T$ which has been randomly positioned in the solution space. The mayflies have a velocity equal to $V = [v_1, v_2, \cdots, v_d]^T$, and their direction relates to both social and individual flying experiences. Then, the candidates tune their position close to their best position (p_{best}), as well as the best position of the other candidates (g_{best}).

By considering x_i as the present position of the candidate with a step equal t, the updated position is obtained by the following equation [16]:

$$x_i(t+1) = x_i(t) + v_i(t+1), \tag{3}$$

where $x_i(0)$ is limited between $x_{\min}$ and $x_{\max}$.

The movement of mayflies on the top of the water to dance is mathematically modeled as follows:

$$\begin{aligned} v_{ij}(t+1) = v_{ij}(t) &+ a_1 \times \exp\left(-\beta r_p^2\right) \times \left(p\text{best}_{ij} - x_{ij}^t\right) \\ &+ a_2 \times \exp\left(-\beta r_g^2\right) \\ &\times \left(g\text{best}_j - x_{ij}^t\right), j = 1, 2, \cdots, n \end{aligned} \tag{4}$$

where β describes the visibility coefficient utilized for restraining mayfly visibility to others; $p\text{best}_i$ defines the ith best position candidate had ever visited; r_p and r_g describe the Cartesian distance between x_i and $p\text{best}_i$ and x_i and $g\text{best}$; x_{ij}^t and $v_{ij}(t)$ represent the position and the velocity of the ith candidate in dimension j, respectively; and a_1 and a_2 signify the constants for positive attraction scaling the involvement of the social and cognitive component, respectively.

The best position for personally succeeding in the time step $t+1$ is as follows:

$$p\text{best}_i = \begin{cases} x_i(t+1), \text{if } f(x_i(t+1)) < f(p\text{best}_i) \\ \text{is kept the same}, O.W., \end{cases} \tag{5}$$

where $f(.)$ describes the objective function to define the quality of the solution. Then, the global best position ($g\text{best}_j$) is achieved as follows:

$$g\text{best} = \min\{f(p\text{best}_1), f(p\text{best}_2), \cdots, f(p\text{best}_N)\}, \tag{6}$$

where N describes the total number of male candidates in the swarm. The Norm 2 equation has been utilized to determine the r_p and r_g as follows:

$$\begin{aligned} r_p &= \sqrt{\sum_{j=1}^{n}\left(x_{ij} - p\text{best}_i\right)}, \\ r_g &= \sqrt{\sum_{j=1}^{n}\left(x_{ij} - g\text{best}\right)}, \end{aligned} \tag{7}$$

where x_{ij} describes the jth component of the candidate i.

For retaining the algorithm with the best candidates, the best mayflies keep to dance and update their velocities by the following equation:

$$v_{ij}(t+1) = v_{ij}(t) + n_d \times \delta, \tag{8}$$

where n_d signifies the nuptial dance coefficient and δ describes a random value in the range $[-1, 1]$.

However, each male mayfly belongs to a special swarm; the females do not belong to groups. They fly around the males for breeding. With assuming $y_i(t)$ as the ith female candidate path, in the solution space, the position has been updated by the following equation:

$$y_i(t+1) = y_i(t) + v_i(t+1). \tag{9}$$

The best male breeds with the best female, the second-best male with the second-best female, etc. Therefore, the velocity has been considered as follows:

$$v_{ij}(t+1) = \begin{cases} v_{ij}(t) + a_2 \times \exp\left(-\beta r_{mf}^2\right) \times \left(x_{ij}^t - y_{ij}^t\right), \text{if } f(y_i) > f(x_i), \\ v_{ij}^t(t) + r_w \times r, f(y_i) \le f(x_i), \end{cases} \tag{10}$$

where β represents a fixed visibility coefficient, a_2 describes a positive attraction constant, r_{mf} describes the Cartesian distance between male and female candidates, y_{ij}^t and $v_{ij}^t(t)$ signify the ith female candidate position and velocity in dimension j at time step t, and r_w defines a random walk coefficient, and r is a random value in the range $[-1, 1]$.

The MFO algorithm uses crossover as the mating process between the male and female candidates such that two candidates are first selected as male and female. The way of selecting the parent is similar to the method of female attraction by males. The new generation of the crossover process has been achieved as follows:

$$\begin{aligned} \text{offspring}_1 &= \zeta \times \text{male} + (1-\gamma) \times \text{female}, \\ \text{offspring}_2 &= \zeta \times \text{female} + (1-\gamma) \times \text{male}, \end{aligned} \tag{11}$$

where ζ describes a random value and male and female describe the parents. And the early velocity of the offspring is set at zero.

The main reason for using the mayfly optimization algorithm is that it combines the major advantages of swarm intelligence and evolutionary algorithms, which makes it

stronger in providing a good balance between exploration and exploitation [17].

3.1. Data Segmentation. The next step after preprocessing of the input images is to segment the pattern data. The present research uses mayfly optimization (MFO) algorithm to develop a new version of the fuzzy C-means technique for optimal clustering (MFO-FCM). The MFO-FCM is also a good tool for noise reduction in this application. By assuming partitioning of a set of data based on this algorithm including N points into C clusters, the best results of the fuzzy C-means are achieved by minimizing the following equation [18]:

$$\min\left\{F=\sum_{j=1}^{c}\sum_{k=1}^{N}\left[a\left(u_{jk}\right)^{m}+b\times\kappa_{jk}\left(t_{jk}\right)^{\eta}\right]L\left(x_k,v_j\right)+\sum_{j=1}^{c}\delta_i\sum_{k=1}^{N}\left(1-t_{jk}\right)^{\eta}\right\}.$$
$$j=1,2,\cdots,c\,;k=1,2,\cdots,N\,;u_{jk}\in[0,1]\,;\sum_{i}^{c}u_{jk}=1. \tag{12}$$

Subject to

$$\sum_{j=1}^{c}u_{jk}=1,0\le u_{jk}\,;t_{jk}\le 1\,;a,b>0\,;m,\eta>1, \tag{13}$$

where κ_{jk} signifies the feature for the jth cluster and the kth the point, m signifies the weight that is set here 2, and a and b describe the effect of the variable on the status of membership and feature, respectively. In the event that $a>b$, membership provides more effect on the data; otherwise, the data reduces the noise effect and $L(x_k,v_j)$ describes a value in the range x_k to the center of the cluster v_j and has been obtained by the following equation:

$$L\left(x_k,v_j\right)=\sqrt{\sum_{l=1}^{p}\left(x_{il}-v_{jl}\right)^2}. \tag{14}$$

As mentioned in Equation (12), the fuzzy C-means minimization is usually performed by the Lagrange multiplier theorem that is performed based on the following equation:

$$\begin{aligned}
u_{jk}&=\sum_{j=1}^{c}\left(\frac{d\left(x_k,v_j\right)}{d\left(x_k,v_j\right)}\right)^{-1/(m-1)},\\
t_{jk}&=\frac{1}{1+\left[\left(b\times\beta_{jk}/\delta_j\right)D\left(x_k,v_j\right)\right]},\\
v_{jl}^{[r]}&=\frac{\left[\sum_{k=1}^{N}\left[a\left(u_{jk}\right)^{m}+b\times\beta_{jk}\times\left(t_{jk}\right)^{\eta}\right]h_{jkl}^{[r]}x_{kl}\right]}{\left[\sum_{k=1}^{N}\left[a\left(u_{jk}\right)^{m}+b\times\beta_{jk}\times\left(t_{jk}\right)^{\eta}\right]h_{jkl}^{[r]}\right]},\\
&k\in[1,N]\,;l\in[1,p]\,;j\in[1,c].
\end{aligned} \tag{15}$$

In this study, the mayfly optimization (MFO) algorithm is used to generate the memberships and possibilities along with using typicality to tune the impact of the outlier. Here, the parameters T, V, and U are updated to minimize the considered function during the iterations. This will be terminated if $\|V^r-V^{r-1}\|\le\varepsilon$.

```
1) Start
2) Initializing parameters of the MFO and the num-
ber of clusters
3) Initializing the cluster centers based on MFO
4) Calculating the objective function
5) Updating the candidate s position using the MFO
algorithm
6) If the termination criteria are satisfied, go
to ((6))
7) Else
      go to ((2)).
8) End
```

Pseudocode 1: The pseudocode of the MFO-FCM

The present study uses within-cluster sum of squares errors (WCSS) as the objective function for the optimization. By considering $S=[s_1,s_2,\cdots,s_N]$ as clusters and $X=[x_1,x_2,\cdots,x_N]$ as data points, the WCSS function is formulated by the following equation:

$$\text{WCSS}=\sum_{j=1}^{k}\sum_{x_j\in s_i}L\left(x_j,v_i\right), \tag{16}$$

where $L(x_j,v_i)$ determines the distance from v_i and x_j, and v_i represents the center of the cluster s_i.

This method contains two parts: the first part is to identify the model parameters and the second part is to centroid clustering. Six variables are considered in the MFO-FCM model. By assuming cluster centroids including $V=[v_1,v_2,\cdots,v_c]$, the centers of the clusters are defined. The pseudocode of the MFO-FCM model is explicated in the following:

3.2. Enhanced Capsule Networks (ECN). The present study uses enhanced capsule networks (ECN) to provide a suitable diagnosis system. In ECN, the fragmented pixel set of the X-ray image is labeled as a set of nerve cells regarding the capsule. The system used a pixel vector as an actuation vector encompassed by an active capsule such that it may be a particular category like healthy or COVID-19 for the X-ray image.

Here, the capsule output and the coupling coefficient have been multiplied by the capsule routing in a layer. The parent capsule resistance for routing defines the value of the coupling coefficient. Based on the *routing-by-agreement* mechanism, the low-level COVID-19 diagnosis has been determined by the high-level capsule activation [19].

With assuming $y_i \in$ [healthy, COVID-19] as the ith output capsule, and we_{ij} as the weight matrix, we have

$$\hat{y}_{(j|i)} = we_{ij} y_i, \tag{17}$$

where $\hat{y}_{(j|i)}$ signifies the detection vector which identifies the output parent capsule j using capsule i.

The value of the weight will be amended if the values will be reduced or the pixels contain probably to the positive class. For going before layer capsules, the softmax function has been utilized and the potential parent capsule as the coefficient is encrypted c_{ij} such that essential logits b_{ij} appears the log past conceivable outcomes of the ith routing capsule within the last layer to the jth capsule within the succeeding layer.

The routing-by-agreement mechanism is mathematically performed by the following equation:

$$c_{ij} = \frac{e^{b_{ij}}}{\sum e^{b_{ij}}}. \tag{18}$$

And the proceeding layer indicates a critical element in the evaluation of input of the parent capsule j as follows:

$$s_j = \sum_i c_{ij} \hat{y}_{(j|i)}. \tag{19}$$

The squashing function has been utilized to accomplish the pixel vector compression in the range $[0, 1)$ as follows:

$$va_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \times \frac{s_j}{\varepsilon + \|s_j\|^2}, \tag{20}$$

where $\varepsilon = 10^{-7}$.

And the capsule in the next layer is formulated by the following equation:

$$a_{ij} = va_j \times \hat{y}_{(j|i)}. \tag{21}$$

The overall classification of the capsules which are considered as individual margin loss (Loss_k) in the categories capsule k for the capsule networks based on the loss is

$$\text{Loss}_k = T_k \max(0, m^+ - \|va_k\|)^2 + \lambda(1 - T_k) \max(0, \|va_k\| - m^-)^2, \tag{22}$$

where T_k describes the instant attendance in category capsule k and m^-, m^+, and λ represent hyperparameter assistances. Finally, the ECN has been trained by 700 iterations using Adam optimizer to get the optimal results of hyperparameters. The learning rate is considered an amount of $1e-6$ [20].

Table 1: Accuracy results based on different methods.

Epochs	Proposed method	FOMPA [21]	MID [22]	4S-DT [23]
100	95.82	93.49	91.08	88.61
200	96.19	94.82	92.46	89.73
300	96.55	94.67	93.32	90.07
400	97.83	95.53	93.95	91.59
500	97.08	95.11	94.70	91.34

Table 2: The precision results based on different methods.

Epochs	Proposed method	FOMPA [21]	MID [22]	4S-DT [23]
100	94.20	90.33	92.64	88.24
200	95.11	91.27	93.29	88.96
300	95.05	91.08	94.08	89.38
400	96.93	92.46	95.13	90.73
500	97.29	93.14	96.42	90.91

Table 3: The sensitivity results based on different methods.

Epochs	Proposed method	FOMPA [21]	MID [22]	4S-DT [23]
100	95.05	92.26	90.43	87.15
200	95.31	93.04	91.38	89.36
300	96.73	94.53	91.07	89.75
400	96.29	95.38	93.40	90.07
500	97.18	95.85	93.66	91.68

Table 4: $F1$-score results based on different methods.

Epochs	Proposed method	FOMPA [21]	MID [22]	4S-DT [23]
100	95.23	92.25	90.05	88.46
200	95.42	93.53	91.42	89.57
300	96.81	94.12	92.16	90.18
400	96.93	95.60	94.57	91.22
500	97.47	96.39	94.18	91.76

4. Results and Discussions

The proposed model is performed on three datasets including a popular resource collected by the Renmin Hospital of Wuhan University and two affiliated hospitals, Sun Yat-sen Memorial Hospital and the Third Affiliated Hospital of the Sun Yat-sen University in Guangzhou with 12 and 76 patients [10]. The method is programmed in MATLAB 2016b 64-bit version and executed on the computation environment of Intel Core i7 CPU 2.00 GHz, 2.5 GHz, 8 GB RAM, and 64-bit operating system. As before mentioned, the key purpose of this study is to design a new CAD-based system for the diagnosis of COVID-19. The model has been analyzed based on four measurement indicators including accuracy, precision, $F1$-score, and sensitivity.

4.1. Accuracy. The precision determines how nearly the measured esteem is to the genuine (real) value. This indicator is accomplished by the proportion of correct identification

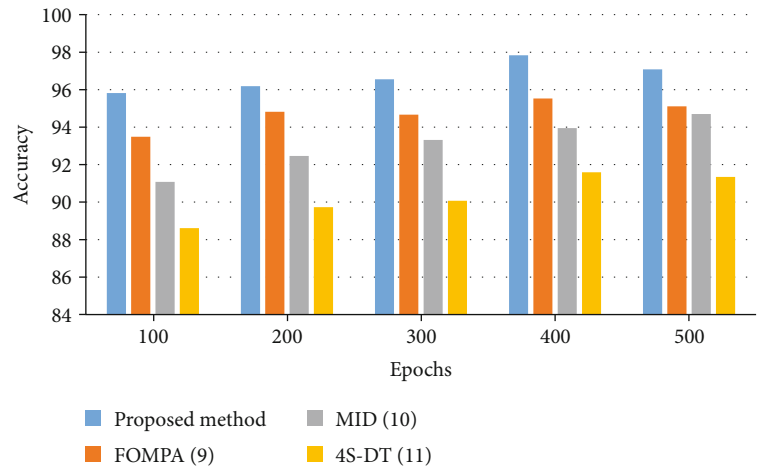


Figure 4: The bar plot of the accuracy results for the studied algorithms.

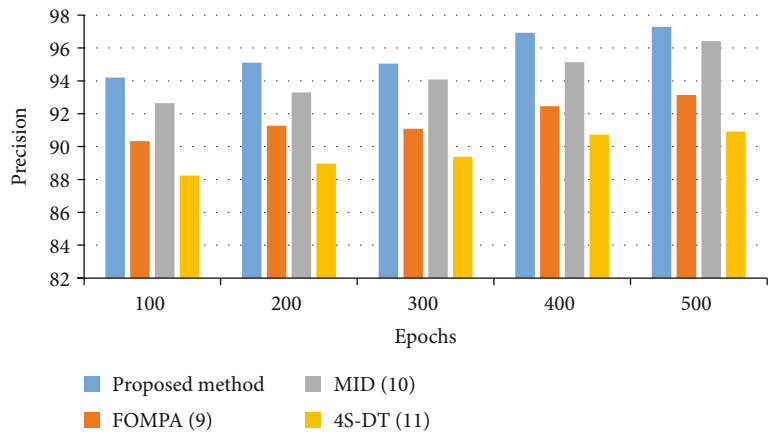


Figure 5: The bar plot of the precision results for the studied algorithms.

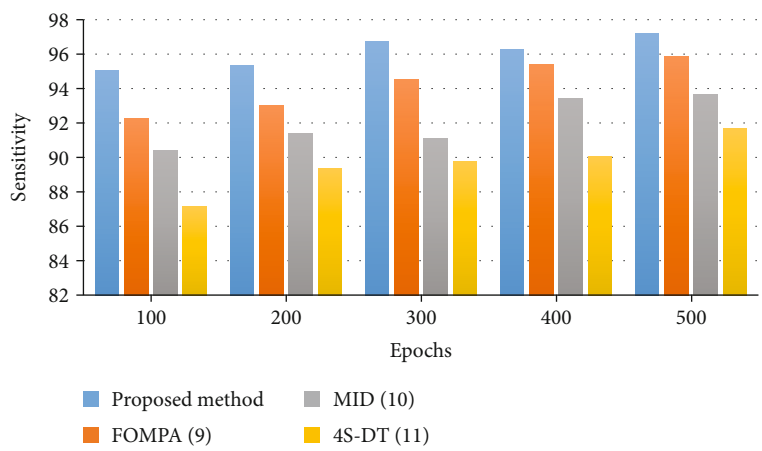


Figure 6: The bar plot of the sensitivity results for the studied algorithms.

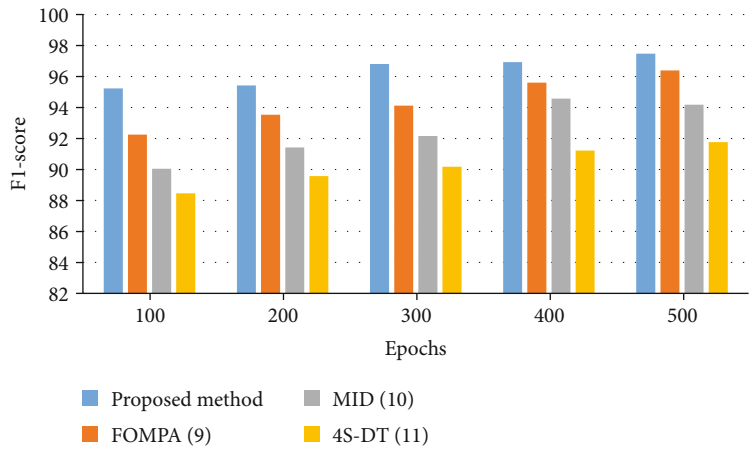


FIGURE 7: The bar plot of the $F1$-score results for the studied algorithms.

values to the whole number of identifications. This can be mathematically described as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^{l}(\text{TP}_i + \text{TN}_i)}{\sum_{i=1}^{l}(\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i)}, \quad (23)$$

where TN and FN describe the true negative and false negative, respectively, and TP and FP represent the true positive and false positive, respectively.

4.2. Precision. Precision defines how near the measured values are to each other. This indicator is accomplished by the proportion of positive identification values to the whole number of identifications. This can be mathematically described as follows:

$$\text{Precision} = \frac{\sum_{i=1}^{l}(\text{TP}_i + \text{FP}_i)}{\sum_{i=1}^{l}(\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i)}. \quad (24)$$

4.3. Sensitivity. Sensitivity is the extent of positives that are accurately recognized (i.e., the extent of those who have a few conditions (influenced) who are accurately recognized as having the condition). This indicator is accomplished by the proportion of true identification values to the true positive and false negative number of identifications. This can be mathematically described as follows:

$$\text{Recall} = \frac{\sum_{i=1}^{l}\text{TP}_i}{\sum_{i=1}^{l}(\text{TP}_i + \text{FN}_i)}. \quad (25)$$

4.4. F1-Score. F-score or F-measure could be the degree of a test's exactness. It is achieved based on the precision and sensitivity of the test. The most noteworthy conceivable value of an F-score is 1.0, showing idealized exactness and review, and the least conceivable value is 0, with the chance that either the precision or the sensitivity is zero. The $F1$ -score is additionally known as the Dice similarity coefficient (DSC). This measure is mathematically defined as follows:

$$F1_{\text{score}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (26)$$

The method analysis of the classification based on the offered enhanced capsule networks (ECN) is reported in Tables 1–4. The method has been compared with three other state-of-the-art methods including FOMPA [21], MID [22], and 4S-DT [23] for better clarification.

To offer a better clarification of the efficiency of the proposed method, a bar plot of the results is shown in Figures 4–7. As can be observed from Figure 4, the simulation shows a 97.08% accuracy with 97.29% precision for the suggested methodology compared to the other studied methods. However, FOMPA, MID, and 4S-DT are in the next ranks.

Figure 5 shows the precision results for the studied algorithms.

Figure 6 shows the bar plot of the sensitivity results for the studied algorithms.

The results show that 500 epochs for all algorithms have been utilized. As can be observed from Figure 6, the proposed classifier offers a higher sensitivity rate to the other state-of-the-art. The designed classifier of the proposed method offers a 97.1% sensitivity rate, whereas FOMPA, MID, and 4S-DT have 95.85%, 93.66%, and 91.68%, respectively, for 500 epochs. Figure 7 illustrates the bar plot of the $F1$-score results for the studied algorithms.

Based on Figure 7, after 500 epochs for all algorithms, the proposed method presents the best better $F1$-score rate to the other methods. As can be observed, the suggested method with a 97.47% $F1$-score rate has the highest value, and the FOMPA, MID, and 4S-DT with 96.39%, 94.18%, and 91.76%, respectively, are in the next ranks.

5. Conclusions

COVID-19 was formed in late 2019 and is spreading rapidly across the world. Early diagnosis of COVID-19 can be so beneficial to the treatment of the disease and to prevent its outbreak. Due to the probability of human error among the experts in finding COVID-19, the application of

machine learning has been recently increased as an auxiliary tool. The present study proposed a method based on image processing for the diagnosis of COVID-19. This study presented an optimal configuration for proper diagnosis of COVID-19 based on an optimized fuzzy *C*-ordered means (FCOM) and an improved version of the enhanced capsule network (ECN). The ECN was improved based on the mayfly optimization (MFO) algorithm. The proposed method was then performed on the chest X-ray COVID-19 images from publicly available datasets. The results were analyzed by comparing some other methods, including FOMPA, MID, and 4S-DT, and the results showed the higher effectiveness of the proposed method. As mentioned, the proposed method has good accuracy in terms of theory. However, due to using complicated methods, using it in real-time applications is not feasible. Therefore, in future work, we will work on using a simplified technique of the proposed method to perform on a microprocessor for real-time applications.

Data Availability

The presented study uses three datasets including a popular resource collected by the Renmin Hospital of Wuhan University and two affiliated hospitals, Sun Yat-sen Memorial Hospital and the Third Affiliated Hospital of the Sun Yat-sen University in Guangzhou, that can be achieved by email to the sources.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] R. Ranjbarzadeh, S. Jafarzadeh Ghoushchi, M. Bendechache et al., "Lung infection segmentation for COVID-19 pneumonia based on a cascade convolutional network from CT images," *BioMed Research International*, vol. 2021, Article ID 5544742, 16 pages, 2021.

[2] R. Ranjbarzadeh and S. B. Saadi, "Automated liver and tumor segmentation based on concave and convex points using fuzzy *c*-means and mean shift clustering," *Measurement*, vol. 150, article 107086, 2020.

[3] R. Ranjbarzadeh, A. Bagherian Kasgari, S. Jafarzadeh Ghoushchi, S. Anari, M. Naseri, and M. Bendechache, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," *Scientific Reports*, vol. 11, no. 1, pp. 1–17, 2021.

[4] N. Razmjooy, M. Ashourian, M. Karimifard et al., "Computer-aided diagnosis of skin cancer: a review," *Current Medical Imaging*, vol. 16, no. 7, pp. 781–793, 2020.

[5] R. M. Pereira, D. Bertolini, L. O. Teixeira, C. N. Silla Jr., and Y. M. G. Costa, "COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios," *Computer Methods and Programs in Biomedicine*, vol. 194, article 105532, 2020.

[6] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Jamalipour Soufi, "Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning," *Medical Image Analysis*, vol. 65, article 101794, 2020.

[7] J. Rasheed, A. A. Hameed, C. Djeddi, A. Jamil, and F. al-Turjman, "A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, no. 1, pp. 103–117, 2021.

[8] M. A. Elaziz, K. M. Hosny, A. Salah, M. M. Darwish, S. Lu, and A. T. Sahlol, "New machine learning method for image-based diagnosis of COVID-19," *PLoS One*, vol. 15, no. 6, article e0235187, 2020.

[9] A. Rehman, T. Sadad, T. Saba, A. Hussain, and U. Tariq, "Real-time diagnosis system of COVID-19 using X-ray images and deep learning," *IT Professional*, vol. 23, no. 4, pp. 57–62, 2021.

[10] R. Kumar, R. Arora, V. Bansal et al., "Accurate prediction of COVID-19 using chest x-ray images through deep feature learning model with smote and machine learning classifiers," 2020, https://www.medrxiv.org/content/10.1101/2020.04.13.20063461v1.

[11] N. Razmjooy and S. Razmjooy, "Skin melanoma segmentation using neural networks optimized by quantum invasive weed optimization algorithm," in *Metaheuristics and Optimization in Computer and Electrical Engineering*, pp. 233–250, Springer, 2021.

[12] M. Ramezani, D. Bahmanyar, and N. Razmjooy, "A new improved model of marine predator algorithm for optimization problems," *Arabian Journal for Science and Engineering*, vol. 696, pp. 1–24, 2021.

[13] S. Nath Datta, "A review of the min-max approach to the solution of relativistic electron wave equation," 2017, https://arxiv.org/abs/1709.07061.

[14] M. Kociołek, M. Strzelecki, and R. Obuchowicz, "Does image normalization and intensity resolution impact texture classification?," *Computerized Medical Imaging and Graphics*, vol. 81, article 101716, 2020.

[15] B. Yu, L. Zhou, L. Wang et al., "Learning sample-adaptive intensity lookup table for brain tumor segmentation," in *23rd International Conference on Medical Image Computing and Computer Assisted Intervention*, Lima, Peru, 2020.

[16] K. Zervoudakis and S. Tsafarakis, "A mayfly optimization algorithm," *Computers & Industrial Engineering*, vol. 145, article 106559, 2020.

[17] N. Razmjooy, M. Khalilpour, and M. Ramezani, "A new meta-heuristic optimization algorithm inspired by FIFA world cup competitions: theory and its application in PID designing for AVR system," *Journal of Control, Automation and Electrical Systems*, vol. 27, no. 4, pp. 419–440, 2016.

[18] N. Dhanachandra and Y. J. Chanu, "An image segmentation approach based on fuzzy c-means and dynamic particle swarm optimization algorithm," *Multimedia tools and applications*, vol. 79, no. 25-26, pp. 18839–18858, 2020.

[19] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules. in Advances in neural information processing systems," *Advances in neural information processing systems*, pp. 3856–3866, 2017.

[20] S. Mehta, C. Paunwala, and B. Vaidya, "CNN based traffic sign classification using Adam optimizer," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 1293–1298, France, 2019.

[21] A. T. Sahlol, D. Yousri, A. A. Ewees, M. A. A. al-qaness, R. Damasevicius, and M. A. Elaziz, "COVID-19 image classification using deep features and fractional-order marine predators algorithm," *Scientific Reports*, vol. 10, no. 1, pp. 1–15, 2020.

[22] M. J. Horry, S. Chakraborty, M. Paul et al., "COVID-19 detection through transfer learning using multimodal imaging data," *IEEE Access*, vol. 8, pp. 149808–149824, 2020.

[23] A. Abbas, M. M. Abdelsamea, and M. Gaber, "4S-DT: self supervised super sample decomposition for transfer learning with application to COVID-19 detection," 2020, https://arxiv.org/abs/2007.11450.

Hindawi
BioMed Research International
Volume 2021, Article ID 3440498, 15 pages
https://doi.org/10.1155/2021/3440498

Research Article

Identification of Key Exosome Gene Signature in Mediating Coronary Heart Disease by Weighted Gene Correlation Network Analysis

Yanbin Fu,[1] Yanzhi Ge,[2] Jianfeng Cao,[3] Zedazhong Su,[1] and Danqing Yu[1,3]

[1]*School of Medicine, South China University of Technology, Guangzhou, China*
[2]*The First Affiliated Hospital, Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China*
[3]*Department of Cardiology, Guangdong Cardiovascular Institute, Guangdong Provincial Key Laboratory of Coronary Heart Disease Prevention, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China*

Correspondence should be addressed to Zedazhong Su; 673946537@qq.com and Danqing Yu; yudanqing2017@126.com

Received 5 August 2021; Revised 7 September 2021; Accepted 18 September 2021; Published 15 October 2021

Academic Editor: Alireza Baghban

Background. Coronary heart disease (CHD) is the most prevalent disease with an unelucidated pathogenetic mechanism and is mediated by complex molecular interactions of exosomes. Here, we aimed to identify differentially expressed exosome genes for the disease development and prognosis of CHD. *Method*. Six CHD samples and 32 normal samples were downloaded from the exoRbase database to identify the candidate genes in the CHD. The differentially expressed genes (DEGs) were identified. And then, weighted gene correlation network analysis (WGCNA) was used to investigate the modules in coexpressed genes between CHD samples and normal samples. DEGs and the module of the WGCNA were intersected to obtain the most relevant exosome genes. After that, the function enrichment analyses and protein-protein interaction network (PPI) were performed for the particular module using STRING and Cytoscape software. Finally, the CIBERSORT algorithm was used to analyze the immune infiltration of exosome genes between CHD samples and normal samples. *Result*. We obtain a total of 715 overlapping exosome genes located at the intersection of the DEGs and key modules. The Gene Ontology enrichment of DEGs in the blue module included inflammatory response, neutrophil degranulation, and activation of CHD. In addition, protein-protein networks were constructed, and hub genes were identified, such as LYZ, CAMP, HP, ORM1, and LTF. The immune infiltration profiles varied significantly between normal controls and CHD. Finally, we found that mast cells activated and eosinophils had a positive correlation. B cell memory had a significant negative correlation with B cell naive. Besides, neutrophils and mast cells were significantly increased in CHD patients. *Conclusion*. The underlying mechanism may be related to neutrophil degranulation and the immune response. The hub genes and the difference in immune infiltration identified in the present study may provide new insights into the diagnostic and provide candidate targets for CHD.

1. Introduction

Coronary heart disease (CHD) is a collective term for disease in which the wall of the coronary arteries becomes narrowed due to fatty material accumulation [1, 2]. As the most common heart disease worldwide, it is estimated that around 200 million people suffer from CHD [3]. With the aging of society, an increasing number of CHD patients may be seen in the future. The poor prognosis of CHD seriously affects the quality of life of patients and brings a heavy burden to society [4]. Therefore, this study mainly explores the hub genes of exosomes in CHD and its regulatory function.

Exosomes are nanometer-sized vesicles (30-150 nm in diameter) secreted by most cells through exocytosis. They are encapsulated in a lipid bilayer and carry a variety of biomolecules, such as proteins, glycan, lipids, metabolites, RNA, and DNA [5]. Exosomes play a critical aspect in several pathological diseases, including cardiovascular disease [6, 7] and acute and chronic inflammation [8]. exoRBase is an exosome library derived from RNA-seq data analysis of

human blood exosomes, including experimental verification from published literature [9]. It helps researchers identify molecular features in blood exosomes and triggers the discovery of new circulating biomarkers and functional implications for human diseases [10]. An important pathological feature of CHD is atherosclerosis. Exosome-mediated intercellular signaling may be a significant aspect of atherosclerotic plaque formation, affecting not only the initiation of CHD but also its progression [11]. Therefore, exosomes may be an ideal biomarker candidate and therapeutic target for CHD.

Recently, the focus has mainly shifted to screening DEGs but not exploring gene interactions [12]. Weighted gene correlation network analysis (WGCNA) is a systematic biology method and widely used to explore the connections between key modules and target disease [13, 14]. The WGCNA method generates a scale-free gene coexpression network based on Pearson's correlation matrix of genes [15]. After constructing the WGCNA network, we observed that similarly expressed genes were in the candidate module. Then, we analyzed the connection between acquired module and DEGs and finally determined the exosome genes with the most significant relation to CHD. Then, based on the genes acquired before, both functional enrichment analysis and protein-protein interaction (PPI) were performed and we revealed the potential transcriptional regulatory network in CHD, aiming to obtain new insights for CHD prevention and therapy. Besides, this is the first time in discovering the relationship between exosome genes and CHD by merged bioinformatic analysis. Therefore, the present study may advance the understanding of the underlying molecular mechanisms of CHD and may contribute to the diagnosis and treatment of CHD.

2. Methods

2.1. Search Strategy. The mRNA expression profiles, which included 12 CHD and 118 normal blood samples, were obtained from the exoRBase (http://www.exorbase.org/exoRBaseV2/download/toIndex) database [9]. The expression matrix was preprocessed using the following included criteria: (1) mRNA was filtered, and lncRNA was deleted; (2) at least one of mRNAs was nonzero in specific gene expression and (3) on the same platform. Finally, a total of 13768 mRNAs (including 6 CHD blood samples and 32 normal blood samples), which had proper expression data, were included for further analysis. Besides, the acquired genes could trace back to the GEO database (https://www.ncbi.nlm.nih.gov/geo/), and gene expression profiles of GSE100206 and GSE99985 were downloaded. Normal peripheral blood samples were collected from Shanghai Jiao Tong University School of Medicine, and CHD peripheral blood samples were collected from Fudan University Shanghai Cancer Center and Biomedical Research Institute. All the gene matrix data were performed by the single platform of Illumina HiSeq 2000 (Homo sapiens).

2.2. Data Preprocessing and DEG Analysis. Using the GEO database and merging two genes expressed matrix (including GSE99985 and GSE100206), the genes were preprocessed to do the next analysis. The limma (https://www.bioconductor.org/packages/release/bioc/html/limma.html) package in the R (version 4.0.5) software was used to normalize the gene expression profile of peripheral blood samples. The expression profile contained 13768 genes and was used for further study. After that, the limma package was employed to calculate DEGs. An adjusted-P value <0.05 and $|\log_2$ fold change $(\log_2 FC)| > 1$ were considered as a threshold.

2.3. WGCNA Construction. WGCNA is a system biology method used to describe the correlation patterns of genes in microarray samples and is commonly used in a variety of system biology analyses [13]. To conduct WGCNA analysis, the biochip platform (GPL11154) annotation information was used to match gene probes and gene names. The coexpression network module was constructed by the R software and the WGCNA package (https://cran.rproject.org/web/packages/WGCNA/index.html). To ensure the reliability of network construction, we first normalized the samples, then eliminated outliers, constructed a hierarchical cluster tree, and divided the genes with high and low coexpression into the same module according to their respective expression levels. Then, the adjacency matrix was transformed into a topological overlap matrix, and the corresponding dissimilarity degree was calculated. Based on hierarchical gene clustering, the modules were identified by the dynamic tree cutting method. The depth segmentation value was 2, and the minimum size cutoff value was 50. Meanwhile, the Pearson correlation matrix and adjacency matrix were to establish the information of the whole common expression network. Commonly, the value of the highest module significance was considered as the significant part.

2.4. Enrichment Analysis. WGCNA is a network-based method concentrating on gene sets other than individual genes, which alleviates the multiple testing problem inherent in microarray data analysis and is available for unweighted correlation networks. To evaluate its biological function, genes of the intersection between DEGs and the most significant module were selected for further functional enrichment analysis. The http://org.hs.eg/. db package (https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html) was selected to map the key genes with ensemble ID. The Clusterprofiler package (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) was used to perform Gene Ontology (GO) functional annotations to explore and determine the potential biological function. RichPlot, colorspace, STRING, dose, and ggplot2 packages were also used as dependent packages, and three parts, including biological processes (BPs), cellular components (CCs), and molecular functions (MFs), were obtained.

2.5. PPI Network Construction and Identification of Hub Genes. At the protein level, the STRING database (https://string-db.org) was employed to construct a protein-protein interaction (PPI) network and then saved as a tsv file. The Cytoscape software (version 3.8.2) network analyzer was utilized to develop the interaction association of the candidate genes encoding in CHD. After that, the CytoHubba plugin

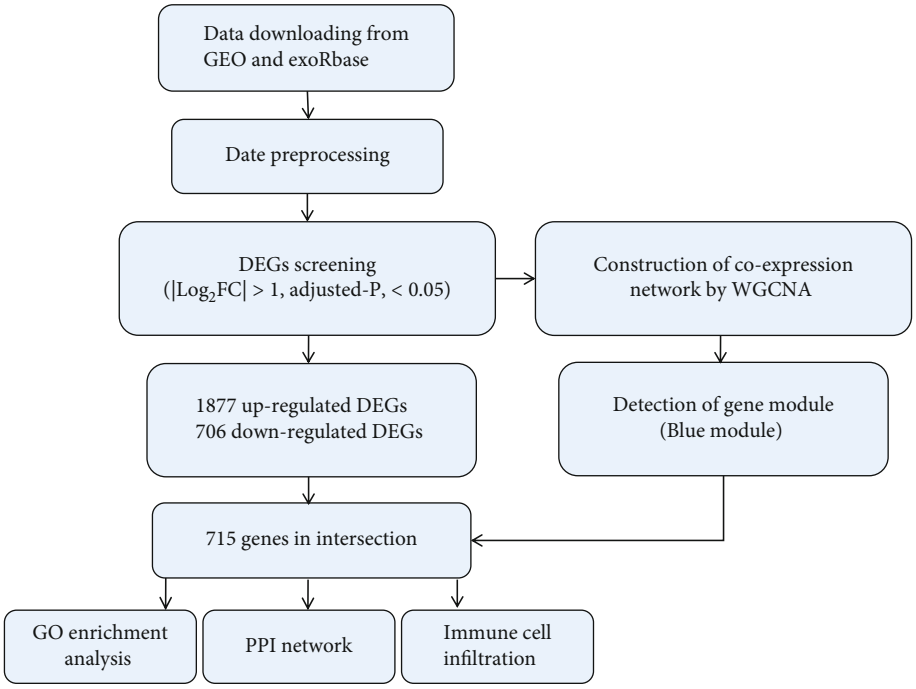


Figure 1: Design and workflow of the whole study. Abbreviations: DEGs: differentially expressed genes; WGCNA: weighted gene correlation network analysis; GO: Gene Ontology; PPI: protein-protein interaction.

was inserted and the Maximal Clique Centrality (MCC) method was calculated to find the top 10 hub genes. Besides, the MCC algorithm performs better performance in predicting hub genes in PPI networks compared with the rest of the topological algorithms. Thus, we selected the MCC algorithm to identify HCC hub genes [16].

2.6. Immune Infiltration Pattern. The proportion of each group of immune cells was estimated using the deconvolution method CIBERSORT (https://cibersort.stanford.edu/). We set to run mode as bulk-mode, disabled quantile normalization, and 100 permutations were set for the following significance analysis. We obtained the gene signatures to identify 22 immune cell populations (B cell naive, B cell memory, plasma cells, T cell CD8, T cell CD4 naive, T cell CD4 memory resting, T cell CD4 memory activated, T cell follicular helper, T cell regulatory (Treg), T cell gamma delta, NK cells resting, NK cells activated, monocytes, macrophage M0, macrophage M1, macrophage M2, dendritic cells resting, dendritic cells activated, mast cells resting, mast cells activated, eosinophils, and neutrophils). After filtration, the corrplot package was employed to generate a correlation heat map. The ggplot2 package was used to compare the normal group with the CHD group. Adjusted-P value <0.05 was considered significant to the corresponding cell type.

3. Results

3.1. Flow Diagram of the Study. Figure 1 shows the workflow of the study. First, the data was obtained from the exoRBase and GEO databases. After conduct batch normalization by limma package, we proceed with DEG screening and WGCNA analysis, respectively. Based on the WGCNA result, the most significant gene modules were detected. Then, taking the intersection DEGs and acquired module and the overlapping genes were regarded as the significant genes we were interested. Then, basing on DEG results, GO analyses were performed to identify the function of hub genes. The Cytoscape software was used, and 10 hub genes of the PPI network were constructed to show the interaction. Last, we analyzed the immune cell infiltrate pattern.

3.2. Identification of DEGs. The GEO dataset of the blood sample was dealt with the R software. We found that there was a significant batch effect between different datasets, which was corrected by performing batch normalization in the limma package (Figure 2). Then, a total of 2583 DEGs were found. Among that, 706 genes were downregulated, and 1877 genes were upregulated in the CHD group compared with the normal group. Figure 3(a) depicts the upregulated, downregulated genes, and non-DEGs in volcanic maps. Meanwhile, the top 50 DEGs ranked with adjusted-P values were used to generate a heat map (Figure 3(b)).

3.3. Weighted Gene Coexpression Networks and Finding Module of Interest. We extracted exosome sequencing genes from CHD blood samples and exosome sequencing genes from normal blood samples for WGCNA to explore coexpression networks. The CHD and normal sample cluster tree diagram is shown in Figure 4(a). We used a scale-free topology index and mean connectivity to determine the soft threshold of WGCNA. The higher the scale-free topology index value equaled to the greater the possibility of the scale-free feature. The correlation coefficient

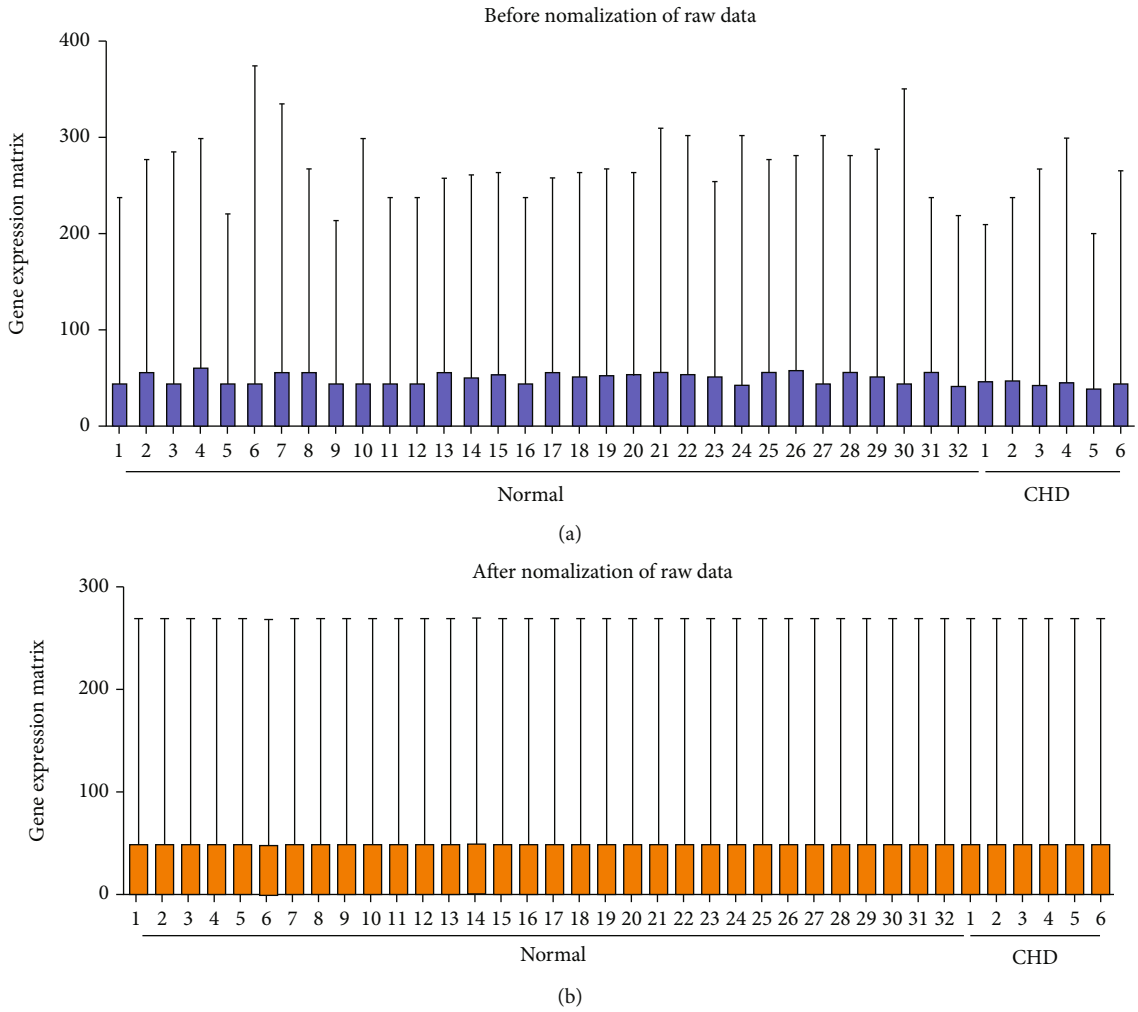


Figure 2: Normalization of CHD expression. (a) Expression microarray datasets of GSE100206 and GSE99985 before normalization. (b) GSE100206 and GSE99985 datasets of normalization. Abbreviations: CHD: coronary heart disease; N: normal samples.

between log (k) and log$P$ (k) was 0.9, and the soft threshold $\beta = 3$ was selected to convert the correlation matrix into a scale-free adjacency matrix (Figure 4(b)). Next, the dynamic tree cutting based on the topological overlap matrix was used to generate the coexpression module, and the coexpression network generated a total of 11 modules (Figure 4(c)). To examine the correlation between different modules and CHD conditions, we calculated the correlation factors for each module (Figure 4(d)). It showed that the correlation coefficients of the blue module, which contains 1994 genes, were greatest (Figure 4(e), correlation coefficient = 0.89, $P < 0.001$, containing 487 genes). The modules obtained from WGCNA were verified with the results of differential gene cluster analysis. To further explore the physiological or pathological pathways associated with CHD, DEGs and the blue module were intersected, and 715 overlapping genes were obtained (Figure 4(f)).

3.4. Functional Annotation. The GO enrichment analysis of overlapping genes is shown in Table 1 and Figure 5(a). As for BP, the analysis showed that these genes were enriched in multiple pathways, including neutrophil degranulation and activation involved in the immune response. With regard to CC, these genes were mainly involved in the formation of vesicle lumen, cytoplasmic lumen, and secretory granule lumen. And for MF, these genes were related to glycosaminoglycan binding, heparin binding, and antioxidant activity. These results suggest that CHD exosome mRNA may play an important role in regulating immune response during the occurrence and development of disease.

3.5. PPI Network Construction and Identification of Hub Genes. A PPI network of 715 overlapping genes was depicted using the STRING database. Hub genes were selected from the PPI network through the MCC algorithm of CytoHubba plugin (Figure 5(b)). The top 10 genes with the highest MCC scores were identified as centers and are shown in Figure 5(c), including lysozyme (LYZ), CAMP, haptoglobin (Hp), ORM1, LTF, CRISP3, PRG3, MMP8, OLFM4, and peptidoglycan recognition protein-1 (PGLYRP1) (Figure 5(d)).

3.6. Immune Cell Infiltration Analysis. Multiple GO functional analyses related to immune processes were identified. Thus, to further discover the relationship between both, we

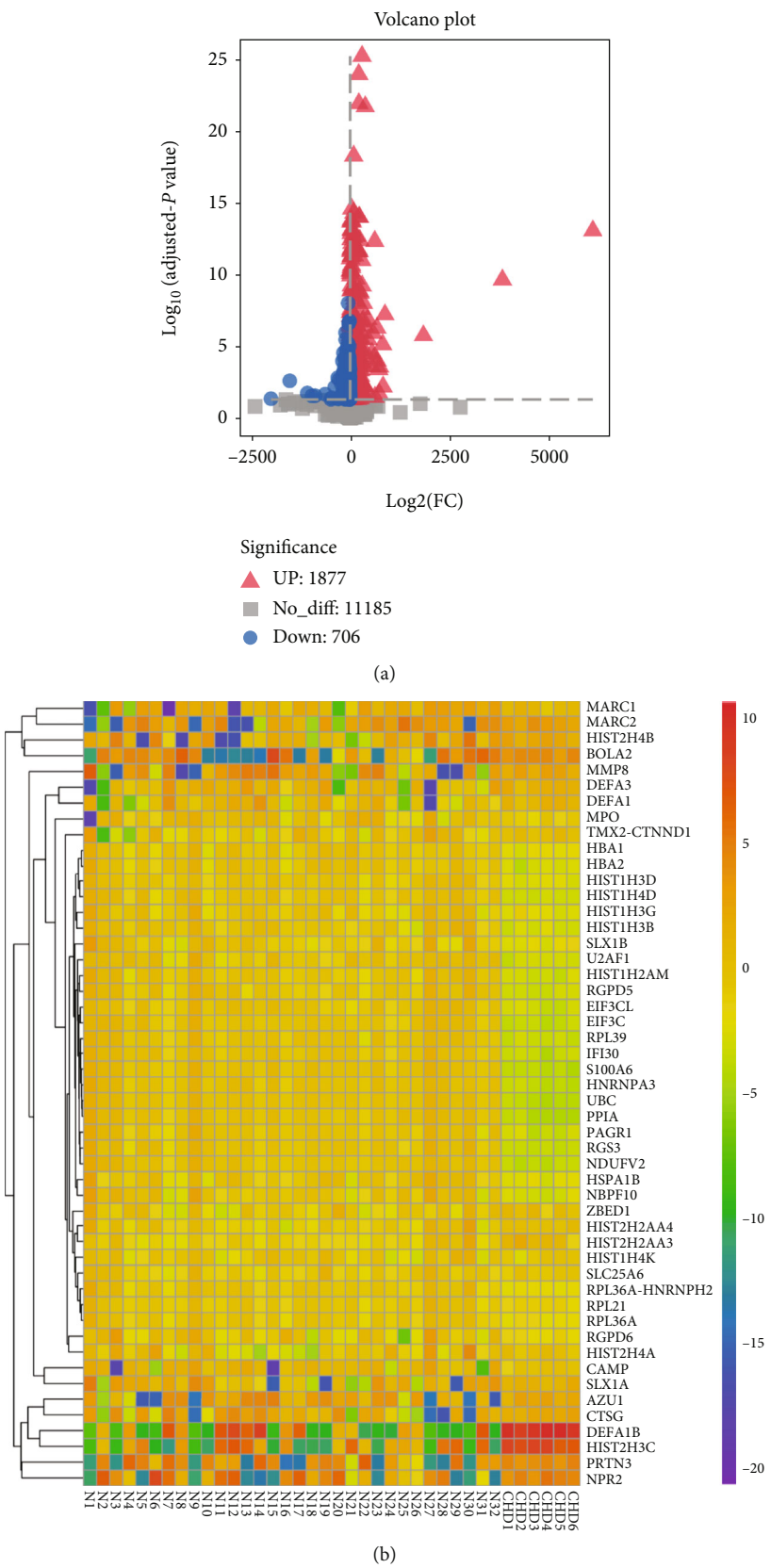
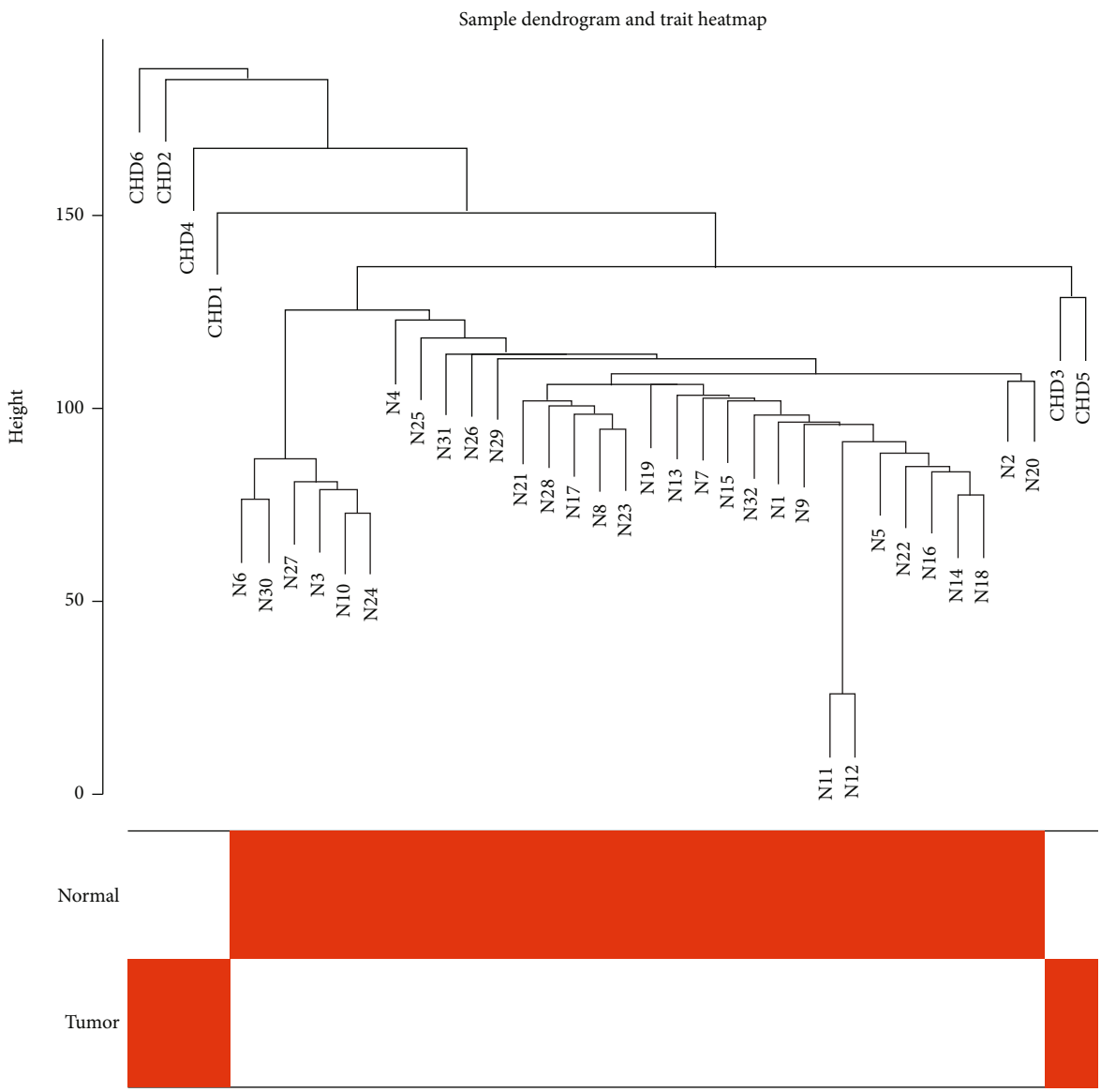
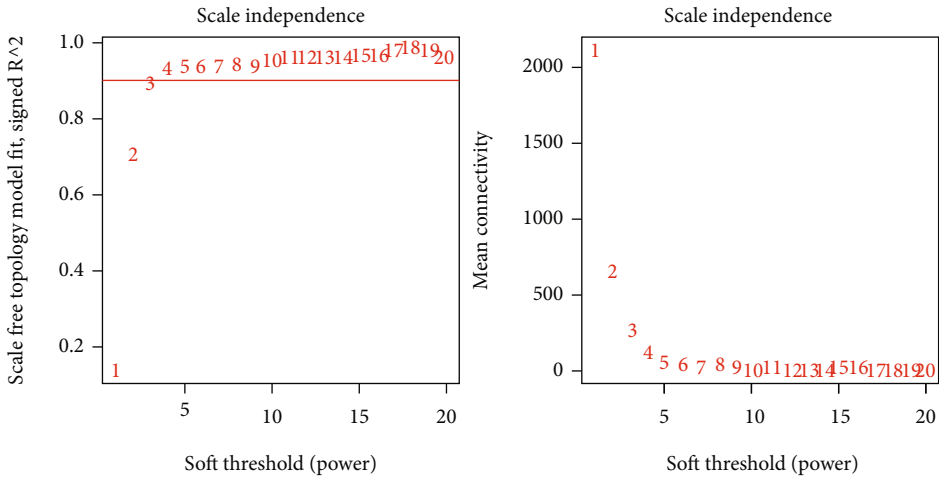


Figure 3: Differential analysis of datasets. (a) Differentially expressed genes screened by the criteria of $|\log_2 FC| > 1$ and adjusted-P value <0.05 and showed by volcano plot. The upregulated and downregulated genes were marked by triangles and circles, respectively. (b) The top 50 DEGs with the smallest adjusted-P value in the upregulated and downregulated clusters were taken out to generate a heat map, respectively. Abbreviations: DEGs: differentially expressed genes; $\log_2$FC: log fold change.

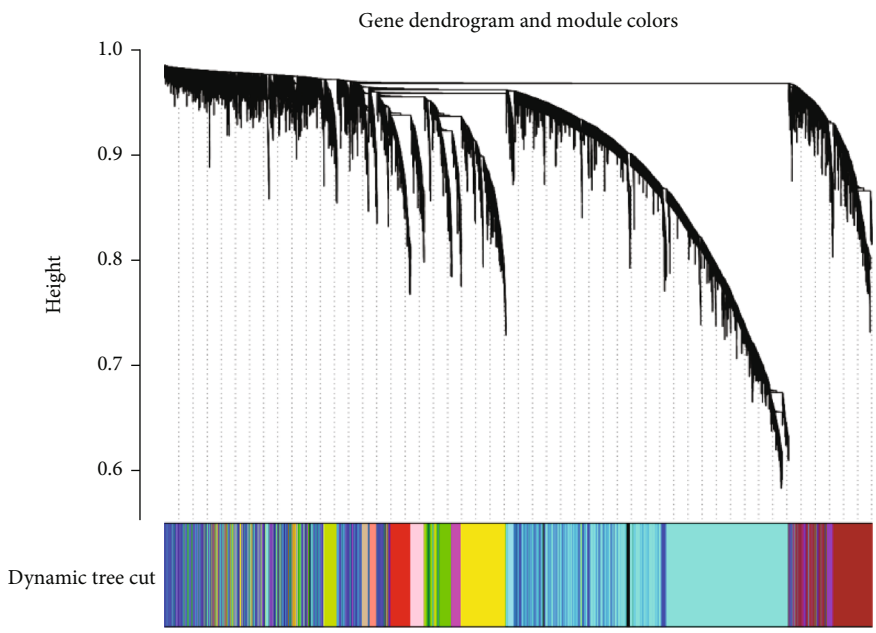


(a)

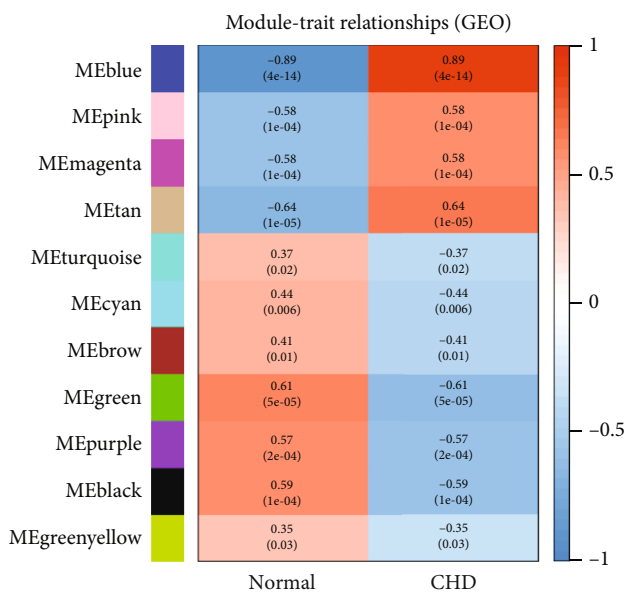


(b)

Figure 4: Continued.



(c)



(d)

Figure 4: Continued.

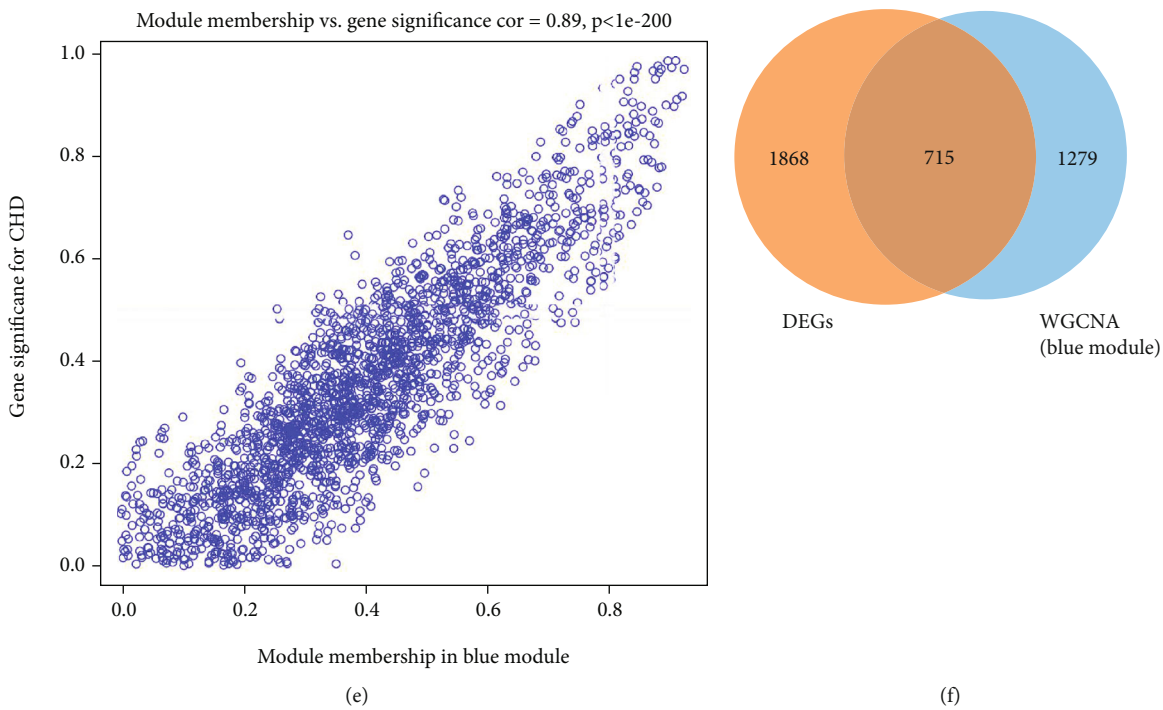


Figure 4: Construction of weighted gene coexpression network of CHD samples. (a) Sample clustering to detect outliers. (b) The cutoff was set to be 0.9, and $\beta = 3$ was chosen to be the soft-threshold power. (c) The gene dendrogram showed that the molecules were classified into different gene modules based on the correlation analysis. Different colors represented the different modules. (d) A heat map of the relationship between module traits showed the correlation between different modules and disease status. The red square represented a positive correlation, and the blue square represented a negative correlation. The common correlation between the module and the disease and the P value was shown in the box. (e) Correlation between module membership of blue and green circles and gene significance (absolute value). (f) Venn diagram of gene crossover between the DEG list and the blue module. A total of 715 overlapping genes were located at the intersection of the DEGs and the blue module. Abbreviation: DEGs: differentially expressed genes.

first examined the pattern of immune infiltration under CHD and normal conditions using the CIBERSORT algorithm. The percent of the 22 immune cells is visually displayed in Figure 6(a). Corheatmap (Figure 6(b)) result showed that mast cells activated and eosinophils had a positive correlation (value = 0.92). B cell memory had a significant negative correlation (value = −0.74) with B cell naive. Besides, in CHD patients, neutrophils and mast cells were significantly increased ($P < 0.05$) and showed in the violplot as below (Figure 6(c)).

4. Discussion

Exosomes had shown a critical effect on the occurrence and development of CHD. Intercellular vesicle information transport of exosomes is one of the important mechanisms of CHD [17]. The recognition and researches of exosomes were penetrating deeply and had got some new progressions for the past few years. Valadi et al. showed many mRNAs were not present in the donor cell [18]. They were passed to another cell and translated through the exosome. It had been proved the translation of exosome mRNAs was functional in vitro. To confirm the potential effects of exosome genes in the development of CHD, we screened the DEGs associated with CHD based on the exoRBase data and the most related module to obtain the candidate genes associated with CHD. In the GO enrichment analyses, most genes were enriched in BP and CC. Neutrophil degranulation, neutrophil activation involved in immune response, secretory granule lumen, cytoplasmic vehicle lumen, and vehicle lumen collagen-containing were the most remarkable categories. In this study, we used the Cytoscape software and the MCC method to further discover the core genes in the network. Then, the plugin Cytoscape was used for node ranking calculation. Chin et al. [19] have reported that CytoHubba provides 11 topological analysis methods including Degree, Edge Percolated Component, Maximum Neighborhood Component, Density of Maximum Neighborhood Component, Maximal Clique Centrality, and six centralities (Bottleneck, EcCentricity, Closeness, Radiality, Betweenness, and Stress) based on shortest paths. Among that, the newly proposed method, MCC, has a better performance on the precision of predicting essential proteins and has been adopted in this study. Finally, ten hub genes were discovered from the PPI network, including LYZ, CAMP, HP, ORM1, LTF, CRISP3, MMP8, OLFM4, and PGLYRP1.

Atherosclerosis was considered to be the pathological foundation of CHD. The development of atherosclerosis is the result of the combined action of chronic inflammation and abnormal lipid metabolism [20]. CAMP was a member of the antimicrobial peptide family with a cathelin domain characterized by a highly conserved N-terminal signal

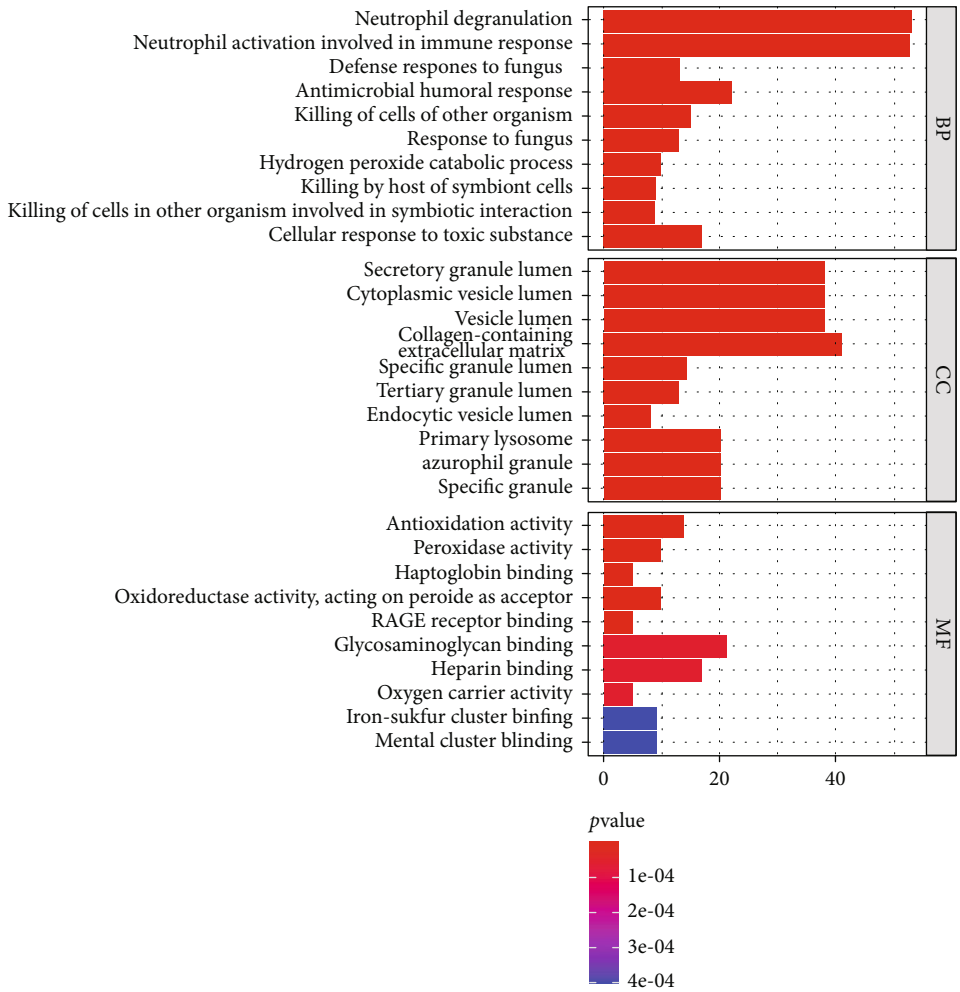
Table 1: GO analysis of overlapping genes.

Category	ID	Description	Adjust- *P* value	Genes
BP	GO:0043312	Neutrophil degranulation	3.25*E*-11	DEFA1B/HSPA1B/CAMP/MPO/AZU1/MMP8/DEFA1/CTSG/PRTN3/PGAM1/MMP9/ELANE/S100A8/HP/EPX/LTF/PRG2/PGLYRP1/CEACAM6/S100A9/CYSTM1/CDA/LYZ/ORM1/FCGR3B/DEFA4/S100A12/BPI/OLFM4/PRG3/LGALS3/RNASE2/HVCN1/A1BG/MS4A3/SLC2A5/ABCA13/HSPA1A/S100P/TNFAIP6/CD59/GPR84/FPR1/ARG1/RNASE3/BRI3/CXCR1/CRISP3/CEACAM8/PRAM1/MME/ALAD/PTGES2
BP	GO:0002283	Neutrophil activation involved in immune response	3.25*E*-11	DEFA1B/HSPA1B/CAMP/MPO/AZU1/MMP8/DEFA1/CTSG/PRTN3/PGAM1/MMP9/ELANE/S100A8/HP/EPX/LTF/PRG2/PGLYRP1/CEACAM6/S100A9/CYSTM1/CDA/LYZ/ORM1/FCGR3B/DEFA4/S100A12/BPI/OLFM4/PRG3/LGALS3/RNASE2/HVCN1/A1BG/MS4A3/SLC2A5/ABCA13/HSPA1A/S100P/TNFAIP6/CD59/GPR84/FPR1/ARG1/RNASE3/BRI3/CXCR1/CRISP3/CEACAM8/PRAM1/MME/ALAD/PTGES2
BP	GO:0050832	Defense response to fungus	4.58*E*-07	DEFA1B/DEFA3/MPO/DEFA1/CTSG/ELANE/S100A8/LTF/S100A9/DEFA4/S100A12/HRG/ARG1
BP	GO:0019730	Antimicrobial humoral response	6.57*E*-07	DEFA1B/CAMP/DEFA3/AZU1/DEFA1/CTSG/PRTN3/ELANE/S100A8/LTF/PGLYRP1/S100A9/LYZ/DEFA4/S100A12/BPI/CXCL9/HRG/BCL3/PGLYRP2/FGB/RNASE3
BP	GO:0031640	Killing of cells of other organism	1.68*E*-06	DEFA1B/CAMP/DEFA3/AZU1/DEFA1/CTSG/ELANE/LTF/PGLYRP1/LYZ/DEFA4/S100A12/HRG/ARG1/APOL1
CC	GO:0034774	Secretory granule lumen	1.77*E*-09	DEFA1B/CAMP/DEFA3/MPO/AZU1/MMP8/DEFA1/CTSG/PRTN3/PGAM1/ELANE/S100A8/HP/EPX/LTF/PGLYRP1/S100A9/CDA/LYZ/ORM1/FN1/DEFA4/S100A12/TIMP3/BPI/OLFM4/PRG3/RNASE2/A1BG/HRG/S100P/FGB/ARG1/RNASE3/APOA1/CRISP3/ALAD/PTGES2
CC	GO:0060205	Cytoplasmic vesicle lumen	1.77*E*-09	DEFA1B/CAMP/DEFA3/MPO/AZU1/MMP8/DEFA1/CTSG/PRTN3/PGAM1/ELANE/S100A8/HP/EPX/LTF/PGLYRP1/S100A9/CDA/LYZ/ORM1/FN1/DEFA4/S100A12/TIMP3/BPI/OLFM4/PRG3/RNASE2/A1BG/HRG/S100P/FGB/ARG1/RNASE3/APOA1/CRISP3/ALAD/PTGES2
CC	GO:0031983	Vesicle lumen	1.77*E*-09	DEFA1B/CAMP/DEFA3/MPO/AZU1/MMP8/DEFA1/CTSG/PRTN3/PGAM1/ELANE/S100A8/HP/EPX/LTF/PGLYRP1/S100A9/CDA/LYZ/ORM1/FN1/DEFA4/S100A12/TIMP3/BPI/OLFM4/PRG3/RNASE2/A1BG/HRG/S100P/FGB/ARG1/RNASE3/APOA1/CRISP3/ALAD/PTGES2
MF	GO:0016209	Antioxidant activity	0.000793131	HBA1/MPO/HBA2/HP/EPX/S100A9/PRDX2/APOE/TXN/HBM/HBZ/TXNDC17/HBG1/PXDNL
MF	GO:0004601	Peroxidase activity	0.002388549	HBA1/MPO/HBA2/EPX/PRDX2/HBM/HBZ/TXNDC17/HBG1/PXDNL
MF	GO:0031720	Haptoglobin binding	0.002388549	HBA1/HBA2/HBM/HBZ/HBG1

Abbreviations: BP: biological process; CC: cellular component; MF: molecular function.

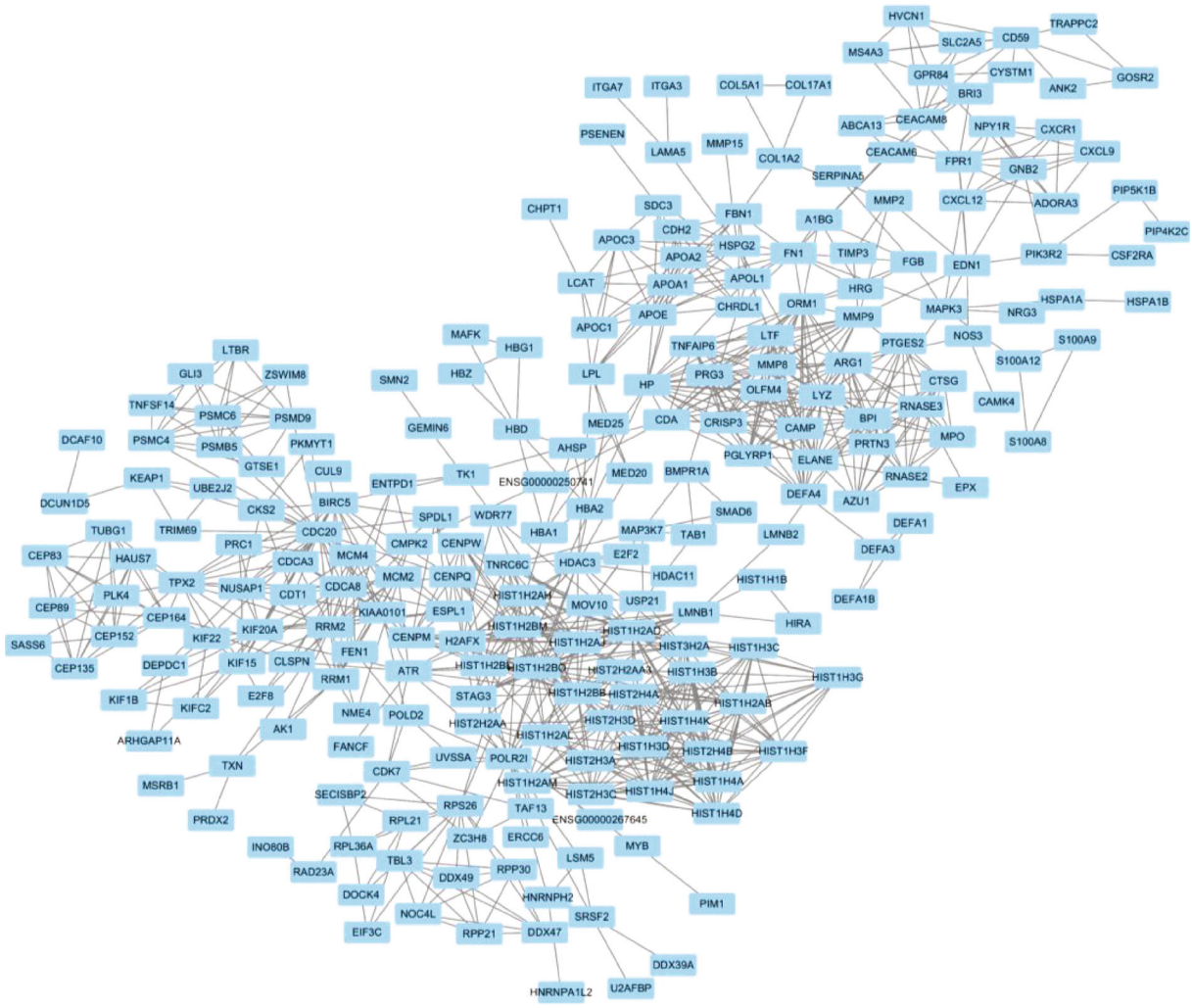
peptide [21, 22], except for its antibacterial activities, the CAMP protein functions in the inflammatory response [23], which was related to the development of CHD. Zhao et al. showed that serum levels of LL-37 (human analog of CAMP) were significantly reduced in CHD patients [24, 25]. Moreover, LL-37 was expressed in atherosclerotic lesions, mainly existed in macrophages and T cells, and functioned in inducing inflammatory gene expression [25]. Amounts of LL-37-mtDNA complex increased in atherosclerotic plasma and plaques, resisted DNase II degradation, and escaped from autophagic recognition in atherosclerosis [26], indicating that CAMP might induce atherosclerosis to increase the risk of CHD.

LYZ was a basic protein interacting with negatively charged phospholipid bilayers [27]. Endogenous LYZ regulated the composition of exosome-related RNA during inflammation, reflecting its role in cell-cell communication signaling in the inflammatory response of CHD [28]. Abey et al. demonstrated the importance of LYZ in epithelial cell migration. Meanwhile, they also found that LYZ therapy altered the expression of proteins, which was associated with signaling networks of inflammation, immune signaling, and atherosclerotic pathways [29]. Abdul-Salam et al. found that LYZ was a potential biomarker for atherosclerotic disease. The elevated LYZ level was closely correlated with disease severity, suggesting its value as a diagnostic tool to assess CHD patients [30].

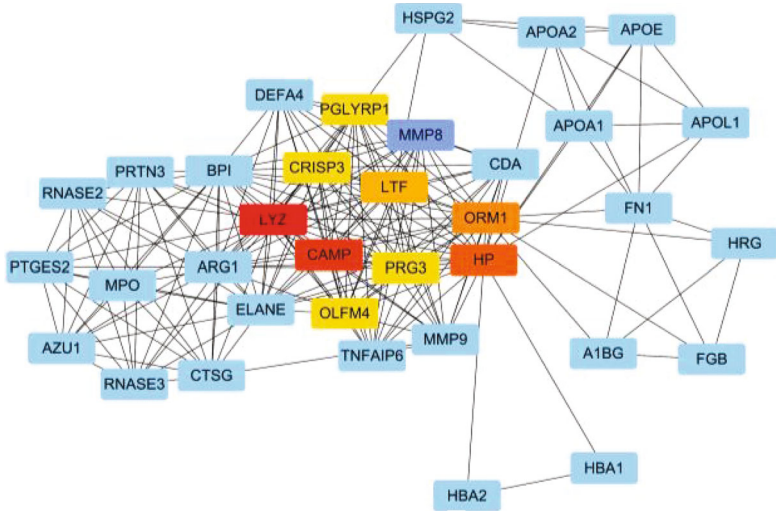


(a)

Figure 5: Continued.



(b)

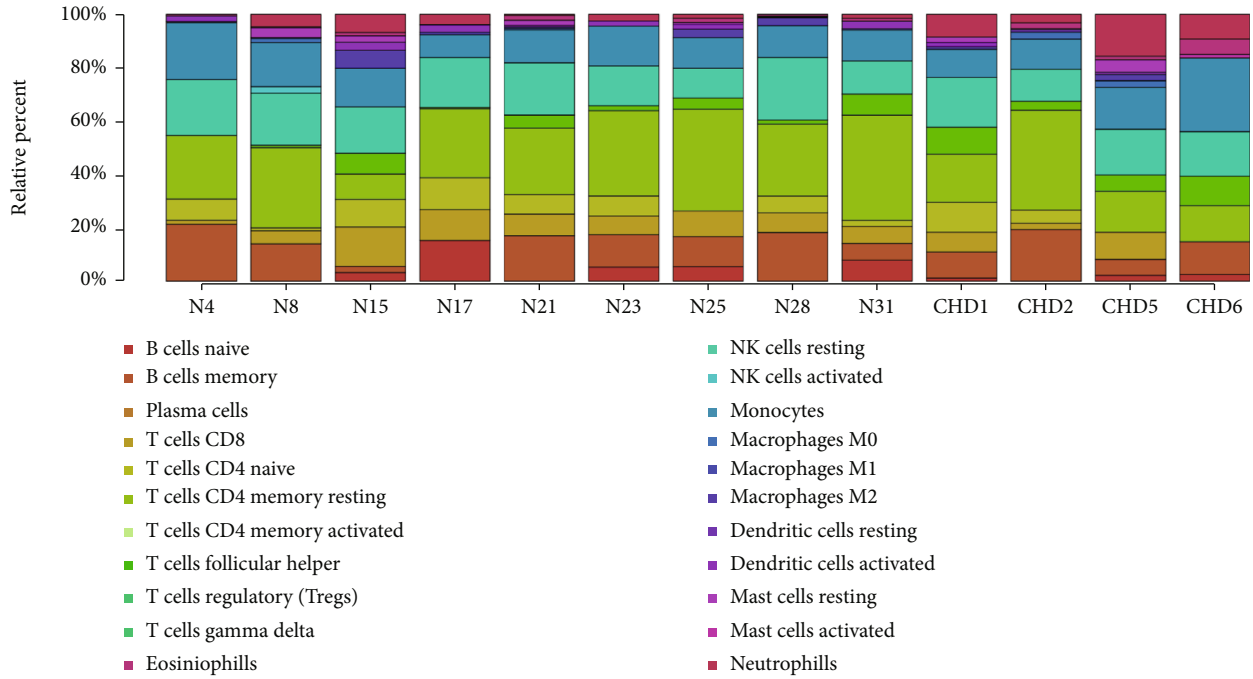


(c)

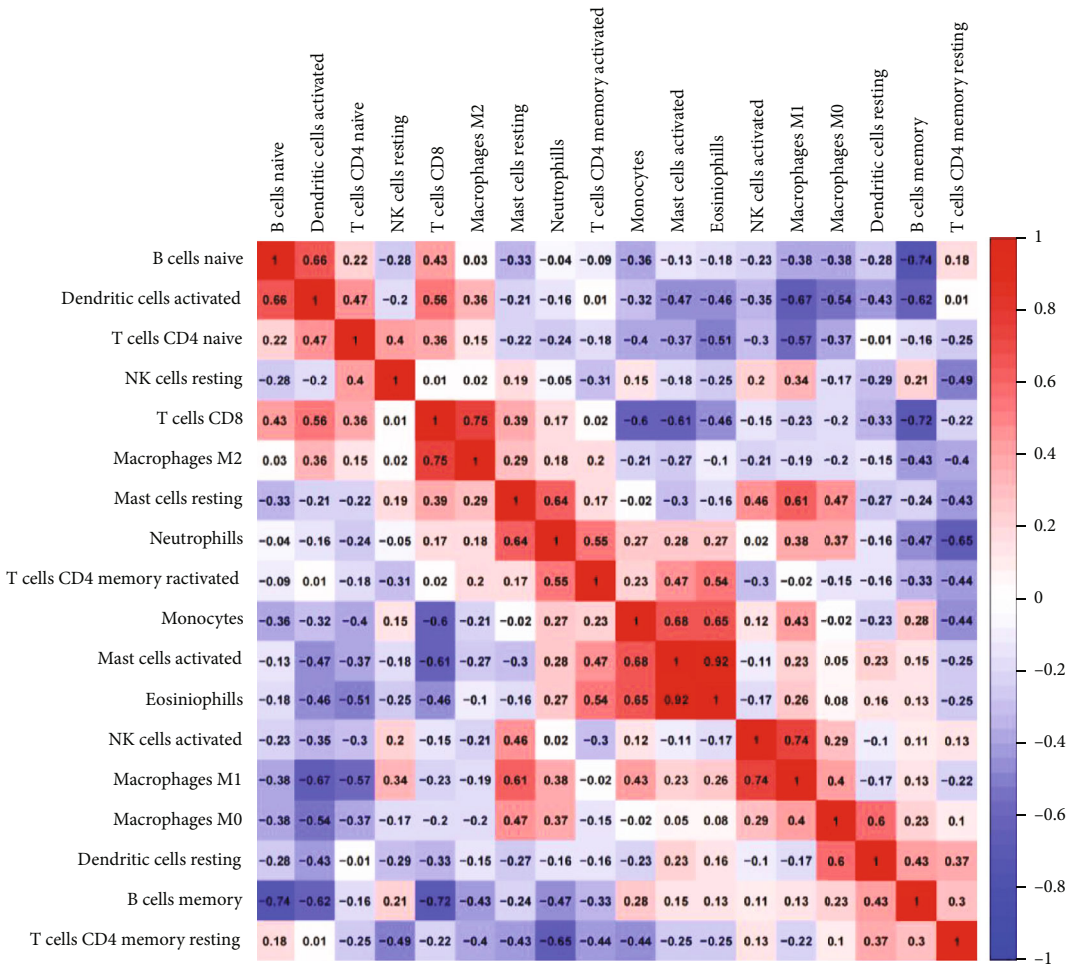
Rank	Node
1	LYZ
2	CAMP
3	HP
4	ORM1
5	LTF
6	CRISP3
6	PRG3
6	MMP8
6	OLFM4
6	PGLYRP1

(d)

Figure 5: Functional enrichment analysis of DEGs. (a) GO enrichment significance items of overlapping DEGs in different functional groups: BP, CC, and MF. (b) Visualization of the PPI network and hub genes of the identified DEGs by Cytoscape of a PPI network of overlapping DEGs. (c) Ten hub genes of the overlapping DEGs marked with different colors. (d) Top 10 DEGs ranked by MCC. The redder, the higher its grades. Abbreviations: DEGs: differentially expressed genes; GO: Gene Ontology; BP: biological process; CC: cellular component; MF: molecular function; PPI: protein-protein interaction.

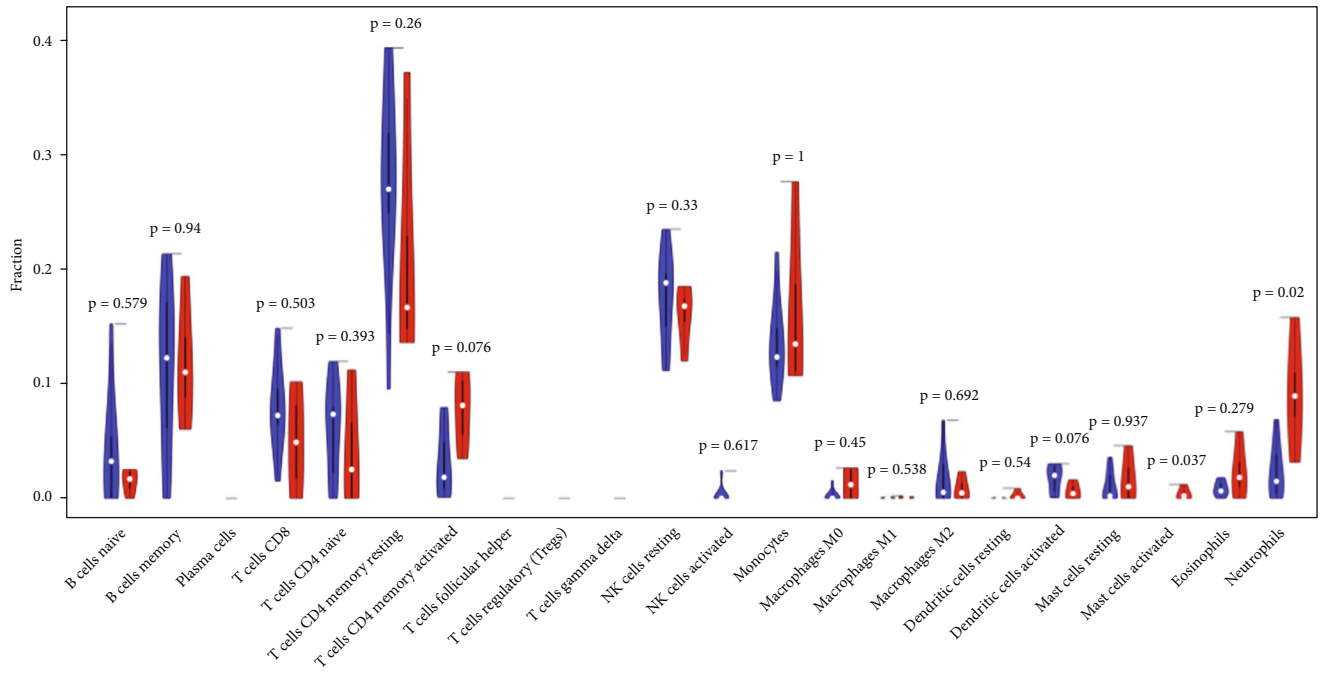


(a)



(b)

Figure 6: Continued.



(c)

Figure 6: Immune cell infiltration analysis. (a) The distribution of immune infiltration in peripheral blood of CHD patients in 22 subpopulations of immune cells. Different colors represent different immune cells, and the length of the bars in the barplot represents the proportion of the immune cell population. (b) A Corheatmap of the correlation matrix for filtered 18 immune cell proportions in CHD. The red color represents the positive relationship between two immune cells, and the blue color represents the negative relationship between two immune cells. The darker the color, the higher the correlation was ($P < 0.05$). (c) The violin plot of immune cells. The blue bar represents normal samples, and the red bar represents the peripheral blood samples of CHD patients. Abbreviations: CHD: coronary heart disease; N: normal samples.

PGLYRP1 was a bacterial wall component known to be present in human atherosclerosis and was found almost exclusively in the secretory granules of neutrophils and eosinophils [31]. Peptidoglycan might promote inflammation by activating innate immune responses, peptidoglycan recognition proteins, chemokines, and proinflammatory cytokines (IL-1, IL-6, and tumor necrosis factor-alpha) in nonmucosal sites [32]. These processes may promote and accelerate the development of atherosclerotic lesions. Our results verified that CAMP, LYZ, and PGLYRP1 were important factors in constituting the atherosclerosis of CHD.

The role of cholesterol in coronary heart disease was undisputed [33, 34]. Hp was a rich plasma protein that played an important role in immune regulation and reversal of cholesterol transport. It is also bound to hemoglobin to protect against oxidative damage [35]. Also, this gene played an important role in CHD pathological process [36, 37]. In Belgian, Hp 1-1 phenotype had a strong association with an increased risk of CHD [38]. It had been reported by Cahill et al. that diabetes mellitus individuals carried with Hp 2-2 allele had more likely to develop CHD [39], which means that Hp was closely related to the occurrence and development of cardiovascular disease.

Other key exosome genes, including LTF, CRISP3, and OLFM4, were mainly associated with acute and chronic inflammation. MMP8, LTF, CRISP3, and OLFM4 were upregulated in the inflammatory process to facilitate leukocyte-mediated migration, neutrophil activation, and degranulation process [40]. MMP8, LTF, and OLFM4 were also known as neutrophil collagenase [41]. MMP8 was expressed and produced by endothelial cells, smooth muscle cells, and macrophages in atherosclerotic plaques [42]. Momiyama et al. proved MMP8 levels were higher in both stable CHD and unstable angina patients. Also, high plasma MMP8 levels suggested that MMP8 might reflect coronary plaque instability, which suggested that MMP8 was a promising biomarker for CHD [43]. OLFM4 and LTF were subpopulations of neutrophils in septic shock. Among that, a high percentage of OLFM4 positive neutrophils were associated with a greater risk of organ failure and death [44]. PRG3 gene was a novel p53 target gene in p53-dependent apoptosis pathway [45].

Infiltrated immune cells constitute important parts of CHD and have been widely studied in recent years. Neutrophil was found to be indicative of responses to plaque formation [46]. It expelled intracellular contents which were rich in uncoagulated chromatin, histones, and active substances [47]. These intracellular contents could participate in plaque erosion, including noxious effects on vascular cells, direct thrombogenic activity, and the promotion of platelet activation/aggregation [47]. Quillard et al. had demonstrated that TLR2 stimulation and neutrophil participation might cause the plaques vulnerable to superficial erosion and thrombotic complications [48]. Therefore, the function of neutrophils

played a critical role in the pathophysiology of CHD, which is consistent with our results.

Some limitations cannot be ignored. First of all, the results in this paper were limited to bioinformatic analysis, and it had not been further proved by experiments. Second, the number of samples was limited because of insufficient databases. Therefore, further verification was needed by collecting more clinical samples. Moreover, studies were needed to explore how these exosome genes work in vivo and in vitro.

In summary, we determined that the activation and degranulation of neutrophils may possess significant roles in mediating the process of CHD. Besides, the underlying mechanism may be related to the immune cell infiltration response. In addition, the core PPI exosome genes identified might be used as biomarkers and therapeutic targets for CHD. This study could provide a new insight to predict, assess, and treat for CHD.

Data Availability

All data can be obtained from the corresponding author Danqing Yu or the first author Yanbin Fu.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yanbin Fu and Yanzhi Ge analyzed and extracted the data and contributed analysis tools. Yanbin Fu and Yanzhi Ge prepared figures and tables. Yanbin Fu, Yanzhi Ge, Zedazhong Su, Danqing Yu, and Jianfeng Cao wrote the main protocol and prepared the manuscript. Danqing Yu and Zedazhong Su conceived and designed the study and approved the final draft. The authors read and approved the final manuscript.

Acknowledgments

This work was supported by the Guangdong Provincial People's Hospital (grant number 8207020467 to author Danqing Yu) and Guangdong Provincial People's Hospital Clinical Research Fund (grant number Y012018085). We thank Dr. Danqing Yu and Dr. Yanzhi Ge for giving us advice during the study.

References

[1] T. L. Ulbricht and D. A. Southgate, "Coronary heart disease: seven dietary factors," *Lancet*, vol. 338, no. 8773, pp. 985–992, 1991.

[2] S. P. V. Petersen, P. Scarborough, and M. Rayner, "British Heart FoundaIn the United Kingdom, coronary heart disease causes almost 114 000 deaths a year, and one in six occurs in womention Health Promotion Research Group," in *Heart & Circulatory Disease Statistics 2020*, British Heart Foundation, Oxford, 2020, wwwheartstatsorg/temp/2020 Statistics Compendium (Tables) pdf.

[3] L. A. Lotta, L. B. L. Wittemans, V. Zuber et al., "Association of genetic variants related to gluteofemoral vs abdominal fat distribution with type 2 diabetes, coronary disease, and cardiovascular risk factors," *Journal of the American Medical Association*, vol. 320, no. 24, pp. 2553–2563, 2018.

[4] J. P. Ferreira, J. G. Cleland, C. S. P. Lam et al., "Heart failure rehospitalizations and subsequent fatal events in coronary artery disease: insights from COMMANDER-HF, EPHESUS, and EXAMINE," *Clinical Research in Cardiology*, vol. 110, 2021.

[5] M. Mathieu, L. Martin-Jaular, G. Lavieu, and C. Théry, "Specificities of secretion and uptake of exosomes and other extracellular vesicles for cell-to-cell communication," *Nature Cell Biology*, vol. 21, no. 1, pp. 9–17, 2019.

[6] S. Ailawadi, X. Wang, H. Gu, and G. C. Fan, "Pathologic function and therapeutic potential of exosomes in cardiovascular disease," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1852, no. 1, pp. 1–11, 2015.

[7] Y. Zhang, Y.-W. Hu, L. Zheng, and Q. Wang, "Characteristics and roles of exosomes in cardiovascular disease," *DNA and Cell Biology*, vol. 36, no. 3, pp. 202–211, 2017.

[8] B. D. Chan, W.-Y. Wong, M. M.-L. Lee et al., "Exosomes in inflammation and inflammatory disease," *Proteomics*, vol. 19, no. 8, article e1800149, 2019.

[9] S. Li, Y. Li, B. Chen et al., "exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes," *Nucleic Acids Research*, vol. 46, no. D1, pp. D106–D112, 2018.

[10] M. Lu, S. Yuan, S. Li, L. Li, M. Liu, and S. Wan, "The exosome-derived biomarker in atherosclerosis and its clinical application," *Journal of Cardiovascular Translational Research*, vol. 12, no. 1, pp. 68–74, 2019.

[11] N. A. Finn, D. Eapen, P. Manocha et al., "Coronary heart disease alters intercellular communication by modifying microparticle-mediated microRNA transport," *FEBS Letters*, vol. 587, no. 21, pp. 3456–3463, 2013.

[12] X. Chen, D. Zhang, F. Jiang et al., "Prognostic prediction using a stemness index-related signature in a cohort of gastric cancer," *Frontiers in Molecular Biosciences*, vol. 7, article 570702, 2020.

[13] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[14] J. Tang, D. Kong, Q. Cui et al., "Prognostic genes of breast cancer identified by gene co-expression network analysis," *Frontiers in Oncology*, vol. 8, p. 374, 2018.

[15] S. Chen, D. Yang, B. Liu et al., "Identification and validation of key genes mediating intracranial aneurysm rupture by weighted correlation network analysis," *Annals of Translational Medicine*, vol. 8, no. 21, p. 1407, 2020.

[16] C. Li and J. Xu, "Identification of potentially therapeutic target genes of hepatocellular carcinoma," *International Journal of Environmental Research and Public Health*, vol. 17, no. 3, p. 1053, 2020.

[17] D. M. Pegtel and S. J. Gould, "Exosomes," *Annual Review of Biochemistry*, vol. 88, no. 1, pp. 487–514, 2019.

[18] H. Valadi, K. Ekström, A. Bossios, M. Sjöstrand, J. J. Lee, and J. O. Lötvall, "Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells," *Nature Cell Biology*, vol. 9, no. 6, pp. 654–659, 2007.

[19] C.-H. Chin, S.-H. Chen, H.-H. Wu, C. W. Ho, M. T. Ko, and C. Y. Lin, "cytoHubba: identifying hub objects and sub-

networks from complex interactome," *BMC Systems Biology*, vol. 8, 8 Supplement 4, 2014.

[20] J. Li, D. H. Lee, J. Hu et al., "Dietary Inflammatory Potential and Risk of Cardiovascular Disease Among Men and Women in the U.S.," *Journal of the American College of Cardiology*, vol. 76, no. 19, pp. 2181–2193, 2020.

[21] D. Fan, L. A. Coughlin, M. M. Neubauer et al., "Activation of HIF-1α and LL-37 by commensal bacteria inhibits _Candida albicans_ colonization," *Nature Medicine*, vol. 21, no. 7, pp. 808–814, 2015.

[22] K.-O. Shin, K. P. Kim, Y. Cho et al., "Both sphingosine kinase 1 and 2 coordinately regulate cathelicidin antimicrobial peptide production during keratinocyte differentiation," *The Journal of Investigative Dermatology*, vol. 139, no. 2, pp. 492–494, 2019.

[23] R. Lande, J. Gregorio, V. Facchinetti et al., "Plasmacytoid dendritic cells sense self-DNA coupled with antimicrobial peptide," *Nature*, vol. 449, no. 7162, pp. 564–569, 2007.

[24] K. Edfeldt, B. Agerberth, M. E. Rottenberg et al., "Involvement of the antimicrobial peptide LL-37 in human atherosclerosis," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 26, no. 7, pp. 1551–1557, 2006.

[25] H. Zhao, H. Yan, S. Yamashita et al., "Acute ST-segment elevation myocardial infarction is associated with decreased human antimicrobial peptide LL-37 and increased human neutrophil peptide-1 to 3 in plasma," *Journal of Atherosclerosis and Thrombosis*, vol. 19, no. 4, pp. 357–368, 2012.

[26] Z. Zhang, P. Meng, Y. Han et al., "Mitochondrial DNA-LL-37 complex promotes atherosclerosis by escaping from autophagic recognition," *Immunity*, vol. 43, no. 6, pp. 1137–1147, 2015.

[27] E. Posse, B. F. De Arcuri, and R. D. Morero, "Lysozyme interactions with phospholipid vesicles: relationships with fusion and release of aqueous content," *Biochimica et Biophysica Acta*, vol. 1193, no. 1, pp. 101–106, 1994.

[28] M. L. Squadrito, C. Baer, F. Burdet et al., "Endogenous RNAs modulate microRNA sorting to exosomes and transfer to acceptor cells," *Cell Reports*, vol. 8, no. 5, pp. 1432–1446, 2014.

[29] S. K. Abey, Y. Yuana, P. V. Joseph et al., "Lysozyme association with circulating RNA, extracellular vesicles, and chronic stress," *BBA Clinical*, vol. 7, pp. 23–35, 2017.

[30] V. B. Abdul-Salam, P. Ramrakha, U. Krishnan et al., "Identification and assessment of plasma lysozyme as a putative biomarker of atherosclerosis," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 30, no. 5, pp. 1027–1033, 2010.

[31] N. K. Brownell, A. Khera, J. A. de Lemos, C. R. Ayers, and A. Rohatgi, "Association between peptidoglycan recognition protein-1 and incident atherosclerotic cardiovascular disease events: the Dallas heart study," *Journal of the American College of Cardiology*, vol. 67, no. 19, pp. 2310–2312, 2016.

[32] A. Rohatgi, C. R. Ayers, A. Khera et al., "The association between peptidoglycan recognition protein-1 and coronary and peripheral atherosclerosis: observations from the Dallas heart study," *Atherosclerosis*, vol. 203, no. 2, pp. 569–575, 2009.

[33] J. Stamler, D. Wentworth, and J. D. Neaton, "Is relationship between serum cholesterol and risk of premature death from coronary heart disease continuous and graded?," *Journal of the American Medical Association*, vol. 256, no. 20, pp. 2823–2828, 1986.

[34] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.

[35] M. K. Jensen, M. L. Bertoia, L. E. Cahill, I. Agarwal, E. B. Rimm, and K. J. Mukamal, "Novel metabolic biomarkers of cardiovascular disease," *Nature Reviews. Endocrinology*, vol. 10, no. 11, pp. 659–672, 2014.

[36] V. V. Bamm, V. A. Tsemakhovich, M. Shaklai, and N. Shaklai, "Haptoglobin phenotypes differ in their ability to inhibit heme transfer from hemoglobin to LDL," *Biochemistry*, vol. 43, no. 13, pp. 3899–3906, 2004.

[37] M. Melamed-Frank, O. Lache, B. I. Enav et al., "Structure-function analysis of the antioxidant properties of haptoglobin," *Blood*, vol. 98, no. 13, pp. 3693–3698, 2001.

[38] D. De Bacquer, G. De Backer, and M. Langlois, "Haptoglobin polymorphism as a risk factor for coronary heart disease mortality," *Atherosclerosis*, vol. 157, no. 1, pp. 161–166, 2001.

[39] L. E. Cahill, M. K. Jensen, S. E. Chiuve et al., "The risk of coronary heart disease associated with glycosylated hemoglobin of 6.5% or greater is pronounced in the haptoglobin 2-2 genotype," *Journal of the American College of Cardiology*, vol. 66, no. 16, pp. 1791–1799, 2015.

[40] A. Banerjee, S. Shukla, A. D. Pandey et al., "RNA-Seq analysis of peripheral blood mononuclear cells reveals unique transcriptional signatures associated with disease progression in dengue patients," *Translational Research*, vol. 186, pp. 62–78.e9, 2017.

[41] R. Almansa, A. Ortega, A. Ávila-Alonso et al., "Quantification of immune dysregulation by next-generation polymerase chain reaction to improve sepsis diagnosis in surgical patients," *Annals of Surgery*, vol. 269, no. 3, pp. 545–553, 2019.

[42] R. Kato, Y. Momiyama, R. Ohmori, H. Taniguchi, H. Nakamura, and F. Ohsuzu, "Plasma matrix metalloproteinase-8 concentrations are associated with the presence and severity of coronary artery disease," *Circulation Journal*, vol. 69, no. 9, pp. 1035–1040, 2005.

[43] Y. Momiyama, R. Ohmori, N. Tanaka et al., "High plasma levels of matrix metalloproteinase-8 in patients with unstable angina," *Atherosclerosis*, vol. 209, no. 1, pp. 206–210, 2010.

[44] M. N. Alder, A. M. Opoka, P. Lahni, D. A. Hildeman, and H. R. Wong, "Olfactomedin-4 is a candidate marker for a pathogenic neutrophil subset in septic shock," *Critical care medicine*, vol. 45, no. 4, pp. e426–e432, 2017.

[45] Y. Ohiro, I. Garkavtsev, S. Kobayashi et al., "A novel p53-inducible apoptogenic gene, PRG3, encodes a homologue of the apoptosis-inducing factor (AIF)," *FEBS Letters*, vol. 524, no. 1-3, pp. 163–171, 2002.

[46] W. C. Schrottmaier, M. Mussbacher, M. Salzmann, and A. Assinger, "Platelet-leukocyte interplay during vascular disease," *Atherosclerosis*, vol. 307, pp. 109–120, 2020.

[47] F. Grégory, "Role of mechanical stress and neutrophils in the pathogenesis of plaque erosion," *Atherosclerosis*, vol. 318, pp. 60–69, 2021.

[48] T. Quillard, H. A. Araújo, G. Franck, E. Shvartz, G. Sukhova, and P. Libby, "TLR2 and neutrophils potentiate endothelial stress, apoptosis and detachment: implications for superficial erosion," *European Heart Journal*, vol. 36, no. 22, pp. 1394–1404, 2015.

Hindawi
BioMed Research International
Volume 2021, Article ID 5122962, 9 pages
https://doi.org/10.1155/2021/5122962

Research Article

COVID-19 Diagnosis from CT Images with Convolutional Neural Network Optimized by Marine Predator Optimization Algorithm

Huaping Jia,[1] Junlong Zhao,[2] and Ali Arshaghi[3]

[1]*College of Computer, Weinan Normal University, Weinan, Shaanxi, China*
[2]*Rehabilitation Medicine Department, Weinan Central Hospital, Shaanxi, China*
[3]*Department of Electrical Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran*

Correspondence should be addressed to Ali Arshaghi; ali.arshagi.eng@iauctb.ac.ir

Received 16 July 2021; Revised 10 August 2021; Accepted 23 September 2021; Published 12 October 2021

Academic Editor: Alireza Baghban

In recent years, almost every country in the world has struggled against the spread of Coronavirus Disease 2019. If governments and public health systems do not take action against the spread of the disease, it will have a severe impact on human life. A noteworthy technique to stop this pandemic is diagnosing COVID-19 infected patients and isolating them instantly. The present study proposes a method for the diagnosis of COVID-19 from CT images. The method is a hybrid method based on convolutional neural network which is optimized by a newly introduced metaheuristic, called marine predator optimization algorithm. This optimization method is performed to improve the system accuracy. The method is then implemented on the chest CT scans with the COVID-19-related findings (MosMedData) dataset, and the results are compared with three other methods from the literature to indicate the method's performance. The final results indicate that the proposed method with 98.11% accuracy, 98.13% precision, 98.66% sensitivity, and 97.26% $F1$ score has the highest performance in all indicators than the compared methods which shows its higher accuracy and reliability.

1. Introduction

During authoring this paper on May 15, 2021, 162,974,265 COVID-19 cases and 3,378,495 deaths are reported by the "Worldometers" organization. This disease was officially named by the World Health Organization as coronavirus disease 2019 (COVID-19) on the 11th of February 2020. The outbreak, originally associated with a city in China, has now become a widespread pandemic, affecting more than 1.2 million people in more than 200 countries and regions around the world.

Several approaches have been introduced for diagnosing COVID-19, including nucleic acid test (NAT), chest radiographs, and CT scan of the lungs. NAT is used to identify specific nucleic acid sequences and species of an organism, mainly viruses or bacteria that cause disease in the blood, tissue, or urine. Although diagnostic kits play an important role in the diagnosis of COVID-19, chest radiographs and CT scans of the lungs are some of the most effective ways to diagnose the severity and degree of pneumonia that may have been transmitted by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Recently, some researches have been done on lung CT scan images for the early detection of COVID-19 based on image processing and artificial intelligence techniques.

Ahuja et al. proposed a method for the diagnosis of COVID-19 based on decomposing the CT scan images into three levels using a stationary wavelet [1]. This three-phase diagnosis system was presented to progress the accuracy diagnosis [2]. The method first used data augmentation using stationary wavelets. Then, COVID-19 was diagnosed based on the pretrained CNN model for abnormality localization in CT scan images. The method used some well-known pretrained architectures, like ResNet18, ResNet50, ResNet101, and SqueezeNet, for the diagnosis. The simulation results showed that the empirical assessment approves that the ResNet18 pretrained transfer learning-based method provides better classification accuracy.

Maghdid et al. introduced a method for diagnosing COVID-19 based on deep learning [3]. Due to the less values of the CT scan dataset for COVID-19, the study built a

general dataset of CT scans and X-ray images from multiple sources to offer a simple and effective diagnosis system for COVID-19.

Then, a simple convolution neural network (CNN) and a modified pretrained AlexNet model have been performed on the datasets. The experimental results indicated that the employed models offer high accuracy for the diagnosis of COVID-19.

Minaee et al. presented a method based on analysis of radiology images for the diagnosis of COVID-19 [4]. The method had been performed on the COVID-19 chest X-ray images on the datasets from the internet. Four different structures of convolutional neural networks, including ResNet18, ResNet50, DenseNet-121, and SqueezeNet, were utilized for the diagnosis. Simulation results showed that all CNN models provide a satisfying accuracy for the diagnosis of COVID-19 disease.

Some other models based on CNN, such as the combined deep convolution networks [5] and unsupervised learning [6], are also presented for the diagnosis of COVID-19, although the method accuracy for the diagnosis of COVID-19, particularly for low-density areas, is low.

In the present study, a new method has been proposed for COVID-19 area segmentation based on a CNN architecture using VGG-16 encoder for semantic and U-Net segmentation methods. The presented methodology does not need more training data owning to the advantages of U-Net, which provides a model to be used on systems with low-strength GPUs. Therefore, the main contribution of this study can be highlighted as follows:

(i) Proposing a new optimal method for the diagnosis of COVID-19 from CT images

(ii) Using a hybrid technique based on convolutional neural network (CNN) and metaheuristic techniques

(iii) Optimizing the CNN based on a newly introduced metaheuristic, called marine predator optimization algorithm

2. Convolutional Neural Network

Since the advent of deep learning, the convolutional neural network (CNN or ConvNet) has been the flagship of ideas in deep learning [7]. The CNN was introduced in 1990, inspired by experiments performed by Hubel and Wiesel on the visual cortex. The CNN is a modified version of an artificial neural network that can be employed for various mathematical learning methods such as backpropagation, gradient descent, and regularization [8, 9]. Due to the CNN's special structure and filter-like state, it is processed in the signal area. This network includes three main concepts of layers with a convolutional layer, pooling layer, and fully connected layer.

In a CNN, different layers perform different tasks with two steps for training: the feed-forward stage and the backpropagation stage. In the first stage, the input image is fed to the network and this action is nothing but multiplying the point between the input and the parameters of each neuron and finally applying a convolution operation in each layer. The network output is then calculated. Here, to adjust the network parameters or in other words the network training, the output result is used to calculate the amount of network error. To do this, the output of the network is compared with the correct solution based on the loss function to calculate the error rate. In the next step, based on the calculated error rate, the backpropagation step begins. In this step, the gradient of each parameter is calculated according to the chain rule and all parameters change according to the effect they have on the error created in the network. After updating the parameters, the next feed-forward step begins. After repeating a proper number of these steps, the network training ends.

The learning process in the CNN is to obtain kernel matrices to generate better features of the problem (here, COVID-19 diagnosis). The backpropagation (BP) technique has been considered for learning and for minimizing the error value of the network. The study uses a sliding window for convolution.

The activation function is a rectified linear unit (ReLU) such that $f(x) = \max(x, 0)$ [10]. The method of scale reduction in this study is max pooling. BP defines a gradient descent technique to minimize the error of the neural network by minimizing the cross-entropy [11] which can be mathematically formulated as follows:

$$L = \sum_{j=1}^{N} \sum_{i=1}^{M} - d_j^{(i)} \log z_j^{(i)}, \tag{1}$$

where N signifies the number of samples, $d_j = (0, \cdots, 0, \underbrace{1, \cdots, 1}_{k}, 0, \cdots, 0)$ describes the desired output vector, and z_j defines the achieved output vector of the mth class that is achieved as follows:

$$z_j^{(i)} = \frac{e^{f_j}}{\sum_{i=1}^{M} e^{f_i}}. \tag{2}$$

The function has been extended based on a weight penalty by considering η term as follows:

$$L = \sum_{j=1}^{N} \sum_{i=1}^{M} - d_j^{(i)} \log z_j^{(i)} + \frac{1}{2} \eta \sum_{K} \sum_{L} \omega_{k,l}^2, \tag{3}$$

where L signifies the total number of layers, K is the layer l connections, and ω_k describes the weight for connection. Figure 1 shows a block diagram of a simple CNN for COVID-19 diagnosis.

Several research works have been proposed to optimize the arrangement of the convolutional neural network. Particularly, the application of optimization algorithms in CNNs indicated satisfying achievements [12]. The present study uses a new optimal technique to provide an optimized CNN. All input images have been resized to 28-by-28 pixel

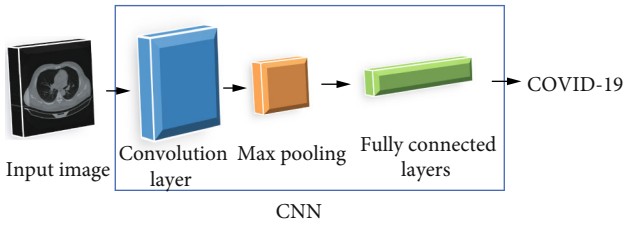


Figure 1: A block diagram of a simple CNN for COVID-19 diagnosis.

images to improve the speed of diagnosis. For the COVID-19 diagnosis problem, the CNN arrangement should be explained by considering some terminology depending on the suggested CNN arrangement:

(i) The input layer of the image in which the input images of the network are normalized by the process

(ii) 2D convolutional layer which implements sliding convolutional filter to convolve with the input image by striding the filter along with the input image horizontally and vertically and evaluates the dot product of the weights and the input image. A bias term is then also added

(iii) Batch normalization layer which is used for normalizing the input channels of the input image crosswise minibatch

(iv) ReLU layer which makes a threshold operation to discard negative values of the image

(v) 2D max pooling layer which makes downsampling by dividing the input image into rectangular pooling regions by calculating the maximum of the regions

(vi) 2D max unpooling layer which unpools the output of the max pooling layer

(vii) Softmax layer which performs the softmax function on the input image

The suggested CNN for COVID-19 diagnosis contains five max pooling and unpooling layers. The main architecture of the CNN is shown in Figure 2.

As can be observed from Figure 2, the layer order for Pooling #1 defines an image input layer, 2D convolution layer, batch normalization layer, ReLU layer, and 2D max pooling layer.

For block Pooling #2, the order is 2D convolution layer, batch normalization layer, ReLU layer, 2D convolution layer, batch normalization layer, ReLU layer, and 2D max pooling layer.

For block Pooling #3, Pooling #4, and Pooling #5, the order is 2D convolution layer, batch normalization layer, ReLU layer, 2D convolution layer, batch normalization layer, ReLU layer, 2D convolution layer, batch normalization layer, ReLU layer, and max pooling 2D layer.

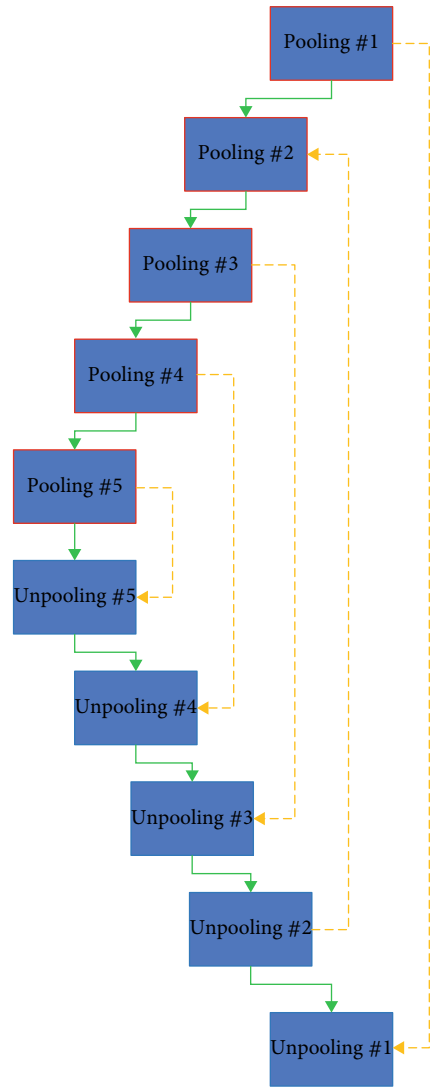


Figure 2: The main architecture of the proposed CNN.

For block Unpooling #5, Pooling #4, and Unpooling #3, the order is 2D max unpooling layer, 2D convolution layer, batch normalization layer, ReLU layer, 2D convolution layer, batch normalization layer, ReLU layer, 2D convolution layer, batch normalization layer, and ReLU layer.

For block Unpooling #2, there are 2D max unpooling layer, 2D convolution layer, batch normalization layer, ReLU layer, 2D convolution layer, batch normalization layer, and ReLU layer. For block Unpooling #1, 2D max unpooling layer, 2D convolution layer, batch normalization layer, ReLU layer, softmax layer, and classification output layer (pixel classification layer) have been used. The presented CNN architecture employed a VGG-16 encoder with U-Net construction.

3. Marine Predator Optimization Algorithm

There are two types of optimization algorithms: exact algorithms and approximate algorithms [13]. Exact algorithms as the first priority present the exact optimal

solutions for optimization problems; thus, they are not well organized for hard optimization problems, such that their execution time improves exponentially based on the problem dimensions [14]. By using approximate algorithms, suitable solutions with a short period can be achieved for optimization problems, even for NP-hard optimization problems that cannot be solved by the exact methods [15]. Metaheuristic algorithms are the best candidates of approximate algorithms [16]. Metaheuristic algorithms define a kind of random algorithm that is employed to provide the optimal solution [17, 18]. Numerous metaheuristic algorithms have been presented in the last decade, e.g., World Cup Optimization (WCO) algorithm [19], Arithmetic Optimization Algorithm (AOA) [20], Ant Lion Optimizer (ALO) algorithm [21], and equilibrium optimizer [22].

Marine predator algorithm (MPA) [23] is another new metaheuristic algorithm that is introduced by Faramarzi et al. The MPA is a new metaheuristic algorithm inspired by marine predators that are used for solving optimization problems. The marine predator algorithm starts with random numbers which are spread uniformly in the search space. This is mathematically modeled as follows:

$$X_0 = X_{\min} + \text{rand} \times (X_{\max} - X_{\min}), \tag{4}$$

where rand describes a uniformly distributed random number in the range $[0, 1]$ and $X_{\min}$ and $X_{\max}$ represent the minimum and maximum boundaries.

The best predators have more intelligence for hunting based on the "survival of the fittest theory" [24]. Accordingly, the best predator is defined as "Elite," which is appropriate for generating a matrix. The prey search has been defined based on matrix arrays using the prey information location. This is defined by the following matrix:

$$E = \begin{bmatrix} X_{1,1}^I & \cdots & X_{1,d}^I \\ \vdots & \ddots & \vdots \\ X_{n,1}^I & \cdots & X_{n,d}^I \end{bmatrix}_{n\times d}, \tag{5}$$

where d signifies the dimensions' number, X^I represents the best predator vector with n simulation to generate the Elite matrix (E), and n is a variable to describe the number of candidates.

Both prey and predator are considered as candidates. This is because when the prey is looking for food, the predator is looking for the prey. At the end of each iteration, the best predator is updated as the new Elite. Furthermore, another matrix with a similar dimension of the Elite is generated as prey, where the position of the predator has been updated by this matrix:

$$P = \begin{bmatrix} X_{1,1} & \cdots & X_{1,d} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,d} \end{bmatrix}_{n\times d}, \tag{6}$$

where $X_{i,j}$ describes the jth dimension for the ith prey. Particularly, the optimization method is associated with these matrixes.

The MPA contains three main units around different speed ratios that are defined as follows:

(i) The prey moves faster (with a higher speed ratio) than the predator

(ii) The predator moves faster (with a lower speed ratio) than the prey

(iii) Both prey and predator move with the same velocity (with equal speed ratio)

Some phases have been clarified by nature principles of prey and predator movement with nature. This description is defined by the following:

(a) This step includes the exploration term of the algorithm which is employed at the initial iterations. If the predator has a higher speed ratio such that $v \geq 10$, the optimum strategy has been used for stopping moving. This is mathematically modeled by the following equation:

$$\begin{aligned} &\text{While Iter} < \frac{1}{3}\,\text{Max}_{\text{Iter}}, \\ &\overrightarrow{\text{stepsize}_i} = \vec{R}_B \otimes \left(\vec{E}_i - \vec{R}_B \otimes \vec{P}_i\right), i = 1, 2, \cdots, n, \\ &\vec{P}_i = \vec{P}_i + P \times \vec{R} \otimes \overrightarrow{\text{stepsize}_i}, \end{aligned} \tag{7}$$

where the sign $\otimes$ describes the entry-wise product and $\vec{R}_B$ represents a vector including some random values that are generated by the Brownian movement [23]. The prey movement has been modeled with the product by prey and $\vec{R}_B$.

P describes a constant value (0.5) and R represents uniformly distributed random values between 0 and 1, and Iter and Max_{Iter} represent the present iteration and the number of iterations.

(b) This step includes the searching of the prey and the predator for the prey. This process is a middle process between the optimization processes. In this step, the exploration attempts to convey the exploitation. Indeed, both exploration and exploitation terms are included in this step. Similarly, the candidate is divided into two parts so that one is employed for exploitation and the other for exploration. In this status, whereas the predator has a Brownian movement, the prey moves in a Lévy movement

$$\text{while}\,\frac{1}{3}\text{Max}_{\text{Iter}} < \text{Iter} < \frac{2}{3}\text{Max}_{\text{Iter}}. \tag{8}$$

Based on this policy, with the exploitation term in the candidate,

$$\begin{aligned}\overrightarrow{\text{stepsize}}_i &= \vec{R}_L \otimes \left(\vec{E}_i - \vec{R}_L \otimes \vec{P}_i\right), \quad i = 1, 2, \cdots, \frac{n}{2},\\ \vec{P}_i &= \vec{P}_i + P \times \vec{R} \otimes \overrightarrow{\text{stepsize}}_i,\end{aligned} \tag{9}$$

where $\vec{R}_L$ describes a random value that is distributed by the Lévy movement [23].

The Lévy movement of the prey has been modeled by multiplying the prey and $\vec{R}_L$ while the prey movement has been modeled. Therefore, for the exploration term in the individual,

$$\begin{aligned}\overrightarrow{\text{stepsize}}_i &= \vec{R}_B \otimes \left(\vec{E}_i - \vec{R}_B \otimes \vec{P}_i\right), \quad i = 1, 2, \cdots, \frac{n}{2},\\ \vec{P}_i &= \vec{E}_i + P \times \text{CF} \otimes \overrightarrow{\text{stepsize}}_i,\end{aligned} \tag{10}$$

where CF defines a modifiable variable to cope with the predator movement that is formulated as follows:

$$\text{CF} = \left(1 - \frac{\text{Itr}}{\text{Max}_{\text{Iter}}}\right)^{(2\times\text{Iter})/\text{Max}_{\text{Iter}}}. \tag{11}$$

(c) The final step is usually allied to improve the exploitation term. Lévy has been performed as the optimum policy for the predator with $v = 0.1$ (low speed ratio). This is modeled as follows:

$$\begin{aligned}&\text{while Iter} > \frac{2}{3}\text{Max}_{\text{Iter}},\\ &\overrightarrow{\text{stepsize}}_i = \vec{R}_L \otimes \left(\vec{R}_L \otimes \vec{E}_i - \vec{P}_i\right), \quad i = 1, 2, \cdots, n,\\ &\vec{P}_i = \vec{E}_i + P \times \text{CF} \otimes \overrightarrow{\text{stepsize}}_i.\end{aligned} \tag{12}$$

By considering the definition of the Fish Aggregating Devices (FADs), above, 80 percent of the time of the sharks was spent close to the FADs and the remaining candidates are employed for longer jumps in various dimensions perhaps for searching the position for exploitation. Therefore, considering the jumps, avoid from stuck in the local optima points. This is formulated as follows:

$$\vec{P}_i = \begin{cases}\vec{P}_i + \text{CF} \times \left(\vec{X}_{\min} + \vec{R} \otimes (X_{\max} - X_{\min})\right) \otimes \vec{U} & \text{if } r \le p_f,\\ \vec{P}_i + \left(p_f \times (1 - r) + r\right) \times \left(\vec{P}_{r_1} - \vec{P}_{r_2}\right) & \text{if } r > p_f,\end{cases} \tag{13}$$

where p_f describes the impact of FADs and is considered 0.2 in this study, $\vec{U}$ describes the binary vector with arrays in the range [0, 1], r describes a randomly distributed value between 0 and 1, r_1 and r_2 represent random indices of the prey matrix, and $X_{\min}$ and $X_{\max}$ represent the vector connecting the minimum and the maximum bounds of dimensions.

4. Optimized CNN

In the present study, we used an optimized technique to improve the efficiency of the CNN architecture and implement a good relationship between the layers for guaranteeing a suitable diagnosis system for SARS-COV-2. As we know, the original CNN uses a gradient descent algorithm for optimizing the model parameters, which includes convolution filters and the weights of fully connected layers. Due to the significance of the last layer in classification, assigning the image into a related class is significant that is accomplished by a proper connection between the weights and the previous layers. To improve the accuracy of the diagnosis system, the last weight vector training should be optimized based on the proposed marine predator optimization algorithm. The number of candidates and the iteration number for the algorithm are considered 100 and 120, respectively.

The objective function for minimizing the CNN is mathematically formulated as follows:

$$E = \frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ji} - O_{ji}\right)^2, \tag{14}$$

where N describes the number of training samples, M represents the number of output layers, and Y_{ji} and O_{ji} represent the desired value and the output value of the CNN.

lHere, the half-value precision function has been established for validation of the optimized diagnosis system. The algorithm then starts to optimize the CNN structure until the stopping criteria have been obtained. The designed system is then validated and verified on a dataset based on the Mean Square Error (MSE). Then, the MSE has been minimized by optimal selection of the weights and biases, i.e.,

$$\begin{aligned}W &= \left(w_1, w_2, \cdots, w_p\right),\\ b_n &= (b_{1n}, b_{2n}, \cdots, b_{Ln}),\\ l &= 1, 2, \cdots, L,\\ n &= 1, 2, \cdots, A,\\ A &= (a_1, a_2, \cdots, a_A),\\ w_n &= (w_{1n}, w_{2n}, \cdots, w_{Ln}),\end{aligned} \tag{15}$$

where l describes the layer index, A defines the total number of candidates, w_{in} represents the value of the weights in the ith layer, L signifies the total number of layers, and n defines the number of the candidates.

5. Dataset Description

The method of authentication has been presented by a standard test case of SARS-CoV-2 dataset. Numerous datasets are proposed for the diagnosis of SARS-CoV-2. The

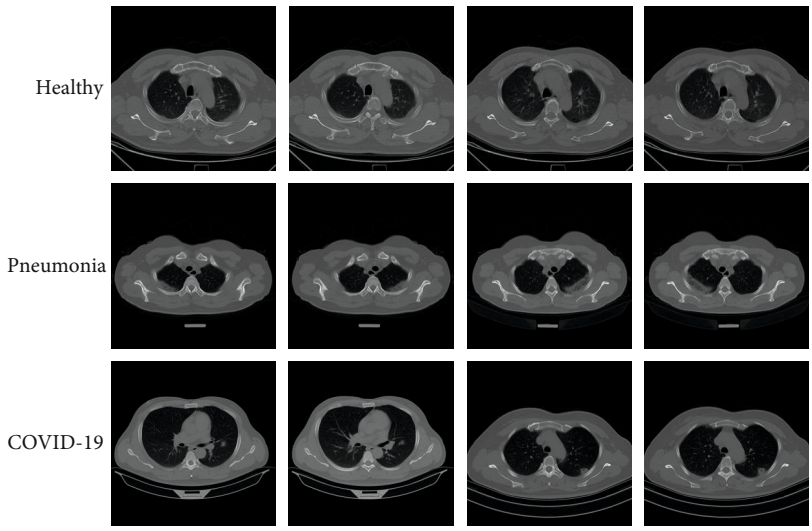


Figure 3: Some examples of the CT scan images collected from the MosMedData dataset [25].

presented study uses chest CT scans with SARS-CoV-2-related findings (MosMedData) for the analysis [25]. The dataset has been collected by the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department (MosMed). 1110 patients are analyzed based on NIfTI format. Figure 3 shows some examples of the CT scan images collected from the dataset.

After data acquisition from the dataset, to improve the quality of the raw data for statistical analysis and for increasing the accuracy of the system, some preprocessing has been done on the raw data. The first preprocessing step is data conversion. This process is a mathematical method employed for modifying variables that do not follow the statistical assumptions of linearity, normality, and uniform scattering or have patterns with uncommon outliers.

Here, data normalization has been employed. This process normalizes data/variables and puts data in the same domain when they are not. In this study, the Min–Max method has been used for the normalization. Based on the Min–Max method, unifying data scale, the data changing edges will be distributed between 0 and 1. Considering attribute X, so that it has a mapping from the dataset in the range $[X_{\min}, X_{\max}]$, the Min–Max normalization ($\bar{X}$) is mathematically given as follows:

$$\bar{X} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}. \tag{16}$$

6. Simulation Results

The present study implements the training process and the proposed COVID-19 diagnosis system on MATLAB 2019b. The system configuration for the computation is Windows 10 Enterprise with Intel® Core™ i7-4720HQ, 1.60 GHz, 16 GB RAM with Intel HD GPU 4600. The main idea is to introduce a new system for the diagnosis of COVID-19. The system is assessed by four measurement indicators that contain precision, accuracy, sensitivity, and $F1$ score.

6.1. Accuracy. The accuracy is a measurement indicator for achieving the rate of similarity of the image with the real value. This is established by the proportion of correct identification values to the total number of identifications. This indicator is mathematically obtained as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^{l}(\text{TP}_i + \text{TN}_i)}{\sum_{i=1}^{l}(\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i)}, \tag{17}$$

where TN and FN define the true negative and false negative, respectively, and TP and FP describe the true positive and false positive, respectively.

6.2. Precision. Precision describes the way of similarity of the measured values to each other. This indicator is established based on the proportion of positive identification values to the total number of identifications. This is mathematically defined by the following equation:

$$\text{Precision} = \frac{\sum_{i=1}^{l}(\text{TP}_i + \text{FP}_i)}{\sum_{i=1}^{l}(\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i)}. \tag{18}$$

6.3. Sensitivity. This indicator shows the extent of positives that are accurately detected. The sensitivity is established by the proportion of true-positive recognition values to the true-positive and false-negative number of recognition. This is mathematically modeled as follows:

$$\text{Sensitivity} = \frac{\sum_{i=1}^{l}\text{TP}_i}{\sum_{i=1}^{l}(\text{TP}_i + \text{FN}_i)}. \tag{19}$$

6.4. F1 Score. This score defines the exactness of the degree of a test set. This measure is achieved by the sensitivity and precision of the test. The most notable value of an F score is 1, which indicates idealized exactness and review, and the least conceivable value is 0, with the chance that either the precision or sensitivity is 0. The $F1$ score is

Table 1: The accuracy results using different techniques.

Epochs	Proposed method	Horry et al.'s [26]	Li et al.'s [27]	Ahuja et al.'s [1]
100	96.12	94.51	92.17	89.22
200	97.05	95.76	93.61	90.43
300	97.16	96.44	94.29	91.39
400	98.32	96.81	94.14	92.08
500	98.11	96.38	95.27	92.19

Table 2: The precision results using different techniques.

Epochs	Proposed method	Horry et al.'s [26]	Li et al.'s [27]	Ahuja et al.'s [1]
100	95.35	91.94	93.46	85.67
200	96.25	91.22	94.17	85.29
300	96.34	92.19	95.08	86.34
400	97.83	93.97	96.33	87.11
500	98.13	94.26	97.39	88.09

Table 3: The sensitivity results using different techniques.

Epochs	Proposed method	Horry et al.'s [26]	Li et al.'s [27]	Ahuja et al.'s [1]
100	96.37	93.46	91.11	85.04
200	96.18	94.29	92.37	87.16
300	97.80	95.81	92.69	87.26
400	97.59	96.53	94.08	88.68
500	98.66	96.74	94.16	89.37

Table 4: The $F1$ score results using different techniques.

Epochs	Proposed method	Horry et al.'s [26]	Li et al.'s [27]	Ahuja et al.'s [1]
100	96.38	93.68	91.28	89.33
200	96.15	94.07	92.15	90.27
300	97.29	95.39	93.69	91.26
400	97.26	96.13	95.09	92.43
500	98.34	97.35	95.02	92.56

moreover recognized as the Dice similarity coefficient (DSC) and is mathematically formulated as follows:

$$F1_{\text{score}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (20)$$

The analysis results of the defined indicators are reported in Tables 1–4. This technique is compared with three state-of-the-art techniques including Horry et al.'s [26], Li et al.'s [27], and Ahuja et al.'s [1] for better clarification.

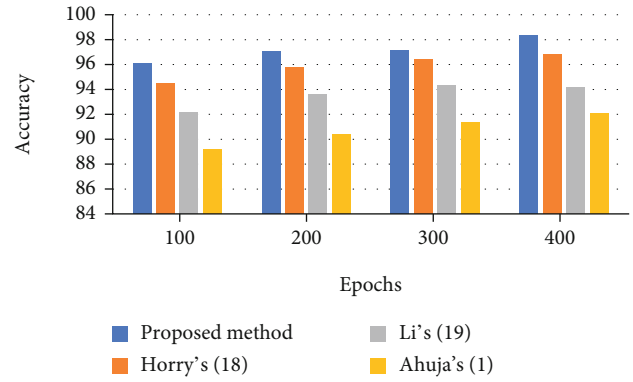


Figure 4: The accuracy bar plot for the assessed algorithms.

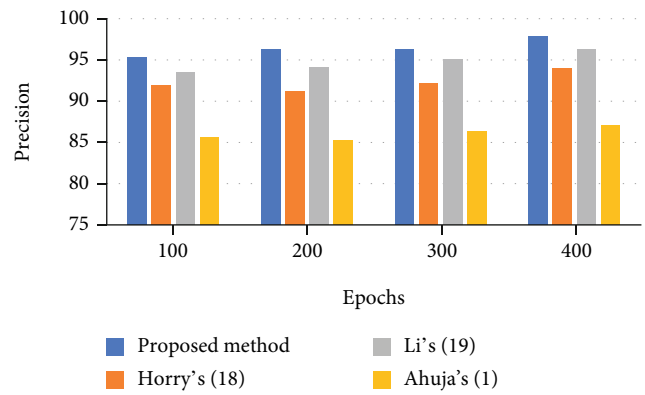


Figure 5: The precision bar plot for the assessed algorithms.

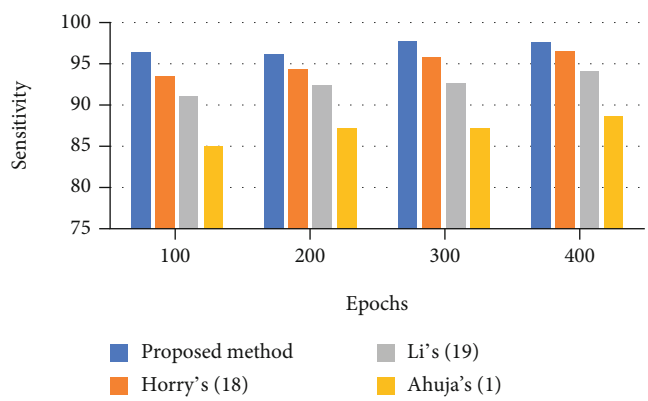


Figure 6: The sensitivity bar plot for the assessed algorithms.

Accuracy is 98.11%, precision is 98.13%, sensitivity is 98.66%, and $F1$ score is 97.26%.

To provide better observation of the system effectiveness, a bar plot of the results is shown in Figures 4–7. It can be observed from Figures 4–6 that there is 98.32% accuracy, 97.83% precision, and 98.66% sensitivity for the proposed technique after 400 epochs compared with the other investigated methods. However, Horry et al.'s, Li et al.'s, and Ahuja et al.'s are in the next ranks.

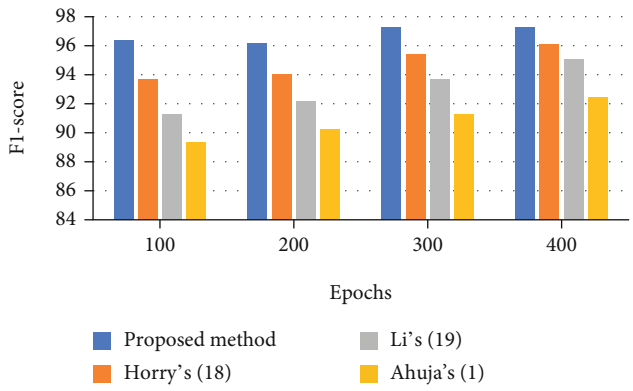


Figure 7: The $F1$ score bar plot for the assessed algorithms.

400 epochs have been implemented for the algorithm. As it is observed from the figures, the suggested method provides better sensitivity to the other comparative methods. The proposed classifier provides a 97.59% sensitivity rate, whereas Horry et al.'s, Li et al.'s, and Ahuja et al.'s have 96.53%, 94.08%, and 8.68%, respectively, for 400 epochs. Figure 7 shows the $F1$ score bar plot for the assessed algorithms.

It is also observed that after 400 epochs, the proposed method provides the highest $F1$ score value than the other comparative methods. As can be observed, the proposed technique with a 97.26% $F1$ score value offers the highest F measure, and Horry et al.'s, Li et al.'s, and Ahuja et al.'s with 96.13%, 95.09%, and 92.43%, respectively, are in the next ranks.

7. Conclusions

The COVID-19 pandemic continues as a dangerous problem for worldwide health. One significant way to stop this pandemic is to diagnose the infected patients efficiently and execute instant isolation. The infected patients with the SARS-CoV-2 virus can be detected by CT images. In the present study, a method based on optimized convolutional neural network based on metaheuristic technique was proposed for proper diagnosis of the COVID-19 CT scan images. The method used a newly introduced metaheuristic called the marine predator optimization algorithm to improve the accuracy of the proposed CNN-based diagnosis system. The proposed method was then performed on the chest CT images with COVID-19-related findings (MosMedData) dataset. Simulation results of the proposed system were compared with three other state-of-the-art methods including Horry et al.'s, Li et al.'s, and Ahuja et al.'s to indicate the method's effectiveness. Final results indicate that the proposed method with 98.11% accuracy, 98.13% precision, 98.66% sensitivity, and 97.26% $F1$ score showed the highest performance in all indicators than the compared methods. In the future work, we will work on applying a modified version of the proposed technique on chest X-ray images to determine the capability of the proposed method for the diagnosis of COVID-19 based on X-ray images and CT scan images.

Data Availability

Chest CT scans with COVID-19-related findings (MosMedData) 2020 are available from https://mosmed.ai/datasets/covid19_1110.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] S. Ahuja, B. K. Panigrahi, N. Dey, V. Rajinikanth, and T. K. Gandhi, "Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices," *Applied Intelligence*, vol. 51, no. 1, pp. 571–585, 2021.

[2] N. Razmjooy, M. Ashourian, M. Karimifard et al., "Computer-aided diagnosis of skin cancer: a review," *Current Medical Imaging*, vol. 16, 2020.

[3] H. Maghdid, A. T. Asaad, K. Z. G. Ghafoor, A. S. Sadiq, S. Mirjalili, and M. K. K. Khan, "Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms," in *Multimodal Image Exploitation and Learning 2021*, Florida, United States, 2021.

[4] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Jamalipour Soufi, "Deep-covid: predicting COVID-19 from chest X-ray images using deep transfer learning," *Medical Image Analysis*, vol. 65, article 101794, 2020.

[5] M. Z. Islam, M. M. Islam, and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," *Informatics in Medicine Unlocked*, vol. 20, p. 100412, 2020.

[6] Y. Karadayi, M. N. Aydin, and A. S. Öğrencí, "Unsupervised anomaly detection in multivariate spatio-temporal data using deep learning: early detection of COVID-19 outbreak in Italy," *IEEE Access*, vol. 8, article 164155, 164177 pages, 2020.

[7] Z. Guo, L. Xu, Y. Si, and N. Razmjooy, "Novelcomputer-aidedlung cancer detection based onconvolutional neural network-based and feature-based classifiers using metaheuristics," *International Journal of Imaging Systems and Technology*, 2021.

[8] K. Roy, K. K. Mandal, and A. C. Mandal, "Ant-lion optimizer algorithm and recurrent neural network for energy management of micro grid connected system," *Energy*, vol. 167, pp. 402–416, 2019.

[9] N. Razmjooy, F. R. Sheykhahmad, and N. Ghadimi, "A hybrid neural network–world cup optimization algorithm for melanoma detection," *Open Medicine*, vol. 13, no. 1, pp. 9–16, 2018.

[10] F. Koehler and A. Risteski, "Representational power of ReLU networks and polynomial kernels: beyond worst-case analysis," 2018, https://arxiv.org/abs/1805.11405.

[11] B. Van Merriënboer, D. Bahdanau, V. Dumoulin et al., "Blocks and fuel: frameworks for deep learning," 2015, https://arxiv.org/abs/1506.00619.

[12] L. Xie and A. Yuille, "Genetic cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017.

[13] M. Ramezani, D. Bahmanyar, and N. Razmjooy, "A new improved model of marine predator algorithm for optimization problems," *Arabian Journal for Science and Engineering*, vol. 46, no. 9, pp. 8803–8826, 2021.

[14] Y. Cao, Y. Li, G. Zhang, K. Jermsittiparsert, and N. Razmjooy, "Experimental modeling of PEM fuel cells using a new improved seagull optimization algorithm," *Energy Reports*, vol. 5, pp. 1616–1625, 2019.

[15] Y. Cao, Y. Wu, L. Fu, K. Jermsittiparsert, and N. Razmjooy, "Multi-objective optimization of a PEMFC based CCHP system by meta-heuristics," *Energy Reports*, vol. 5, pp. 1551–1559, 2019.

[16] Z. Yuan, W. Wang, H. Wang, and N. Razmjooy, "A new technique for optimal estimation of the circuit-based PEMFCs using developed sunflower optimization algorithm," *Energy Reports*, vol. 6, pp. 662–671, 2020.

[17] Y. Guo, X. Dai, K. Jermsittiparsert, and N. Razmjooy, "An optimal configuration for a battery and PEM fuel cell-based hybrid energy system using developed krill herd optimization algorithm for locomotive application," *Energy Reports*, vol. 6, pp. 885–894, 2020.

[18] D. Yu, Y. Wang, H. Liu, K. Jermsittiparsert, and N. Razmjooy, "System identification of PEM fuel cells using an improved Elman neural network and a new hybrid optimization algorithm," *Energy Reports*, vol. 5, pp. 1365–1374, 2019.

[19] N. Razmjooy, V. V. Estrela, R. Padilha, and A. C. B. Monteiro, *World cup optimization algorithm: an application for optimal control of pitch angle in hybrid renewable PV/wind energy system*, Springer, 2018.

[20] L. Abualigah, A. Diabat, S. Mirjalili, M. Abd Elaziz, and A. H. Gandomi, "The arithmetic optimization algorithm," *Computer Methods in Applied Mechanics and Engineering*, vol. 376, article 113609, 2021.

[21] M. Mani, O. Bozorg-Haddad, and X. Chu, "Ant lion optimizer (ALO) algorithm," in *Advanced Optimization by Nature-Inspired Algorithms*, Springer, 2018.

[22] A. Faramarzi, M. Heidarinejad, B. Stephens, and S. Mirjalili, "Equilibrium optimizer: a novel optimization algorithm," *Knowledge-Based Systems*, vol. 191, article 105190, 2020.

[23] A. Faramarzi, M. Heidarinejad, S. Mirjalili, and A. H. Gandomi, "Marine predators algorithm: a nature-inspired metaheuristic," *Expert Systems with Applications*, vol. 152, article 113377, 2020.

[24] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, M. G. E. da Luz, E. P. Raposo, and H. E. Stanley, "Optimizing the success of random searches," *Nature*, vol. 401, no. 6756, pp. 911–914, 1999.

[25] Chest CT, "Scans with COVID-19 related findings (MosMedData)," 2020, https://mosmed.ai/datasets/covid19_1110.

[26] M. J. Horry, S. Chakraborty, M. Paul et al., "COVID-19 detection through transfer learning using multimodal imaging data," *IEEE Access*, vol. 8, pp. 149808–149824, 2020.

[27] K. Li, Y. Fang, W. Li et al., "CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19)," *European Radiology*, vol. 30, no. 8, pp. 4407–4416, 2020.

Hindawi
BioMed Research International
Volume 2021, Article ID 2204021, 9 pages
https://doi.org/10.1155/2021/2204021

Research Article

Improved Estimation of Bio-Oil Yield Based on Pyrolysis Conditions and Biomass Compositions Using GA- and PSO-ANFIS Models

Zhimin Li,[1] **Deyin Zhao**,[1] **Linbo Han**,[2] **Li Yu**,[2] **and Mohammad Mahdi Molla Jafari**[3]

[1]*Research Institute of Petroleum Engineering and Technology, Sinopec Northwest Oilfield Company, Urumqi 830011, China*
[2]*College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen 518118, China*
[3]*Department of Petroleum Engineering, Ahwaz Faculty of Petroleum Engineering, Petroleum University of Technology (PUT), Ahwaz, Iran*

Correspondence should be addressed to Li Yu; yuli@sztu.edu.cn
and Mohammad Mahdi Molla Jafari; mohammad.molajafari@afp.put.ac.ir

Received 12 August 2021; Accepted 21 September 2021; Published 5 October 2021

Academic Editor: Alireza Baghban

This paper incorporates the adaptive neurofuzzy inference system (ANFIS) technique to model the yield of bio-oil. The estimation of this parameter was performed according to pyrolysis conditions and biomass compositions of feedstock. For this purpose, this paper innovates two optimization methods including a genetic algorithm (GA) and particle swarm optimization (PSO). Primary data were gathered from previous studies and included 244 data of biodiesel oils. The findings showed a coefficient determination (R^2) of 0.937 and RMSE of 2.1053 for the GA-ANFIS model, and a coefficient determination (R^2) of 0.968 and RMSE of 1.4443 for PSO-ANFIS. This study indicates the capability of the PSO-ANFIS algorithm in the estimation of the bio-oil yield. According to the performed analysis, this model shows a higher ability than the previously presented models in predicting the target values and can be a suitable alternative to time-consuming and difficult experimental tests.

1. Introduction

Bioenergy is by far the most successful and sustainable future path [1]. The primary source of energy today is fossil fuels that have enormously negative environmental consequences, causing many issues around the world [2]. Thankfully, biomass energy with neutral carbons is a viable means of addressing both energy needs and environmental issues [3, 4]. In addition, a massive quantity of potential supplies is analyzed each year [5–7]. Thermochemical and biochemical conversions are the procedures appropriate for efficient biomass consumption, which are currently being researched. Thermochemical conversion has drawn the attention of researchers in recent years due to the elevated level of conversion performance besides minimal costs [8]. Pyrolysis is a thermochemical conversion method that involves heating feedstock in an inert environment or oxygen-deficient atmosphere to generate biochar, bio-oils, and noncondensable gas [9]. Bio-oils are liquid substances that typically contain over 350 chemicals, including several materials in short supply [10]. Furthermore, provided bio-oil is properly improved, it has the potential to be a viable alternative energy source, compared to fossil fuels. Moreover, the bio-oil hydrogen content typically represents the heating efficiency and chemical composition (i.e., bio-oil efficiency), whereas the yield refers to the amount of bio-oil. The quality of bio-oils and their quantity are primarily determined by biomass feedstock as well as pyrolysis circumstances [11]. Using proximate and ultimate analyses can generally produce data on various biomasses. The proximate analysis could be used to establish the concentration of fixed carbon, ash, and volatile material. The organic compartment in the biomass is determined by the amount of fixed carbon and volatile material, whereas the ash usually reflects inorganic salts. In the

meantime, the content of basic elements (i.e., C–H–N–O) is primarily defined by ultimate analysis. The pyrolysis circumstances include the temperature, the size of the particles, heating rate, and residence period during the pyrolysis process. As a result, several inquiries have been launched in the field of study. Akhtar and Amin reported that the intermediate pyrolysis temperature (500–550°C) normally increased the yield of bio-oil to the maximum extent [12]. Gholizadeh and colleagues reported that the production of bio-oil from herbaceous biomass was smaller compared with that obtained from woody biomass. In addition, the amount of hydrogen in bio-oil was usually greater than in feedstock [13]. Chiodo et al. realized that bio-oils extracted from woody biomass possess more amounts of hydrogen than that from algae, resulting in a greater thermal output [14]. Nonetheless, the relationship between bio-oil characteristics containing biomass compositions and operational parameters remains unclear, due to experimental and financial constraints. The linear regression approach is the most commonly used method for detecting variable correlation. Li et al. investigated the relationship between the distribution of bio-oil compounds and feedstock features using linear regression [15]. Oasmaa et al. established the association between the organic and ash amounts [16]. While the output of linear regression can be undesirable in the presence of a nonlinear association between variables, after the emergence of artificial intelligence, several new approaches were applied to conventional studies and yielded suitable results [17–24]. Cao et al. used a least-squares support vector machine (LM-SVM) and an artificial neural network (ANN) to reliably estimate biochar yields from cattle manure [25]. Sun et al. used the Levenberg Marquardt ANN approach to specifically assess the significance of every variable for the gas yield [26]. Satisfactorily, the ANN method was used by Naqvi et al. to research the mechanism of the reaction according to data related to copyrolysis thermal decomposition [27]. SVM and ANN models were developed by Xing et al. to thoroughly make an estimation of the biomass heating rate by proximate and ultimate analyses [28]. The entire models aided researchers in evaluating a specific outcome without running tests, besides expanding their understanding of the biomass pyrolysis mechanism. Nonetheless, these models were mostly concerned with estimation, leaving the finer knowledge to be retrieved. Random forest (RF) is defined as an ensemble study approach focused on tree predictors that can solve regression and classification problems [29]. Zhu et al. skillfully and accurately predicted biochar yields by the use of the RF approach, and, at the same time, they established associations between biochar production, biomass structural details, and pyrolysis circumstances [30]. Using the RF model, Xing et al. accurately predicted the biomass chemical composition from the ultimate analysis [31]. Due to the properties of the ensemble analysis, the RF approach can achieve higher training rates and superior productivity than other estimation techniques. Further, high-dimensional properties and feature correlations can be addressed and established using the RF procedure.

In this paper, for the first time, attempts have been made to estimate models using the two models GA- and PSO-ANFIS. For this purpose, first, the relevant input data affecting the output parameter were collected, and then, this issue was modeled. Finally, in order to evaluate the strength of these models, various statistical analyses were used.

2. Theory

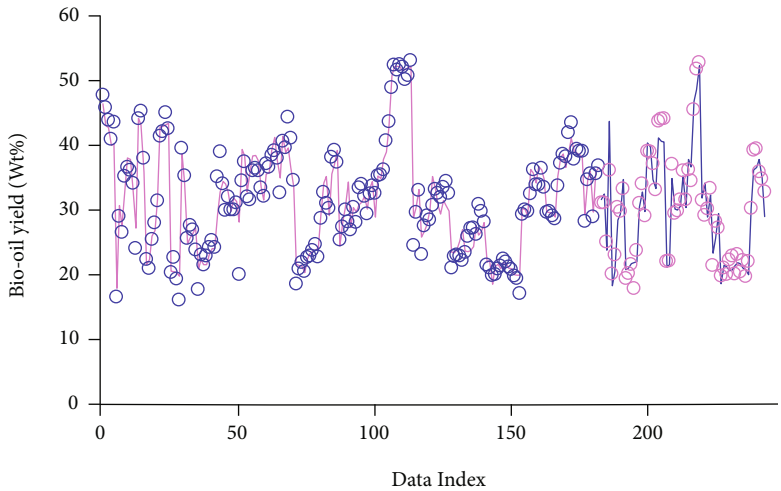
2.1. The Adaptive Neurofuzzy Inference System. As a general guideline, a Takagi-Sugeno fuzzy rule and input-output variables form the basis of an adaptive neurofuzzy inference system (ANFIS). Generally, an adaptive neurofuzzy inference system (ANFIS) involves input-output variables and a Takagi-Sugeno fuzzy rule [32–34]. An adaptive, multilayer, and feed-forward network described by ANFIS can be simplified by expressing it as two inputs (x, y) and one output (z). The ANFIS model is an adaptive, multilayer, and feed-forward network that, for the sake of simplicity, can be expressed with two inputs (x, y) and one output (z). Following that, two different if-then fuzzy rules are set for a first-order Sugeno fuzzy model to determine the matching principle. Next, the matching principle is set with two different if-then fuzzy rules for a first-order Sugeno fuzzy model:

$$\begin{aligned} &\text{Rule 1 : if } x \text{ is } A_1 \text{ and } y \text{ is } B_1, \text{ then } Z_1 = P_1 x + q_1 y + r_1, \\ &\text{Rule 2 : if } x \text{ is } A_2 \text{ and } y \text{ is } B_2, \text{ then } Z_2 = P_2 x + q_2 y + r_2. \end{aligned} \tag{1}$$

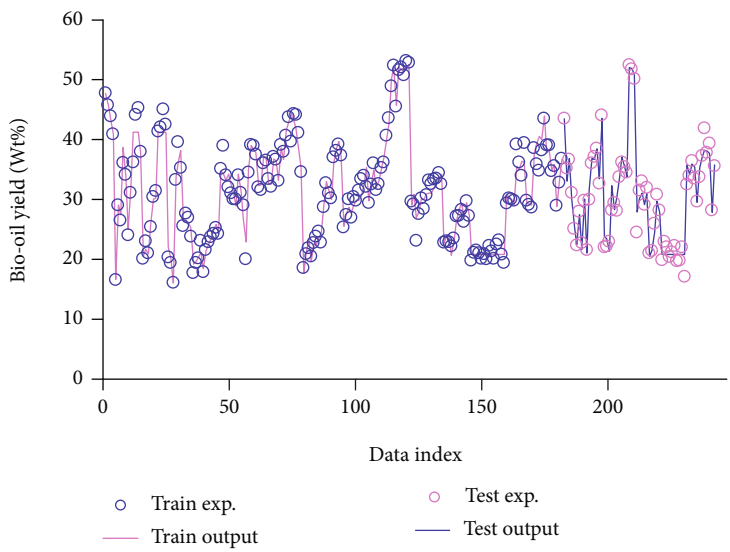
This equation evaluates entries through linguistic B1 as well as A1 variable entries of this equation are evaluated through linguistic A1 and B1 variables. In order to calculate the outcome of every rule, the inputs with the constant term (r) can be linearly combined. The results of each rule can be calculated using a linear combination of the inputs with the constant term (r). There are five layers in ANFIS architecture, with the first layer undergoing fuzzification to map the x and y variables (inputs) into fuzzy sets (that is, A1, A2, B1, and B2). The membership grades for square nodes are determined by node functions. Each square node generates membership grades using node functions. An alternative to linguistic labels (for example, high and low) uses symbols such as A and B. Instead of linguistic labels (e.g., high and low), characters such as A and B are used. The labels are classified according to their membership functions; for example, different membership functions serve to characterize the labels, e.g., the sigmoid, triangular, and generalized bell functions. Various sets of fuzzy inputs and firing strength are used in layer two. There exist combinations of fuzzy sets of inputs in layer two and the use of firing strength. The fuzzy conjunction "and" is successfully utilized by the G-norm operator to locate the output in this layer. In this layer, a G-norm operator successfully performs the fuzzy conjunction "and" to find the output. Calculation of the ith rule ratio is done at the third layer. The third layer involves calculating the ratio of the ith rule. At the fourth layer, a function of the Sugeno fuzzy rule is multiplied by the output of the three previous layers. Next, in the fourth layer, the output of the three former layers is multiplied by the function of the Sugeno fuzzy rule. One node in the fifth and last layer is responsible for computing and summarizing

Table 1: The values of different statistical parameters obtained for the models.

Model	Phase	R^2	MRE (%)	MSE	RMSE	STD
GA-ANFIS	Train	0.937	5.077	4.244909156	2.0603	1.4186
	Test	0.937	5.693	4.432311766	2.1053	1.3085
	Total	0.937	5.231	4.291759808	2.1053	1.3910
PSO-ANFIS	Train	0.968	3.323	2.180267641	1.4766	1.0671
	Test	0.969	3.876	2.086124383	1.4443	0.9936
	Total	0.968	3.461	2.156731826	1.4443	1.0473



(a)



(b)

Figure 1: Simultaneous and visual comparison between actual and modeled output data for models (a) GA-ANFIS and (b) PSO-ANFIS.

each rule output from the previous layer. The fifth and final layer, which contains a single node, involves the summation and calculation of the outputs associated with each rule from the fourth layer. The next step is the application of the weight average method to perform defuzzification. Next, the weight averaged approach is incorporated to carry out the defuzzification process. During this process, fuzzy outputs are transformed into crisp ones. This process results in a crisp output by transforming the fuzzy outputs. ANFIS parameters fall into the consequent and the premise parts depending on whether linear or nonlinear parameters are used. ANFIS parameters can be classified into two categories: linear parameters in the consequent part and nonlinear in the premise part. The parameters can be optimized by gradient descent or steepest descent, among other methods. These parameters can be optimized through a variety of

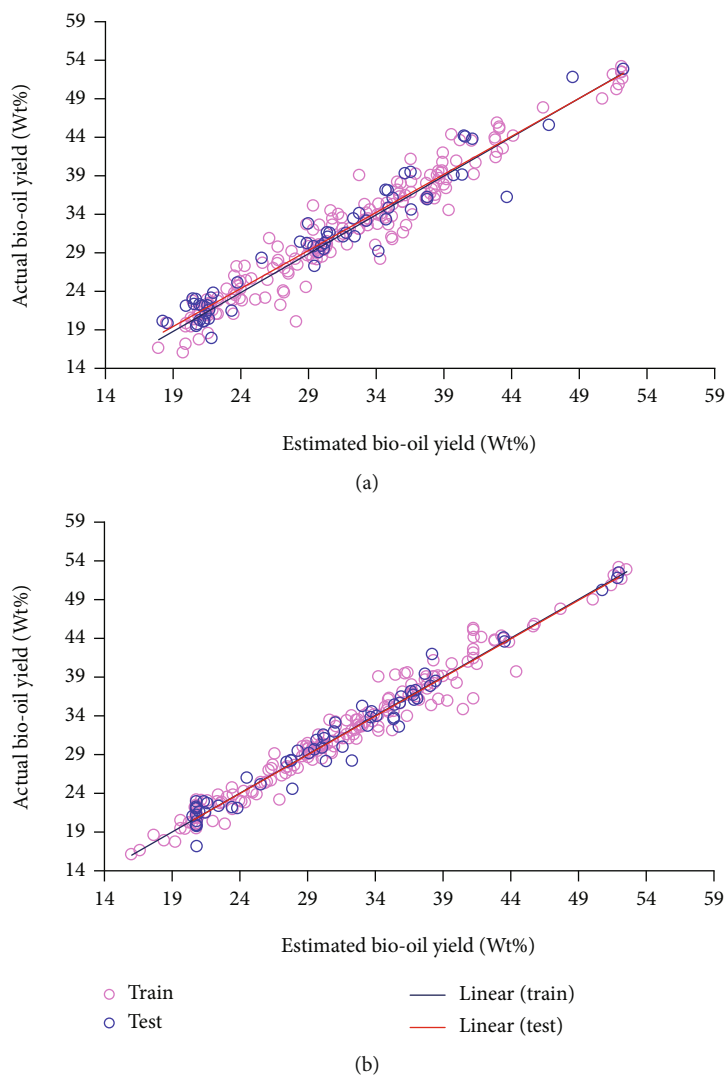


Figure 2: Cross-plot diagrams obtained using different models: (a) GA-ANFIS and (b) PSO-ANFIS.

methods, such as gradient descent and steepest descent methods. However, the hybrid learning method is much more effective. Yet, much higher efficiency can be achieved through the hybrid learning method [32].

2.2. Particle Swarm Optimization (PSO). The fundamental knowledge for PSO came from configuring natural populations (e.g., birds) [35, 36]. In PSO, the optimizing problem is the particle and the answer is obtained through generation update. The swarm denotes the total number of particles [37]. This way, the particle is considered as an individual and the swarm as a population. The above expressions also exist in most other evolutionary methods, such as genetic algorithms (GA) [38]. However, the evolutionary type operators (e.g., mutations) do not exist in PSO [37]. Particles, during the process of finding the optimal answer, search for the problem domain and, in the meantime, are affected by their topological neighborhoods (e.g., queen, physical, and social) [39]. Equation (2) calculates the ith particle velocity. In this equation, $vi(t)$ depicts the velocity vector and $xi(t)$ represents the position vector [38, 40].

$$v_{id}(t+1) = c_1 r_1 (p_{\text{best}}, id(t)) - x_{id}(t)) + w v_{id}(t) + c_2 r_2 (g_{\text{best}}, id(t) - x_{id}(t)), d = 1, 2, \cdots, D. \quad (2)$$

Additionally, p_{best}, id represents the best position, w is representative of the inertia's weight, and g_{best}, id represents the best global position of the ith particle. Random coefficients are represented by r_1 and r_2, together with the degrees of learning by c_1 and c_2 [41]. In Equation (2), the first term is a cognitive element directing the movements in particles and the second term denotes the previous movement route memory, and finally, the last term serves to assess the particle action in comparison with its neighborhood [35, 38]. Equation (3) provides an integrating process that helps calculate the position vector.

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1), d = 1, 2, \cdots, D. \quad (3)$$

2.3. Genetic Algorithm (GA). An evolutionary heuristic algorithm such as GA imitates the natural process of evolution to optimization. To resolve optimization issues, the GA can be

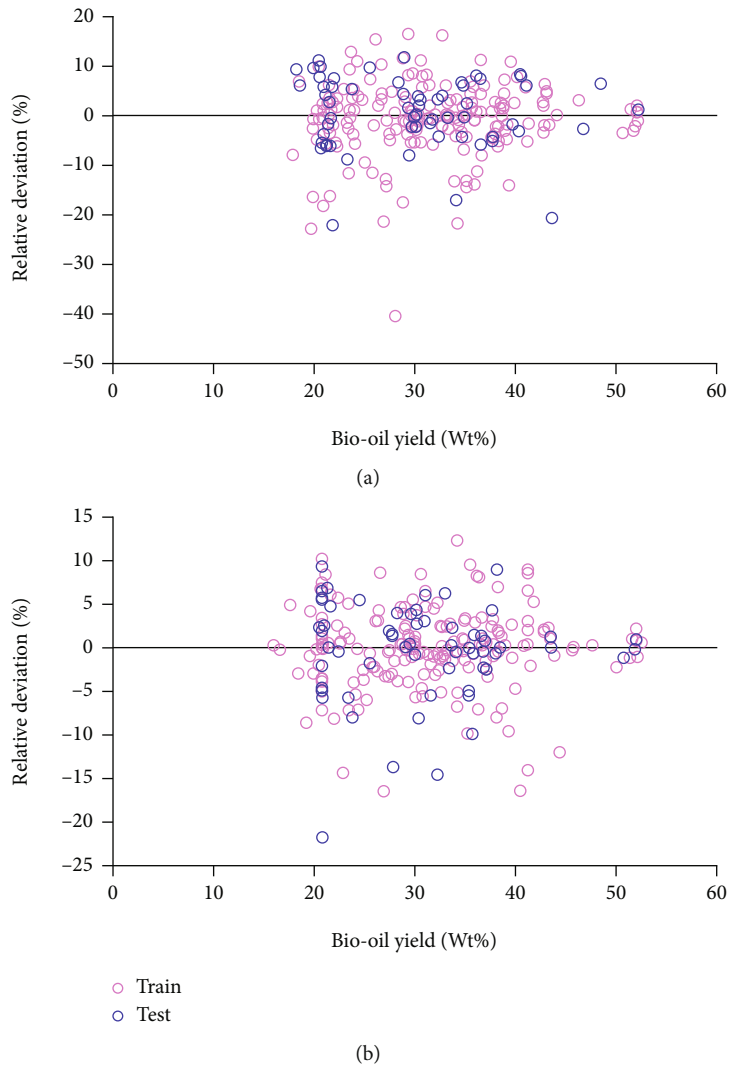


Figure 3: Relative derivation diagrams of (a) GA-ANFIS and (b) PSO-ANFIS models to evaluate their accuracy.

used to calculate the best solution. Holland developed GA when he utilized a common functional framework in 1975 [42, 43]. The development of the algorithm was inspired by Darwin's natural selection theory. In fact, the GA method makes it possible to renew the genetic behavior observed in a biological population. Generally, chromosomes, also known as individuals, are referred to as candidate answers to a particular problem in the GA which typically comprises a linear array of genes. By randomly using the generated design populations, the search process is started. The search process does not require the definition of starting points because it is iterative. In the GA technique, the multiplication from one generation to the subsequent generation is performed by three operators during the optimization. When GA takes into account the theory of a greater chance of survival in order to generate design solutions, the "Selection" operator is the first operand. At all stages of the process of selection, these solutions must be compatible with their environment. "Crossover" is the second operator, and it triggers mating among the biological populations. Crossover operators ensure that fitting surviving characteristics are transferred from the current to subsequent populations. With this method, it is more likely that arbitrarily surviving will be included in the population. "Mutation" is the third operator responsible for creating heterogeneity in the characteristics of the population. According to Dutch (1975), Hasan, and Cohanim et al. (2005), mutation operator performs the worldwide search in the search space and also does not allow the genetic algorithm located in local minima.

Table 2: Statistical comparison of the performance of different models in assessing the target values.

Model	R^2	RMSE
RF	0.87	3.05
MLR	0.284	7.96
PSO-ANFIS	0.968	1.4443

3. Data Bank

From previous researches, a total of 244 samples involving biodiesel oil yield on the basis of pyrolysis conditions and

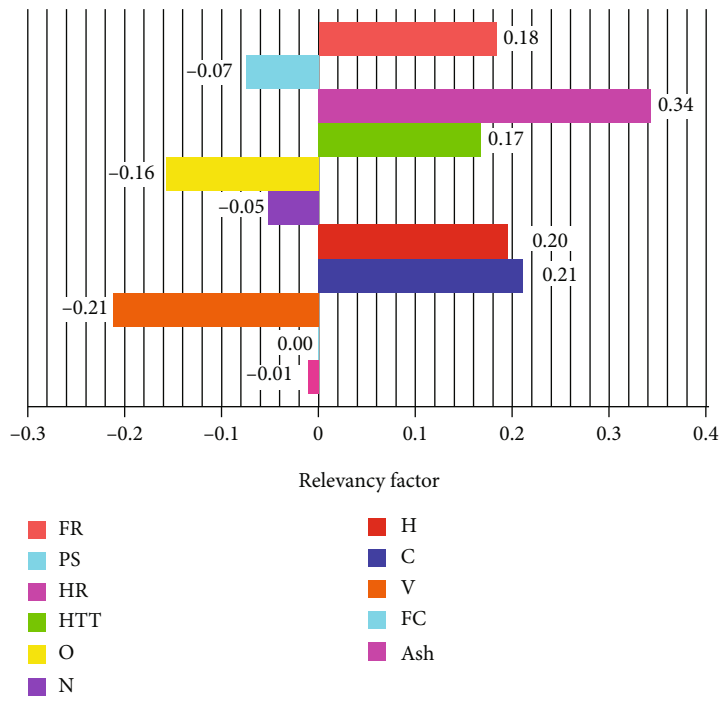


Figure 4: Sensitivity diagram on all input parameters affecting the output parameter.

biomass compositions of feedstock were gathered [44]. The samples were categorized into a training cluster (183 samples) and a test group (61 samples).

4. Model Results

Methods such as STD, MSE, RMSE, MRE %, and R^2 were used to analyze the obtained yield values (Table 1) against real data. The statistical parameters were derived from the formulas as follows [45–49]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left[x_i^{\text{sim}} - x_i^{\text{exp}}\right]^2}{\sum_{i=1}^{n}\left[x_i^{\text{sim}} - x^m\right]^2}, x_m = \frac{\sum_{i=1}^{n} x_i^{\text{exp}}}{n}, \quad (4)$$

$$\text{MRE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|x_i^{\text{exp}} - x_i^{\text{sim}}\right|}{x_i^{\text{exp}}}, \quad (5)$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^{\text{exp}} - x_i^{\text{sim}}\right)^2, \quad (6)$$

$$\text{RMSE} = \left(\frac{1}{N}\sum_{i=1}^{N}\left(x_i^{\text{exp}} - Y x_i^{\text{sim}}\right)^2\right)^{0.5}, \quad (7)$$

$$\text{STD} = \sqrt{\sum_{i=1}^{n}\left(\frac{\left(x_i^{\text{sim}} - x_m\right)^2}{n}\right)}. \quad (8)$$

In Equations (4)–(8), the character x_i^{exp} denotes the experimental target value, and x_i^{sim} represents the simulated value. The number of experimental data is shown by n. Table 1 displays the data calculated for these parameters. A more favorable model has smaller RMSE, MRE, MSE, STD, and larger R^2. As is observed, the GA-ANFIS method is not as precise in training, testing, and total datasets compared with the PSO-ANFIS model (see Table 1).

Figure 1 outlines empirical data and the estimated yield values to represent predictive capability and the liability of the models. As can be observed, the concordance between the obtained and actual data regarding the efficiency of the models is exceptional.

Figure 2 displays the yield values obtained through models and experimental data. It shows a nearly straight line at the angle of 45° which confirms the ability of the model in producing accurate results. The displayed data indicates a higher level of R^2 for the PSO-ANFIS model.

Figure 3 displays the relative derivations of both models. In estimating the yield values of diesel oils, the maximum absolute relative derivations of GA-ANFIS is 40% and of PSO-ANFIS is 23%. The corresponding values for biodiesel oils are 28 and 21, respectively. The statistical parameters indicate that the PSO-ANFIS model performs with the highest efficiency.

The present study, by comparing and assessing previous models developed by Tang et al. employing the same dataset on biodiesel oils [44], concluded that the PSO-ANFIS model performs more favorably in estimating the yield values. As seen in Table 2, this model boasts a more precise performance in estimation results than other models. The R^2 and RMSE values for different models are as follows.

4.1. Sensitivity Analysis. In Equation (9), the relevancy factor examines the input parameters affecting yield values [50, 51].

$$r = \frac{\sum_{i=1}^{n}\left(X_{k,i} - \bar{X}_k\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{k,i} - \bar{x}_k\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}. \quad (9)$$

The ith output is shown by Y_i, the output average is shown by $\bar{Y}$, $X_{k,i}$ denotes the kth input, and x_k denotes the input. The r value for each parameter is continually less than unity. The relevancy factor of biodiesel oil yield is shown in Figure 4. It can be observed that PS, O, N, and V have a negative effect on yield, and the effect of FR, HR, HTT, H, and C on the output is positive. This means that yield values of biodiesel oils are decreased by reducing the later parameters. In this figure, the relevancy factor of diesel oil yield is displayed. The largest impact on diesel oil yield is indicated for HR and the lowest for V.

5. Conclusion

The current paper is aimed at estimating the yield for diesel and biodiesel oils according to pyrolysis conditions and biomass compositions. For this purpose, the present study designed models using PSO-ANFIS and GA-ANFIS techniques and became the first to succeed in employing these techniques to estimate the output values. The PSO-ANFIS model boasts the most precise prognostication of target values. The raw data incorporated in this study was gathered from previous accredited researches, and statistical parameters (e.g., R^2, %MRE, RMSE, MSE, and STD) in association with graphical valuations were employed in the testing and training stages of the model development. The findings attest to the high quality of performance and accuracy of the proposed PSO-ANFIS model. Therefore, it can be used to estimate output values with high accuracy in all related industries and processes.

Data Availability

The data used to support the findings of this study are provided within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (Grant No. 61801301), General Support Projects of Shenzhen Colleges and Universities (Grant No. SZWD2021002), and the Natural Science Foundation of Top Talent of SZTU (Grant No. 2019203).

References

[1] S. Supriya, V. S. Bhat, T. J. Jayeoye, T. Rujiralai, K. F. Chong, and G. Hegde, "An investigation on temperature-dependant surface properties of porous carbon nanoparticles derived from biomass," *Journal of Nanostructure in Chemistry*, pp. 1–17, 2021.

[2] A. Coppola, "Latour and balloons: Gaïa Global Circus and the theater of climate change," *Configurations*, vol. 28, no. 1, pp. 29–49, 2020.

[3] P. Bharti, B. Singh, and R. Dey, "Process optimization of biodiesel production catalyzed by CaO nanocatalyst using response surface methodology," *Journal of Nanostructure in Chemistry*, vol. 9, no. 4, pp. 269–280, 2019.

[4] F. Khanbolouk, M. Akia, H. Arandiyan, F. Yazdani, and Y. Dortaj, "Utilization of spray-dried nanoporous gamma alumina support in biodiesel production from waste cooking oil," *Journal of Nanostructure in Chemistry*, vol. 7, no. 2, pp. 191–200, 2017.

[5] A. Demirbas, "Potential applications of renewable energy sources, biomass combustion problems in boiler power systems and combustion related environmental issues," *Progress in Energy and Combustion Science*, vol. 31, no. 2, pp. 171–192, 2005.

[6] C. Gokcol, B. Dursun, B. Alboyaci, and E. Sunan, "Importance of biomass energy as alternative to other sources in Turkey," *Energy Policy*, vol. 37, no. 2, pp. 424–431, 2009.

[7] R. A. Sheldon, "Green and sustainable manufacture of chemicals from biomass: state of the art," *Green Chemistry*, vol. 16, no. 3, pp. 950–963, 2014.

[8] M. Tripathi, J. N. Sahu, and P. Ganesan, "Effect of process parameters on production of biochar from biomass waste through pyrolysis: a review," *Renewable and Sustainable Energy Reviews*, vol. 55, pp. 467–481, 2016.

[9] A. Bridgwater and G. Peacocke, "Fast pyrolysis processes for biomass," *Renewable and Sustainable Energy Reviews*, vol. 4, no. 1, pp. 1–73, 2000.

[10] D. M. Alonso, J. Q. Bond, and J. A. Dumesic, "Catalytic conversion of biomass to biofuels," *Green Chemistry*, vol. 12, no. 9, pp. 1493–1513, 2010.

[11] Q. Zhang, J. Chang, T. Wang, and Y. Xu, "Review of biomass pyrolysis oil properties and upgrading research," *Energy Conversion and Management*, vol. 48, no. 1, pp. 87–92, 2007.

[12] J. Akhtar and N. S. Amin, "A review on operating parameters for optimum liquid oil yield in biomass pyrolysis," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 7, pp. 5101–5109, 2012.

[13] M. Gholizadeh, X. Hu, and Q. Liu, "A mini review of the specialties of the bio-oils produced from pyrolysis of 20 different biomasses," *Renewable and Sustainable Energy Reviews*, vol. 114, p. 109313, 2019.

[14] V. Chiodo, G. Zafarana, S. Maisano, S. Freni, and F. Urbani, "Pyrolysis of different biomass: direct comparison among Posidonia oceanica, lacustrine alga and white-pine," *Fuel*, vol. 164, pp. 220–227, 2016.

[15] J. Li, Y. Chen, H. Yang et al., "Correlation of feedstock and bio-oil compound distribution," *Energy & Fuels*, vol. 31, no. 7, pp. 7093–7100, 2017.

[16] A. Oasmaa, Y. Solantausta, V. Arpiainen, E. Kuoppala, and K. Sipilä, "Fast pyrolysis bio-oils from wood and agricultural residues," *Energy & Fuels*, vol. 24, no. 2, pp. 1380–1388, 2010.

[17] S. Anbazhagan, V. Thiruvengatam, and K. Kulanthai, "Adaptive neuro-fuzzy inference system and artificial neural network modeling for the adsorption of methylene blue by novel adsorbent in a fixed-bed column method," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 39, no. 6, pp. 75–93, 2020.

[18] A. Tarjomannejad, "Prediction of the liquid vapor pressure using the artificial neural network-group contribution method," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 34, no. 4, pp. 97–111, 2015.

[19] M. R. Ehsani, H. Bateni, and G. Razi Parchikolaei, "Modeling of oxidative coupling of methane over Mn/Na2WO4/SiO2 catalyst using artificial neural network," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 32, no. 3, pp. 107–114, 2013.

[20] A. Erdal Tümer, S. Edebali, and Ş. Gülcü, "Modeling of removal of chromium (VI) from aqueous solutions using artificial neural network," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 39, no. 1, pp. 163–175, 2020.

[21] B. Kavitha and D. Sarala Thambavani, "Artificial neural network optimization of adsorption parameters for Cr (VI), Ni (II) and Cu (II) ions removal from aqueous solutions by riverbed sand," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 39, no. 5, pp. 203–223, 2020.

[22] M. Esmaeili, M. Ahmadi, and A. Kazemi, "A generalized DEIM technique for model order reduction of porous media simulations in reservoir optimizations," *Journal of Computational Physics*, vol. 422, p. 109769, 2020.

[23] V. Vijayaraghavan, A. Garg, C. H. Wong, K. Tai, and Y. Bhalerao, "Predicting the mechanical characteristics of hydrogen functionalized graphene sheets using artificial neural network approach," *Journal of Nanostructure in Chemistry*, vol. 3, no. 1, pp. 1–5, 2013.

[24] B. Tanhaei, M. Esfandyari, A. Ayati, and M. Sillanpää, "Neuro-fuzzy modeling to adsorptive performance of magnetic chitosan nanocomposite," *Journal of Nanostructure in Chemistry*, vol. 7, no. 1, pp. 29–36, 2017.

[25] H. Cao, Y. Xin, and Q. Yuan, "Prediction of biochar yield from cattle manure pyrolysis via least squares support vector machine intelligent approach," *Bioresource Technology*, vol. 202, pp. 158–164, 2016.

[26] Y. Sun, L. Liu, Q. Wang, X. Yang, and X. Tu, "Pyrolysis products from industrial waste biomass based on a neural network model," *Journal of Analytical and Applied Pyrolysis*, vol. 120, pp. 94–102, 2016.

[27] S. R. Naqvi, Z. Hameed, R. Tariq et al., "Synergistic effect on co-pyrolysis of rice husk and sewage sludge by thermal behavior, kinetics, thermodynamic parameters and artificial neural network," *Waste Management*, vol. 85, pp. 131–140, 2019.

[28] J. Xing, K. Luo, H. Wang, Z. Gao, and J. Fan, "A comprehensive study on estimating higher heating value of biomass from proximate and ultimate analysis with machine learning approaches," *Energy*, vol. 188, p. 116077, 2019.

[29] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] X. Zhu, Y. Li, and X. Wang, "Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions," *Bioresource Technology*, vol. 288, p. 121527, 2019.

[31] J. Xing, K. Luo, H. Wang, and J. Fan, "Estimating biomass major chemical constituents from ultimate analysis using a random forest model," *Bioresource Technology*, vol. 288, p. 121541, 2019.

[32] J.-S. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[33] A. Dashti, M. Raji, A. Azarafza, A. Baghban, A. H. Mohammadi, and M. Asghari, "Rigorous prognostication and modeling of gas adsorption on activated carbon and zeolite-5A," *Journal of Environmental Management*, vol. 224, pp. 58–68, 2018.

[34] E. Akkaya, "ANFIS based prediction model for biomass heating value using proximate analysis components," *Fuel*, vol. 180, pp. 687–693, 2016.

[35] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, pp. 39–43, Nagoya, Japan, 1995, Ieee.

[36] A. Rostami, M. Arabloo, S. Esmaeilzadeh, and A. H. Mohammadi, "On modeling of bitumen/n-tetradecane mixture viscosity: application in solvent-assisted recovery method," *Asia-Pacific Journal of Chemical Engineering*, vol. 13, no. 1, article e2152, 2018.

[37] O. Castillo, *Type-2 Fuzzy Logic in Intelligent Control Applications*, vol. 272, Springer, 2012.

[38] J. E. Onwunalu and L. J. Durlofsky, "Application of a particle swarm optimization algorithm for determining optimum well location and type," *Computational Geosciences*, vol. 14, no. 1, pp. 183–198, 2010.

[39] A. Sharma and G. Onwubolu, "Hybrid particle swarm optimization and GMDH system," in *Hybrid self-organizing modeling systems*, pp. 193–231, Springer, 2009.

[40] M.-Y. Chen, "A hybrid ANFIS model for business failure prediction utilizing particle swarm optimization and subtractive clustering," *Information Sciences*, vol. 220, pp. 180–195, 2013.

[41] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*, pp. 69–73, Anchorage, AK, USA, 1998, IEEE.

[42] J. Holland, "Adaptation in Natural and Artificial Systems: An Introductory Analysis with Application to Biology," *Control and Artificial Intelligence*, 1975.

[43] R. Hassan, B. Cohanim, O. De Weck, and G. Venter, "A comparison of particle swarm optimization and the genetic algorithm," in *46th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference*, p. 1897, Austin, Texas, 2005.

[44] Q. Tang, Y. Chen, H. Yang et al., "Prediction of bio-oil yield and hydrogen contents based on machine learning method: effect of biomass compositions and pyrolysis conditions," *Energy & Fuels*, vol. 34, no. 9, pp. 11050–11060, 2020.

[45] A. Baghban, T. Kashiwao, M. Bahadori, Z. Ahmad, and A. Bahadori, "Estimation of natural gases water content using adaptive neuro-fuzzy inference system," *Petroleum Science and Technology*, vol. 34, no. 10, pp. 891–897, 2016.

[46] N. Kardani, A. Bardhan, S. Gupta et al., "Predicting permeability of tight carbonates using a hybrid machine learning approach of modified equilibrium optimizer and extreme learning machine," *Acta Geotechnica*, pp. 1–17, 2021.

[47] M. R. Kaloop, A. Bardhan, N. Kardani, P. Samui, J. W. Hu, and A. Ramzy, "Novel application of adaptive swarm intelligence techniques coupled with adaptive network-based fuzzy inference system in predicting photovoltaic power," *Renewable and Sustainable Energy Reviews*, vol. 148, p. 111315, 2021.

[48] N. Kardani, M. Hedayati Marzbali, K. Shah, and A. Zhou, "Machine learning prediction of the conversion of

lignocellulosic biomass during hydrothermal carbonization," *Biofuels*, pp. 1–13, 2021.

[49] N. Kardani, A. Zhou, S. L. Shen, and M. Nazem, "Estimating unconfined compressive strength of unsaturated cemented soils using alternative evolutionary approaches," *Transportation Geotechnics*, vol. 29, p. 100591, 2021.

[50] M. H. Ahmadi, A. Baghban, M. Sadeghzadeh et al., "Evaluation of electrical efficiency of photovoltaic thermal solar collector," *Engineering Applications of Computational Fluid Mechanics*, vol. 14, no. 1, pp. 545–565, 2020.

[51] A. Rostami, A. Baghban, A. H. Mohammadi, A. Hemmati-Sarapardeh, and S. Habibzadeh, "Rigorous prognostication of permeability of heterogeneous carbonate oil reservoirs: smart modeling and correlation development," *Fuel*, vol. 236, pp. 110–123, 2019.

Hindawi
BioMed Research International
Volume 2021, Article ID 9202127, 8 pages
https://doi.org/10.1155/2021/9202127

Research Article

On the Prediction of Biogas Production from Vegetables, Fruits, and Food Wastes by ANFIS- and LSSVM-Based Models

Yong Yang,[1,2] **Shuaishuai Zheng**,[1,2] **Zhilu Ai**,[1,2] **and Mohammad Mahdi Molla Jafari**[3]

[1]*College of Food Science and Technology, Henan Agricultural University, Zhengzhou, Henan 450002, China*
[2]*Key Laboratory of Staple Grain Processing, Ministry of Agriculture and Rural Affairs, Zhengzhou, Henan 450002, China*
[3]*Department of Petroleum Engineering, Ahwaz, Faculty of Petroleum Engineering, Petroleum University of Technology (PUT), Ahwaz, Iran*

Correspondence should be addressed to Zhilu Ai; zhilafood@163.com
and Mohammad Mahdi Molla Jafari; mohammad.molajafari@afp.put.ac.ir

Received 5 August 2021; Revised 17 August 2021; Accepted 21 August 2021; Published 24 September 2021

Academic Editor: Alireza Baghban

This study is aimed at modeling biodigestion systems as a function of the most influencing parameters to generate two robust algorithms on the basis of the machine learning algorithms, including adaptive network-based fuzzy inference system (ANFIS) and least square support vector machine (LSSVM). The models are assessed utilizing multiple statistical analyses for the actual values and model outcomes. Results from the suggested models indicate their great capability of predicting biogas production from vegetable food, fruits, and wastes for a variety of ranges of input parameters. The values that are calculated for the mean relative error (MRE %) and mean squared error (MSE) were 29.318 and 0.0039 for ANFIS, and 2.951 and 0.0001 for LSSVM which shows that the latter model has a better ability to predict the target data. Finally, in order to have additional certainty, two analyses of outlier identification and sensitivity were performed on the input parameter data that proved the proposed model in this paper has higher reliability in assessing output values compared with the previous model.

1. Introduction

The main disposal pathways for FW (food waste) residues are treating via incineration or disposal in landfills. Given the very fast biodegradability of the food wastes in the presence of contaminating microorganisms, their disposal in landfills is very problematic [1, 2]. In addition, biodegradation within landfills necessitates a vast area, and greenhouse gases, e.g., methane, are generated with no profit gained via the energy produced through the biomass [3]. Thus, many nations have prohibited such a disposal option. From another viewpoint, given the high moisture content (>70%) of the organic matter, the incineration requires intensive amounts of energy with no energy recovery in some situations [4, 5]. Both options impose adverse impacts on human health and the environment [6, 7].

Therefore, since waste-to-energy techniques support reduced environmental impacts and partial replacement of fossil reserves, they are studied for disposal of organic wastes. One of the feasible approaches is AD (anaerobic digestion), which is an environmentally friendly technology for transforming liquid or solid organic wastes into biogas, which is convertible to beneficial energies (heat and/or electricity) [8, 9]. Anaerobic digestion is a complicated multistep biochemical degradation procedure conducted in the absence of O_2, through which the microorganisms transform organic compounds into a gaseous mix mainly composed of CO_2 (carbon dioxide), N_2 (nitrogen), and CH_4 (methane). Nonetheless, one can find other compounds in the composition, including H_2 (hydrogen), H_2S (hydrogen sulfide), O_2 (oxygen), CO (carbon monoxide), and NH_3 (ammonia). Also, trace amounts of siloxanes, dust particles, and halogenated and aromatic compounds are found in biogas; some of which can increase emissions, corrosion, and biohazards for human health. Biogas is also saturated

with water. Also, this process generates sludge residues (or digests), which can be utilized as a fuel for energy generation subsequent to a drying treatment or directly for remediation of soil [10].

The conversion yield, the biogas composition, and the production rate are affected by the biomass nature, the configuration of biodigester, and the process characteristics [11, 12]. Given the diversity of handling and processing techniques, resources, local seasons and climates, and eating behaviors, the same kind of FWs may provide extremely variable features [13]. The amount of TSs (total solids) found in FW ranges from <2%w in liquid FWs to >90%w in solid FWs. The organic content (often ~90%) shown by the VS/TS ratio (in which the VS (volatile solids) represents the weight fraction convertible into gaseous materials) makes biomass a suitable candidate for anaerobic digestion. The C/N ratio varies between 3 and 55, paving the way for modifying the mixture C/N ratio to reach the optimum biodegradability for the food biomass. For most of the FWs, the acidity necessitates adding chemicals (e.g., alkali reagents) to stabilize biodigesters' pH. For each kind of FWs, the highest potential of methane production is within the 0.31–1.1 m^3 CH_4/kg range when volatile solids are added [14]. Thus, the electricity generated per 1 ton of fresh substances is within the 151.6-224.6 m^3/t range for FVW and FW, which is nearly the same value acquired from chicken and cattle dungs (257.3 and 122.5 m^3/t, respectively) [15].

As observed in [16–18] and others, the parameters contributing to biogas generation have been investigated. Nonetheless, only a few studies have simultaneously investigated more than a single factor. The simultaneous investigation of the interacting impact of a number of experimental scenarios presents invaluable data for optimization and prediction of the key features used in the experimental procedure of the entire studied scenarios. The problem is finding plausible standard experimental techniques for dataset compilation. Concerning optimization and prediction, the most routine methods pave the way for creating polynomial models correlating the response to the procedure irrespective of variables and their associated interactions [19]. Seman et al. concentrated on developing an association between parameters and evaluating the interactions between factors [20]. Therefore, the parameter optimization and response prediction of the process were founded on the basis of polynomial regression modeling (second-order model).

Another modeling instrument employed for better prediction is ANNs (artificial neural networks) [21]. The independent variables employed by Beltramo et al. in the artificial neural network model to assess the rate of biogas generation were TS, VFA, VS, acid detergent lignin, acid detergent fiber, ammonium nitrogen, neutral detergent fiber, OLR, and HRT [22]. The prediction error of the model was 6.24%, and the authors considered a coefficient of determination, $R^2 = 0.9$, as the optimum result. By using an artificial neural network, Ghatak et al. optimized and modeled the prediction of particular biogas generation via the parameters including temperature, duration, and composition [23]. The neural model could anticipate the creation of biogas with an accuracy of 99.7%.

In this paper, we have tried to predict the values of biogas production from vegetables, fruits, and food wastes using two new models, ANFIS and LSSVM. First, a wide range of actual output data and input parameters affecting them were collected. Then, these two models were constructed and statistically evaluated, and compared. Finally, the results of these models were compared with the previously proposed models (in terms of accuracy), and the best model was proposed.

2. Description of Models

2.1. ANFIS. The adaptive network-based fuzzy inference system (ANFIS) algorithm is defined as a class of neural network techniques to address problems involving function approximation [24]. To put it in another way, an ANFIS structure is a combined information acquired from the fuzzy logic system and artificial neural network, and it consists of several membership function (MF) parameters optimized utilizing optimization algorithms [25]. Accordingly, the ANFIS structure, because of being particular, is significantly precise, and its reliance on real values is less than other machine learning algorithms, for instance, the artificial neural networks [25].

A typical ANFIS structure includes five layers, each of which has a number of nodes defined by their node functions [26]. Layers' association can be established using internal connections. The outputs of the previous layer are used as the inputs of the next layer. It is worth noting that the fuzzy inference system is utilized in the ANFIS technique as a fuzzy system. More specifically, for the inputs with two parameters, x and y, and output with a single parameter, f_i, the rules governing an ANFIS structure are expressed as follows [27].

First rule: if x is $M1$ and y is $N1$ then z is $f1(x, y)$

First rule: if x is $M2$ and y is $N2$ then z is $f2(x, y)$

where fuzzy sets indicated by M and N and f_i (x, y) is representative of the first-order fuzzy inference system output.

The adaptive nodes included in the first layer are specified as follows:

$$O_i^1 - \mu_{M_i}(x), \quad \text{for } i = 1, 2, \tag{1}$$

$O_i^1 - \mu_{M_{i-2}}(x)$, for $i = 3, 4$,where $\mu(y)$ and $\mu(x)$ indicate the membership functions.

Each node denoted by π is constant in the following layer.

$$O_i^2 = W_i = \mu_{M_i}(x)\mu_{N_i}(y), \quad \text{for } i = 1, 2, \tag{2}$$

where W_i indicates the firing strength of the rule.

The third layer has constant nodes denoted by N. The corresponding node functions are applied to normalize the firing by dividing the ith node's firing strength value by the all firing strength values' summation [28].

$$O_i^3 = \frac{W_i}{\sum W_i}, \quad \text{for } i = 1, 2. \tag{3}$$

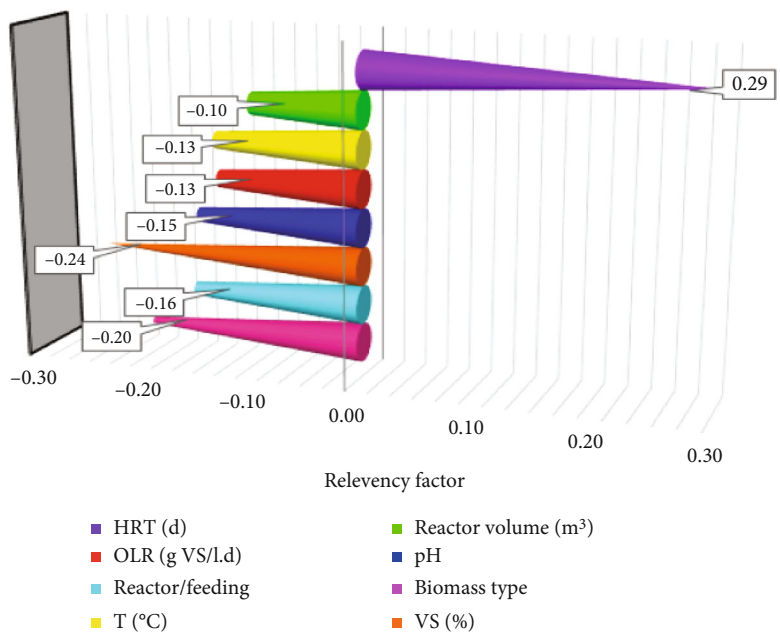


Figure 1: Sensitivity on various input parameters.

The fourth layer includes adaptive nodes indicated by the square shapes.

$$O_i^4 = \overline{W_i} f_i = \overline{W_i}(m_i X_1 + n_i X_2 + r_i), \quad \text{for } i = 1, 2, \tag{4}$$

where f_1 and f_2 indicate the fuzzy if-then rules defined as follows [29].

Rule 1: if x is M_1 and y is N_1 then $f_1 = p_1 x + q_1 y + r_1$

Rule 1: if x is M_2 and y is N_2 then $f_2 = p_2 x + q_2 y + r_2$

where p_i, q_i, and r_i indicate the consequential terms.

The overall output in the last layer is given by:

$$O_i^5 = Y = \sum_i \overline{W_i} f_i = \overline{W_1} f_1 + \overline{W_2} f_2 = \frac{\sum W_i f_i}{\sum W_i}. \tag{5}$$

Totally, the output is described as a linear combined consequential term [30].

2.2. LSSVM. The supervised least square support vector machine (LSSVM) algorithm developed in 1999 by Suykens et al. for solving problems stemmed from the regression together with function approximation. For the inputs denoted by X_i and the output denoted by Y_i, the usual LSSVM nonlinear function is given as follows [31].

$$f(x) = \omega^T \phi(x) + b, \tag{6}$$

where f indicates the connections between the target output and inputs, w denotes the m-dimensional weight vector, and b denotes the bias. The following equation is commonly used to solve the regression problems concerning the minimization theory [32]:

$$\min J(\omega, e) = \frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{k=1}^{N} e_k^2. \tag{7}$$

The following boundary conditions need to be considered:

$$y_k = \omega^T \phi(x_k) + b + e_k \; k = 1, 2, \cdots, N, \tag{8}$$

where c indicates the margin parameter and e_k indicates the error variable of x_k. The LSSVM straightforward derivations lead to

$$f(x) = \sum_{k=1}^{N} a_k K(x, x_k) + b. \tag{9}$$

The radial basis function is commonly used as a kernel function in regression faults due to its great efficiency, which is given by [33]

$$K(x, x_k) = e^{\left(-\left(\|x - x_k\|^2 / \sigma^2\right)\right)}. \tag{10}$$

The σ^2 in this equation indicates the squared bandwidth that needs to be estimated using optimization.

3. Materials and Methods

3.1. Sensitivity Analysis. In order to analyze the effects of individual inputs on the output value, a sensitivity analysis was carried out. Thus, the relevancy factor was decided as

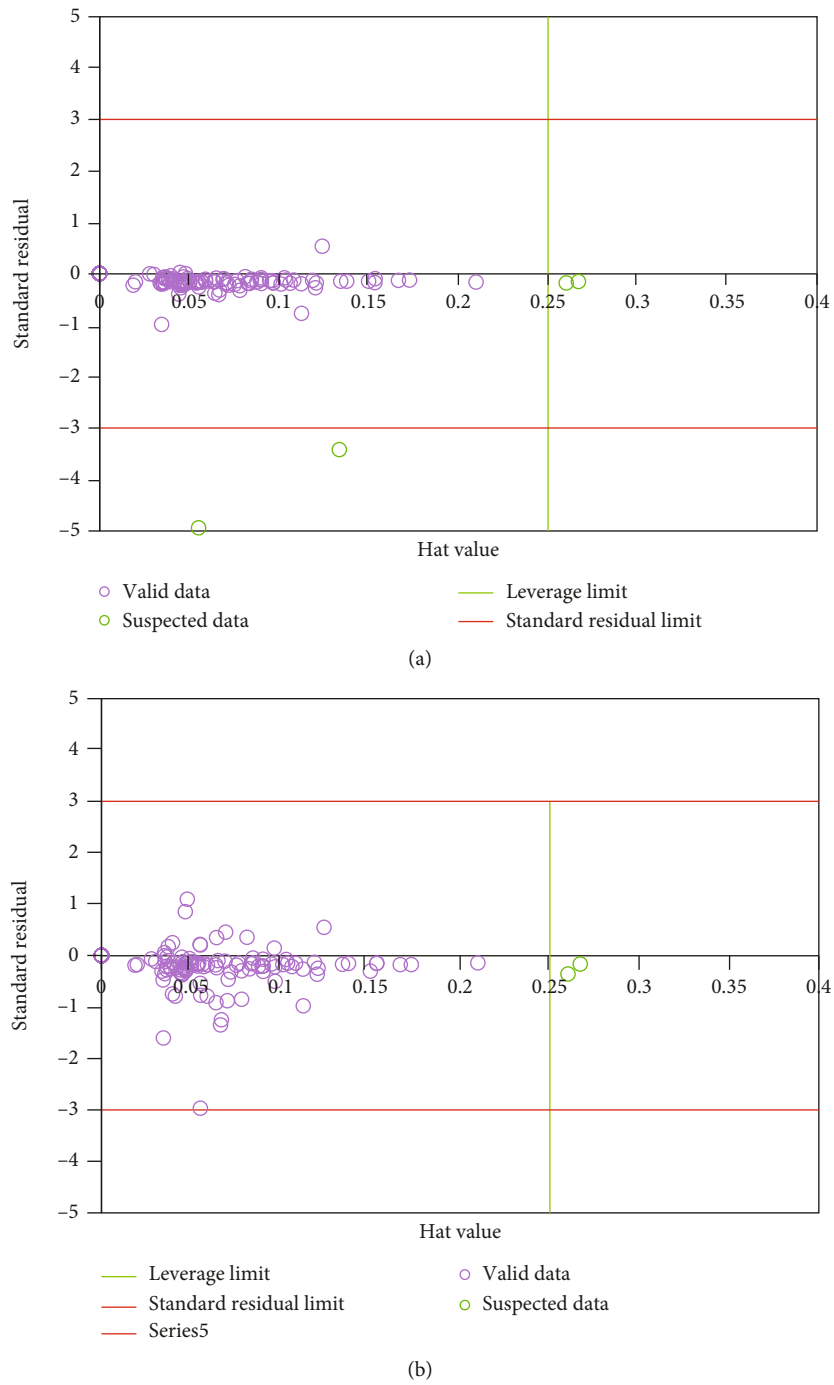


Figure 2: Analysis to identify outlier's data in models (a) LSSVM and (b) ANFIS.

presented in the following to discover the effect of the individual inputs [34].

$$r = \frac{\sum_{i=1}^{n}\left(X_{k,i-}\overline{X_k}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_{k,i} - \overline{X_k}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}. \quad (11)$$

In this equation, $\bar{X}_k$ and Y stand for the average of the input, and the average of the k_{th} output, Y_i represents the i_{th} output, N represents the entire number of data points, and $X_{k,i}$ stands for the i_{th} input value of the k_{th} parameter. Also, the r values range between -1 and 1. The less absolute value is interpreted as the fact that the input is less effective on the output parameter. In addition, the positivity or negativity of r is regarded as direct or reverse impacts of the concerned inputs; i.e., by increasing an input with negative r values, the target parameter is decreased; however, for the inputs with negative values of r, it is increased.

This research examined eight inputs that reflected a direct impact on the discussed target. Figure 1 presents the

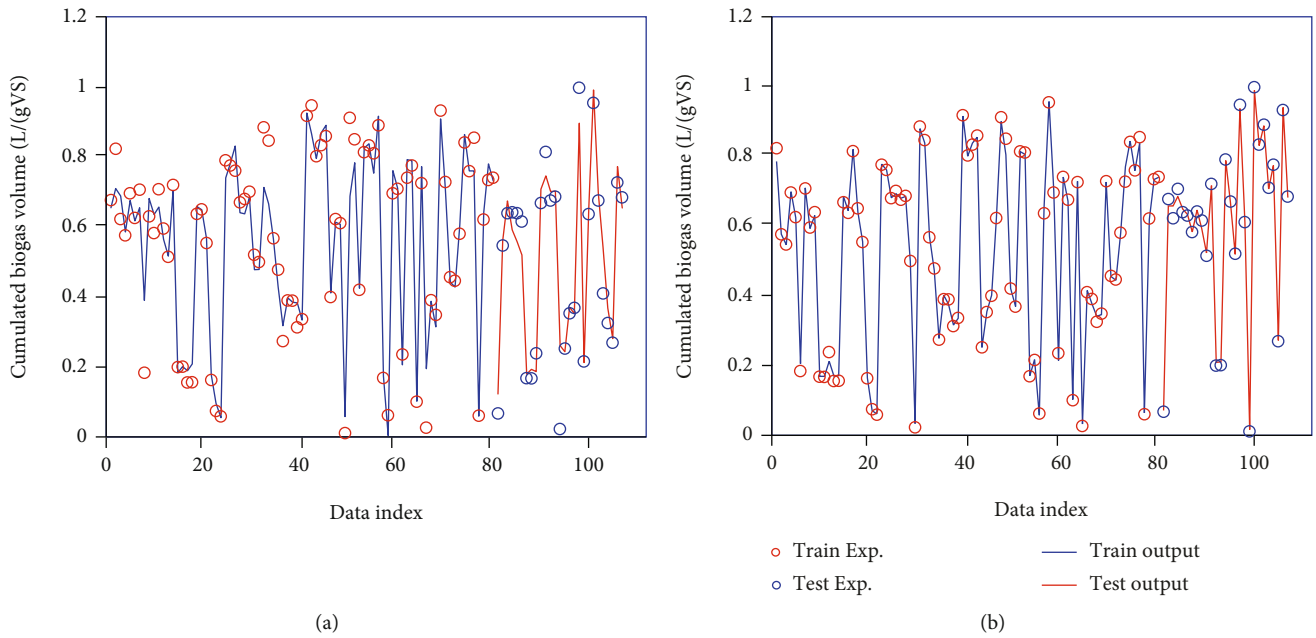


Figure 3: Simultaneous viewing of real and simulated output data using models (a) LSSVM and (b) ANFIS.

analysis results, where the highest eigenvalue n is of the positive value of $r = 0.29$ that refers to HRT (d).

3.2. Data Gathering and Preanalysis Phase. In this phase, our study employed two different techniques to evaluate and predict the output parameters from the employed algorithms. Then, the data acquired in the experiments conducted in this research were employed for training the above-mentioned algorithms [35], whilst of the whole data points, about 25% of the same data points were used for the validation of those algorithms. Also, the dataset was subjected to the normalization procedure:

$$D_k = 2\frac{x - x_{\min}}{x_{\max} - x_{\min}} - 1. \quad (12)$$

In the above equation, D_k represents the normalized value and x stands for the input value.

3.3. Identification of Outliers. Outlier or suspected data points featuring behaviors that differ from the major part of the databank show up in a large dataset typically. However, the same data points may affect model's accuracy and reliability. Hence, it seems necessary then to try to find such data in the proposed models, in particular for the training datasets. In case of neglecting some unrecognized impacts, some restrictions may be encountered in the model. In the other words, the analysis of outliers may provide us with an insight on the same restrictions, which are the benefits of the discussed analysis. In order to eliminate the outlier data, the leverage technique was used, which requires determining the deviation of the predictive tool from the concerned real data [34, 36]. The deviation which is also termed as standardized crossvalidated residuals creates a Hat matrix, which can be determined on the basis of the equation below in this study:

$$H = X\left(X^t X\right)^{-1} X^t, \quad (13)$$

where X stands for an $N \times P$ matrix. N and P, respectively, represent the entire number of data points and the input parameters. T and -1 are called transpose and inverse operators, respectively. In addition, the equation below was used to explicate a warning leverage value:

$$H^* = \frac{3n}{(p+1)}. \quad (14)$$

The practical region is delineated within $0 \le H \le H*$ and $R = \pm 3$ rectangular area. According to the red points observed in Figure 2, only a number of 20 suspected data were discovered amongst the entire dataset.

4. Results and Discussion

The two computational techniques developed in this work are ANFIS and LSSVM used to estimate the target values. Upon splitting the dataset into the testing and training datasets, of the whole data points, 75% were employed to make use of the above-mentioned model for determining the outputs. Then, the training process performance is expressible through a comparison made between the real values and the predicted ones in this step. Alternatively, the comparison made in the testing phase presents a better idea about the model's accuracy in unclear circumstances, which is called model generalization. Figure 3 presents the simultaneous comparison of the experimental and determined targets for the whole models trained in testing and training databanks. Also, as Figure 3

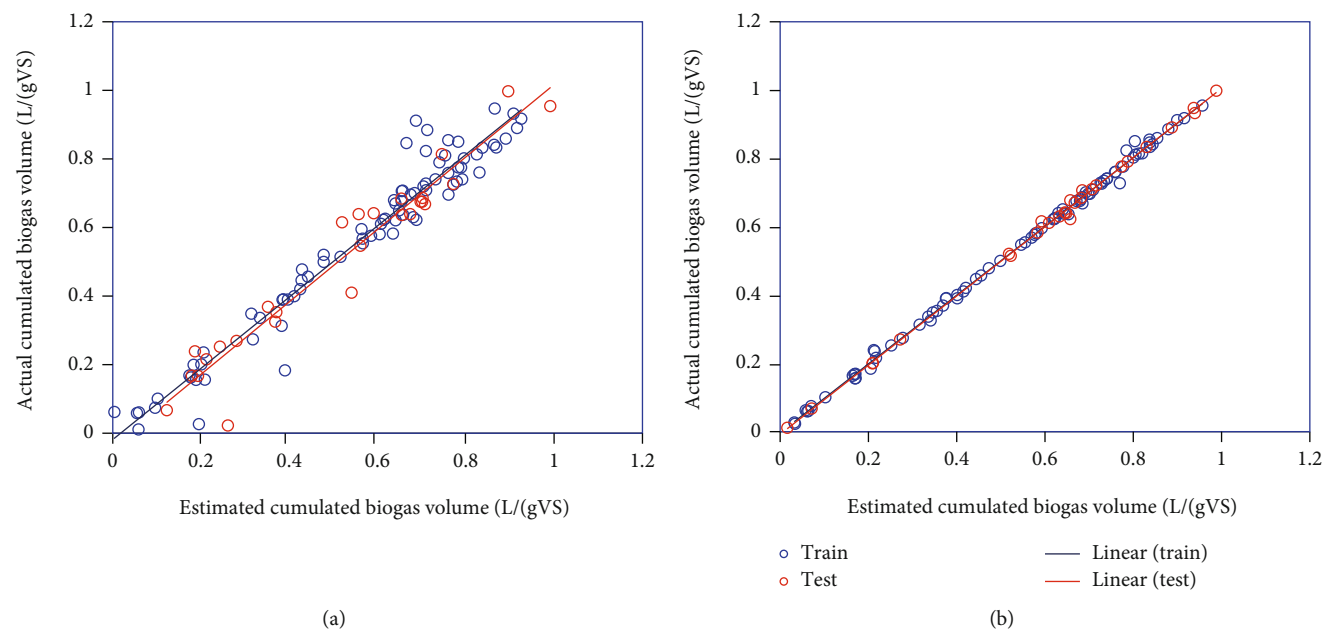


Figure 4: Linear regression diagrams to determine the accuracy of models (a) LSSVM and (b) ANFIS.

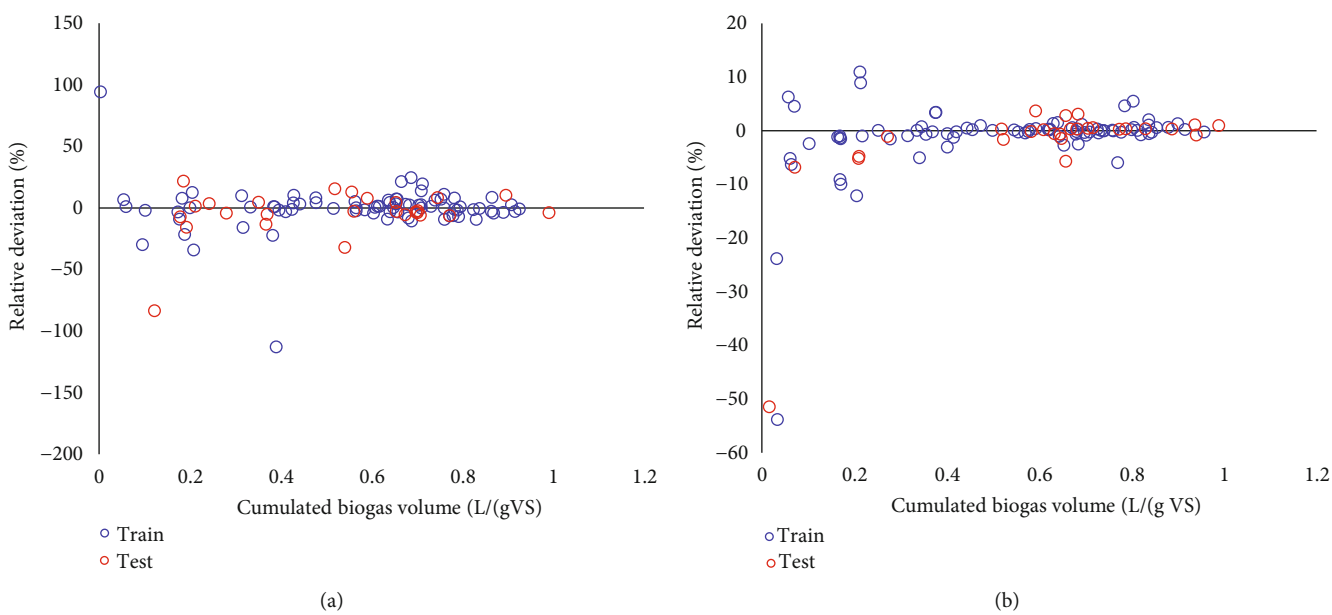


Figure 5: The deviation plots for the (a) LSSVM and (b) ANFIS models.

suggests, the calculated constants may include the experimental points featuring plausible performances.

Thus, it is evident that the proposed LSSVM model can estimate outputs with higher accuracy. In order to evaluate the usability of the proposed models, various mathematical and graphical techniques were used. In Figure 4, the cross-plots or regression are presented exhibiting the capability of the suggested algorithms in estimating the target values. It is clear that the data points are highly concentrated around the bisector line.

Figure 5 presents the deviation plots for the ANFIS and LSSVM models and shows the outputs vs. relative deviation for testing and training steps. Most of the determined deviations are found in the vicinity of the zero error line. In addition, the deviations compaction in the LSSVM model is clearer than the other model. The same compaction reflects the accuracy of prediction for this model.

Table 1 reports the mathematical indexes determined for the presented models. The higher values of R^2, and also, the lower values of RMSE, MRE, STD, and MSE are observable for the proposed models, reflecting their good capability in estimating the output values.

Also, the models of ANFIS and LSSVM and the rest of the techniques found in the literature to decide the target

Table 1: Results of various statistical analyzes to determine the accuracy of the two models ANFIS and LSSVM in predicting real values.

Model	Phase	R^2	MRE (%)	MSE	RMSE	STD
LSSVM	Train	0.998	2.762	0.0001	0.0113	0.0091
	Test	0.998	3.521	0.0001	0.0111	0.0082
	Total	0.998	2.951	0.0001	0.0111	0.0089
ANFIS	Train	0.949	22.070	0.0036	0.0598	0.0464
	Test	0.936	51.064	0.0047	0.0683	0.0494
	Total	0.946	29.318	0.0039	0.0683	0.0471

values, e.g., those presented by the authors such as Neto and colleagues were compared. In 2021, they used the artificial neural network method to predict this parameter [35]. Compared to other models, the LSSVM model with $R^2 = 0.998$ features the most optimal performance. The same comparison reveals that the minimum accuracy is attributable to the ANN model with $R^2 = 0.6167$.

5. Conclusions

In this study, two accurate techniques, i.e., ANFIS and LSSVM, were presented successfully to estimate biogas production. The developed instruments used for estimation may help the scholars in suggesting a new efficient measurement technique. According to the statistical analyses, the LSSVM model can lead to the most accurate results with the best values of STD, RMSE, R^2, MSE, and MRE. Given the above results, compared to the rest of the computational techniques, the LSSVM model presented a superb performance in terms of validity, accuracy, and generalization. Additionally, a sensitivity analysis was conducted in order to reflect the effect of input parameters on the target values which showed that HRT (d) has the greatest effect on the output parameter.

Data Availability

The data used to support the findings of this study are provided within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Key R&D Program of China (project number: 2018YFD0400605).

References

[1] P. Viswanath, S. S. Devi, and K. Nand, "Anaerobic digestion of fruit and vegetable processing wastes for biogas production," *Bioresource Technology*, vol. 40, no. 1, pp. 43–48, 1992.

[2] O. Kuczman, M. V. D. Gueri, S. N. M. de Souza et al., "Food waste anaerobic digestion of a popular restaurant in Southern Brazil," *Journal of Cleaner Production*, vol. 196, pp. 382–389, 2018.

[3] S. J. Grimberg, D. Hilderbrandt, M. Kinnunen, and S. Rogers, "Anaerobic digestion of food waste through the operation of a mesophilic two- phase pilot scale digester - assessment of variable loadings on system performance," *Bioresource Technology*, vol. 178, pp. 226–229, 2015.

[4] M. M. V. Leme, M. H. Rocha, E. E. S. Lora, O. J. Venturini, B. M. Lopes, and C. H. Ferreira, "Techno-economic analysis and environmental impact assessment of energy recovery from municipal solid waste (MSW) in Brazil," *Resources, Conservation and Recycling*, vol. 87, pp. 8–20, 2014.

[5] M. L. N. Carneiro and M. S. P. Gomes, "Energy, exergy, environmental and economic analysis of hybrid waste-to-energy plants," *Energy Conversion and Management*, vol. 179, pp. 397–417, 2019.

[6] D. Palaniswamy, "Optimising biogas from food waste using a neural network model," in *in Proceedings of the Institution of Civil Engineers-Municipal Engineer*, Thomas Telford Ltd, 2017.

[7] C. Zhang, H. Su, J. Baeyens, and T. Tan, "Reviewing the anaerobic digestion of food waste for biogas production," *Renewable and Sustainable Energy Reviews*, vol. 38, pp. 383–392, 2014.

[8] C. Mao, Y. Feng, X. Wang, and G. Ren, "Review on research achievements of biogas from anaerobic digestion," *Renewable and Sustainable Energy Reviews*, vol. 45, pp. 540–555, 2015.

[9] J. Zhang, K. C. Loh, W. Li, J. W. Lim, Y. Dai, and Y. W. Tong, "Three-stage anaerobic digester for food waste," *Applied Energy*, vol. 194, pp. 287–295, 2017.

[10] E. A. Scano, C. Asquer, A. Pistis, L. Ortu, V. Demontis, and D. Cocco, "Biogas from anaerobic digestion of fruit and vegetable wastes: experimental results on pilot-scale and preliminary performance evaluation of a full-scale power plant," *Energy Conversion and Management*, vol. 77, pp. 22–30, 2014.

[11] Q. Sun, H. Li, J. Yan, L. Liu, Z. Yu, and X. Yu, "Selection of appropriate biogas upgrading technology-a review of biogas cleaning, upgrading and utilisation," *Renewable and Sustainable Energy Reviews*, vol. 51, pp. 521–532, 2015.

[12] I. U. Khan, M. H. D. Othman, H. Hashim et al., "Biogas as a renewable energy fuel - a review of biogas upgrading, utilisation and storage," *Energy Conversion and Management*, vol. 150, pp. 277–294, 2017.

[13] Y. Meng, S. Li, H. Yuan et al., "Evaluating biomethane production from anaerobic mono- and co-digestion of food waste and floatable oil (FO) skimmed from food waste," *Bioresource Technology*, vol. 185, pp. 7–13, 2015.

[14] F. Xu, Y. Li, X. Ge, L. Yang, and Y. Li, "Anaerobic digestion of food waste - challenges and opportunities," *Bioresource Technology*, vol. 247, pp. 1047–1058, 2018.

[15] S. Achinas, V. Achinas, and G. J. W. Euverink, "A technological overview of biogas production from biowaste," *Engineering*, vol. 3, no. 3, pp. 299–307, 2017.

[16] V. V. Nair, H. Dhar, S. Kumar, A. K. Thalla, S. Mukherjee, and J. W. C. Wong, "Artificial neural network based modeling to evaluate methane yield from biogas in a laboratory-scale anaerobic bioreactor," *Bioresource Technology*, vol. 217, pp. 90–99, 2016.

[17] A. Sharma, N. A. Ansari, A. Pal, Y. Singh, and S. Lalhriatpuia, "Effect of biogas on the performance and emissions of diesel engine fuelled with biodiesel-ethanol blends through response

surface methodology approach," *Renewable Energy*, vol. 141, pp. 657–668, 2019.

[18] M. Safari, R. Abdi, M. Adl, and J. Kafashan, "Optimization of biogas productivity in lab-scale by response surface methodology," *Renewable Energy*, vol. 118, pp. 368–375, 2018.

[19] P. Tsapekos, B. Khoshnevisan, M. Alvarado-Morales, A. Symeonidis, P. G. Kougias, and I. Angelidaki, "Environmental impacts of biogas production from grass: role of co-digestion and pretreatment at harvesting time," *Applied Energy*, vol. 252, p. 113467, 2019.

[20] S. Z. A. Seman, I. Idris, A. Abdullah, I. K. Shamsudin, and M. R. Othman, "Optimizing purity and recovery of biogas methane enrichment process in a closed landfill," *Renewable Energy*, vol. 131, pp. 1117–1127, 2019.

[21] P. Sakiewicz, K. Piotrowski, J. Ober, and J. Karwot, "Innovative artificial neural network approach for integrated biogas - wastewater treatment system modelling: effect of plant operating parameters on process intensification," *Renewable and Sustainable Energy Reviews*, vol. 124, p. 109784, 2020.

[22] T. Beltramo, M. Klocke, and B. Hitzmann, "Prediction of the biogas production using GA and ACO input features selection method for ANN model," *Information Processing in Agriculture*, vol. 6, no. 3, pp. 349–356, 2019.

[23] M. D. Ghatak and A. Ghatak, "Artificial neural network model to predict behavior of biogas production curve from mixed lignocellulosic co-substrates," *Fuel*, vol. 232, pp. 178–189, 2018.

[24] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [book review]," *IEEE Transactions on Automatic Control*, vol. 42, no. 10, pp. 1482–1484, 1997.

[25] A. Baghban, J. Sasanipour, P. Haratipour, M. Alizad, and M. Vafaee Ayouri, "ANFIS modeling of rhamnolipid breakthrough curves on activated carbon," *Chemical Engineering Research and Design*, vol. 126, pp. 67–75, 2017.

[26] M. Mir, M. Kamyab, M. J. Lariche, A. Bemani, and A. Baghban, "Applying ANFIS-PSO algorithm as a novel accurate approach for prediction of gas density," *Petroleum Science and Technology*, vol. 36, no. 12, pp. 820–826, 2018.

[27] A. Baghban, A. Jalali, M. Shafiee, M. H. Ahmadi, and K. W. Chau, "Developing an ANFIS-based swarm concept model for estimating the relative viscosity of nanofluids," *Engineering Applications of Computational Fluid Mechanics*, vol. 13, no. 1, pp. 26–39, 2019.

[28] S. Shamshirband, M. Hadipoor, A. Baghban, A. Mosavi, J. Bukor, and A. Várkonyi-Kóczy, "Developing an ANFIS-PSO model to predict mercury emissions in combustion flue gases," *Mathematics*, vol. 7, no. 10, p. 965, 2019.

[29] M. S. Zaghloul, R. A. Hamza, O. T. Iorhemen, and J. H. Tay, "Comparison of adaptive neuro-fuzzy inference systems (ANFIS) and support vector regression (SVR) for data-driven modelling of aerobic granular sludge reactors," *Journal of Environmental Chemical Engineering*, vol. 8, no. 3, p. 103742, 2020.

[30] Z. X. Li, F. L. Renault, A. O. C. Gómez et al., "Nanofluids as secondary fluid in the refrigeration system: experimental data, regression, ANFIS, and NN modeling," *International Journal of Heat and Mass Transfer*, vol. 144, p. 118635, 2019.

[31] J. A. Suykens, *Least squares support vector machines*, World scientific, 2002.

[32] J. Ye and T. Xiong, *SVM versus least squares SVM*, In Artificial Intelligence and Statistics, PMLR, 2007.

[33] X. Zhang and Z. Ge, "Local parameter optimization of LSSVM for industrial soft sensing with big data and cloud implementation," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 2917–2928, 2019.

[34] M. H. Ahmadi, A. Baghban, M. Sadeghzadeh et al., "Evaluation of electrical efficiency of photovoltaic thermal solar collector," *Engineering Applications of Computational Fluid Mechanics*, vol. 14, no. 1, pp. 545–565, 2020.

[35] J. G. Neto, L. V. Ozorio, T. C. C. de Abreu, B. F. dos Santos, and F. Pradelle, "Modeling of biogas production from food, fruits and vegetables wastes using artificial neural network (ANN)," *Fuel*, vol. 285, p. 119081, 2021.

[36] A. Bemani, A. Baghban, and A. H. Mohammadi, "An insight into the modeling of sulfur content of sour gases in supercritical region," *Journal of Petroleum Science and Engineering*, vol. 184, p. 106459, 2020.

Hindawi
BioMed Research International
Volume 2021, Article ID 6480804, 20 pages
https://doi.org/10.1155/2021/6480804

Research Article

3D-QSAR-Based Pharmacophore Modeling, Virtual Screening, and Molecular Docking Studies for Identification of Tubulin Inhibitors with Potential Anticancer Activity

Salimeh Mirzaei,[1] **Razieh Ghodsi**,[2,3] **Farzin Hadizadeh**,[2,3] **and Amirhossein Sahebkar**[3,4,5]

[1]*Department of Medicinal Chemistry, Faculty of Pharmacy, Hormozgan University of Medical Sciences, Bandar Abbas, Iran*
[2]*Department of Medicinal Chemistry, School of Pharmacy, Mashhad University of Medical Sciences, Mashhad, Iran*
[3]*Biotechnology Research Center, Pharmaceutical Technology Institute, Mashhad University of Medical Sciences, Mashhad, Iran*
[4]*Applied Biomedical Research Center, Mashhad University of Medical Sciences, Mashhad, Iran*
[5]*School of Pharmacy, Mashhad University of Medical Sciences, Mashhad, Iran*

Correspondence should be addressed to Salimeh Mirzaei; mirzaeis@hums.ac.ir and Farzin Hadizadeh; hadizadehf@mums.ac.ir

Salimeh Mirzaei and Farzin Hadizadeh contributed equally to this work.

Received 11 June 2021; Accepted 22 July 2021; Published 25 August 2021

Academic Editor: Alireza Baghban

In this study, we aimed to develop a pharmacophore-based three-dimensional quantitative structure activity relationship (3D-QSAR) for a set including sixty-two cytotoxic quinolines (1-62) as anticancer agents with tubulin inhibitory activity. A total of 279 pharmacophore hypotheses were generated based on the survival score to build QSAR models. A six-point pharmacophore model (AAARRR.1061) was identified as the best model which consisted of three hydrogen bond acceptors (A) and three aromatic ring (R) features. The model showed a high correlation coefficient ($R^2 = 0.865$), cross-validation coefficient ($Q^2 = 0.718$), and F value (72.3). The best pharmacophore model was then validated by the Y-Randomization test and ROC-AUC analysis. The generated 3D contour maps were used to reveal the structure activity relationship of the compounds. The IBScreen database was screened against AAARRR.1061, and after calculating ADMET properties, 10 compounds were selected for further docking study. Molecular docking analysis showed that compound STOCK2S-23597 with the highest docking score (-10.948 kcal/mol) had hydrophobic interactions and can form four hydrogen bonds with active site residues.

1. Introduction

Cancer is a disease identified by the uncontrolled proliferation of cells. As cancer is the second deadliest disease worldwide, discovering new methods and drugs to treat this disease is very important [1, 2]. Nowadays, scientists are looking to find many novel targets for developing anticancer drugs, due to the critical need for new anticancer agents. Microtubules are polymers composed of α- and β-tubulin heterodimers [3]. They control several cellular functions including motility regulation, cell signaling, secretion, and cell architecture in interphase [4, 5]. Thereby, tubulin and microtubules are important targets for antitumor therapy [6]. The α- and β-tubulin heterodimers are in dynamic equilibrium with microtubules. The impairment of the dynamic equilibrium of microtubules leads to mitotic arrest and accordingly apoptosis [7, 8]. Therefore, microtubule targeting agents that interfere with dynamic microtubules can be effective in the treatment of cancer. In general, the binding sites for paclitaxel, vinblastine, and colchicine are well determined in tubulin [9–11].

Agents that bind to the vinca alkaloid binding site (e.g., Vincristine) or the colchicine binding site (e.g., colchicine and podophyllotoxin) are determined as microtubule destabilizing agents or inhibitors of tubulin assembly. In contrast, agents that bind to the paclitaxel binding site (e.g., paclitaxel) are tubulin promotors or microtubule stabilizing agents [12, 13].

Table 1: The score of different parameters of the hypotheses.

S. No.	Hypothesis	Survival score	Survival inactive	Site	Vector	Matches	Activity	Inactive
1	AAARRR.1061	3.870	1.270	0.98	0.992	18	5.640	2.599
2	AAAHRR.319	3.863	1.258	0.98	0.991	18	5.895	2.605
3	AAAHRR.311	3.863	1.258	0.98	0.999	18	5.895	2.605
4	AAHRRR.1415	3.862	1.280	0.97	0.999	18	5.640	2.582
5	AAAHRR.327	3.861	1.280	0.97	0.999	18	5.640	2.581
6	AAAHRR.326	3.861	1.280	0.97	0.999	18	5.513	2.851

A: acceptor; H: hydrophobic; R: aromatic ring.

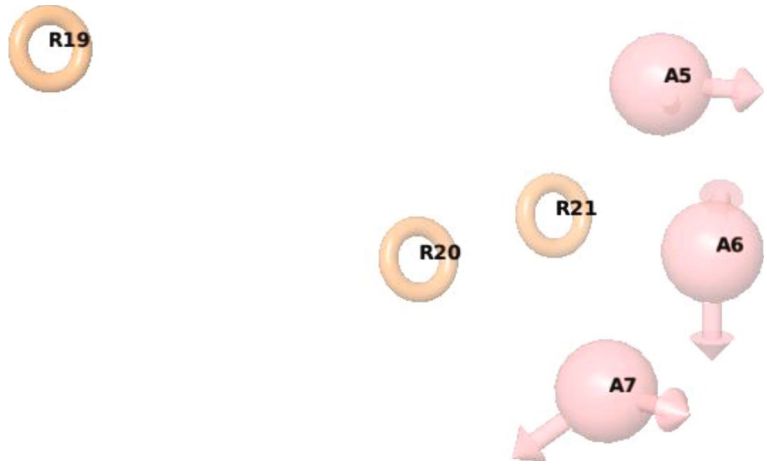


Figure 1: Common pharmacophore hypothesis AAARRR.1061. Pink spheres with vectors A5, AR, and A7 are hydrogen bond acceptor features, and orange open circles R19, R20, and R21 are aromatic ring features.

Antimitotic agents such as vinca alkaloids and taxanes have been used for the clinical treatment of different cancerous patients over the past decades [14, 15]. However, the use of these agents is limited due to drug resistance and associated side effects [16, 17]. Recently, different small molecules have been designed as tubulin inhibitors and anticancer agents [5, 8, 12, 18, 19].

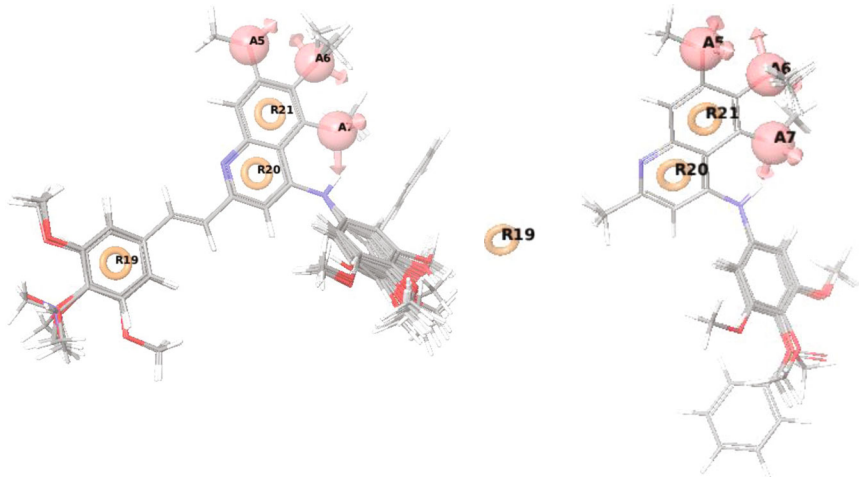
Despite the design and synthesis of promising compounds, antiproliferative and tubulin inhibitory activities of these compounds could not be confirmed until experimentally evaluated, which are time-consuming and expensive besides ethical limitations in using and sacrificing animals. Therefore, computational methods (*in silico* tools) such as pharmacophore modeling, drug screening, and design are essential for activity prediction and greatly reduce the time and cost of drug development [20, 21].

One advantage of *in silico* methodologies is that they can be used to identify new compounds with desirable properties as "druggable" targets before they are synthesized and thus reduce the need for time-consuming and expensive animal and in vitro laboratory work [21, 22].

The relationships between the biological activity and physicochemical properties of a set of compounds could be analyzed with three-dimensional quantitative structure activity relationships (3D-QSARs). QSAR techniques by generating three-dimensional alignment of molecules facilitate design and synthesis of new compounds with superior activity [23]. In this paper, quinolines with cytotoxic activities were used to develop new potent anticancer agents. We selected some cytotoxic quinolines (1-62) as tubulin inhibitors from our previous works [5, 8], generated a training set and test set, and created 279 pharmacophore models with

Table 2: Intersite distances between the pharmacophoric sites of AAARRR.1061.

Entry	Site1	Site2	Distance (Å)
AAARRR.1061	A5	A6	2.702
AAARRR.1061	A5	A7	4.830
AAARRR.1061	A5	R19	10.134
AAARRR.1061	A5	R20	5.045
AAARRR.1061	A5	R21	2.803
AAARRR.1061	A6	A7	2.832
AAARRR.1061	A6	R19	11.316
AAARRR.1061	A6	R20	5.053
AAARRR.1061	A6	R21	2.778
AAARRR.1061	A7	R19	10.384
AAARRR.1061	A7	R20	3.761
AAARRR.1061	A7	R21	2.808
AAARRR.1061	R19	R20	6.632
AAARRR.1061	R19	R21	8.546
AAARRR.1061	R20	R21	2.423



(a) (b)

Figure 2: Mapping of the (a) active compounds and (b) inactive compounds on the pharmacophore.

Table 3: PLS statistical parameters of the model AAARRR.1061.

PLS	SD	R^2	F	P	Stability	RMSE	Q^2	Pearson R
1	0.4189	0.6288	81.3	6.698e-12	0.9411	0.3486	0.5412	0.8756
2	0.3627	0.7276	62.8	5.339e-14	0.8681	0.3198	0.6137	0.874
3	0.2995	0.8182	69	4.698e-17	0.6923	0.3733	0.4737	0.7448
4*	0.2606	0.8653	72.3	5.278e-19	0.6111	0.334	0.7186	0.7787
5	0.2238	0.9029	81.8	3.836e-21	0.5238	0.4009	0.393	0.6463
6	0.1941	0.9286	93.2	5.226e-23	0.4545	0.4409	0.2658	0.5743

SD: standard deviation of regression; R^2: regression coefficient; F: ratio of the model variance to the observed activity variance (variance ratio); P: significance level of variance ratio; Q^2: cross-validated correlation coefficient for the test set; RMSE: the RMS error in the test set predictions. *Best model.

activity prediction ability. Then, we screened the database based on the best pharmacophore model and docked the selected compounds into the colchicine binding site of tubulin. Finally, we selected molecules with the highest docking score as tubulin inhibitor candidates and anticancer agents.

2. Materials and Methods

2.1. Data Set and Ligand Preparation. For the preparation of common pharmacophore through 3D-QSAR studies, a set including sixty-two quinolines from our previous studies, with cytotoxic activity against A2780 (human ovarian carcinoma) cell line, was selected and the pIC_{50} ($pIC_{50} = -\log IC_{50}$) values were calculated [24].

The data set was randomly divided into training and test sets for generation and validation of the model, respectively [25]. The 3D structures of ligands were generated using the builder panel in Maestro and successively optimized using LigPrep module (v4.3, Schrodinger 2016-1) [26]. The energy minimization was done using OPLS_2005 (optimized potentials for liquid simulations) with an implicit distance-dependent dielectric solvation treatment [27].

2.2. Pharmacophore 3D-QSAR Modeling. For the generation of pharmacophore and 3D-QSAR models for anticancer agents, Phase (v4.3, Schrodinger 2016-1 was used [28]. The data set ligands were categorized into active, with the threshold of $pIC_{50} > 5.5$, and inactive, with the threshold of $pIC_{50} < 4.7$ for the generation of common pharmacophore hypotheses [29, 30].

The default settings were used to generate acceptable conformations, and a maximum of 100 conformers were generated. Alignment was done, and a maximum of one conformer was retained for every ligand. Random selection was used for assigning training and test set for 62 compounds. Twelve (12) compounds were selected as a test set, and the remaining (50 compounds) were used as a training set. Pharmacophore sites were produced based on the pharmacophore features in the Phase module. There are six built-in pharmacological features in Phase, namely, hydrogen bond receptor (A), hydrogen bond donor (D), hydrophobic group (H), negatively charged group (N), positively charged group (P), and aromatic ring (R) [31]. In the generated hypotheses, six common sites were found for all selected compounds. Six of the best resulting hypotheses were scored and ranked by their vector, volume, site scores, survival scores, and survival actives showed in Table 1 [28]. AAARRR.1061 hypothesis which consisted of three hydrogen bond acceptors (A) and three aromatic ring (R) features was selected as the best model for further study (Figure 1).

Table 4: Structures and properties of train and test ligands.

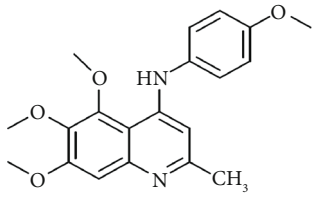
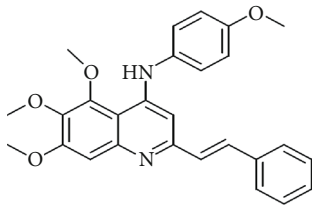
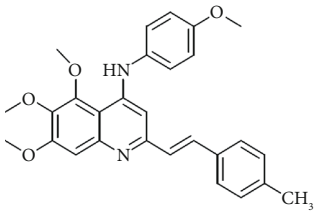
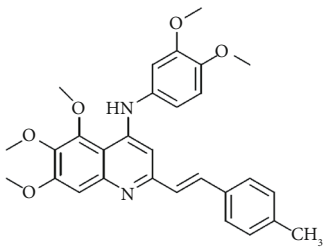
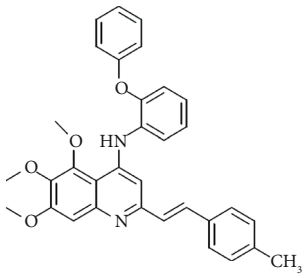
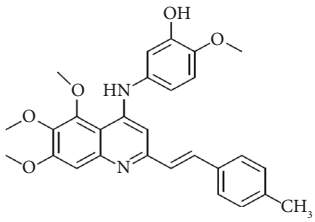
Ligand	Structure	Experimental activity (pIC_{50})	Predicted activity (pIC_{50})	Residual activity	Fitness
1		4.415	4.56	-0.145	2.29
2		5.122	5.62	-0.498	2.87
3		5.339	5.35	0.011	2.90
4		5.513	5.68	-0.167	2.84
5		4.961	4.72	0.241	2.71
6[t]		5.633	5.41	0.223	2.86

Table 4: Continued.

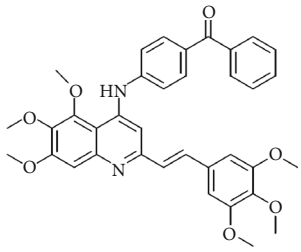
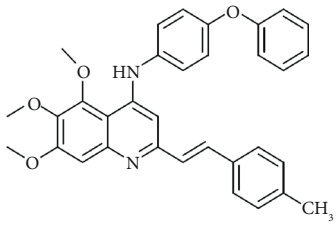
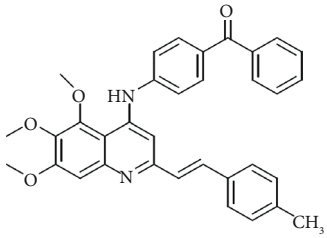
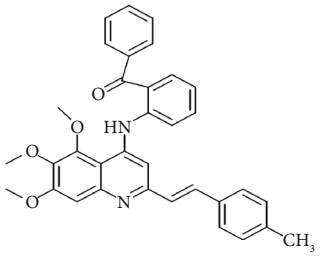
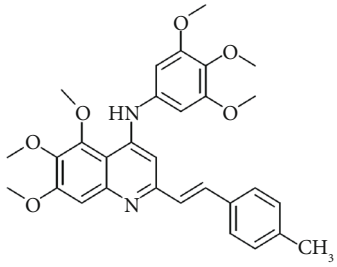
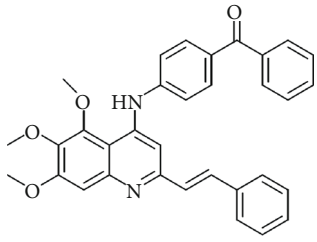
Ligand	Structure	Experimental activity (pIC_{50})	Predicted activity (pIC_{50})	Residual activity	Fitness
7^t		4.845	5.05	-0.205	2.75
8		4.922	5.17	-0.248	2.80
9		4.526	4.92	-0.394	2.80
10^t		5.140	5.24	-0.1	2.67
11		5.879	5.73	0.149	2.83
12		4.178	4.11	0.068	2.76

Table 4: Continued.

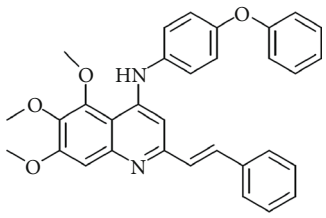
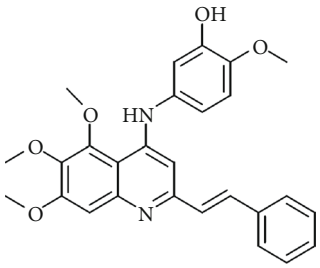
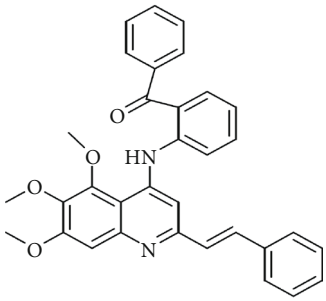
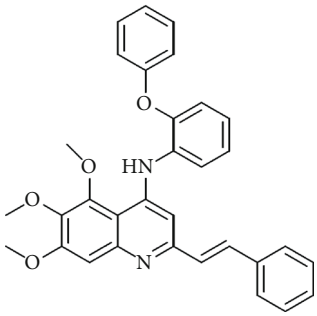
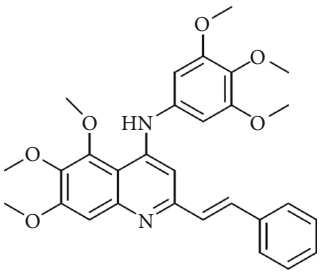
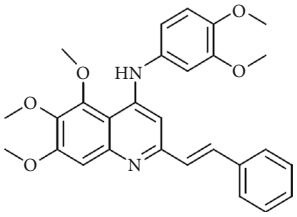
Ligand	Structure	Experimental activity (pIC_{50})	Predicted activity (pIC_{50})	Residual activity	Fitness
13[t]		4.970	5.01	-0.04	2.78
14		5.660	5.36	0.3	2.88
15		4.992	5.35	-0.358	2.69
16		5.037	4.95	0.087	2.68
17		5.479	5.77	-0.291	
18		5.254	5.35	-0.096	2.89

Table 4: Continued.

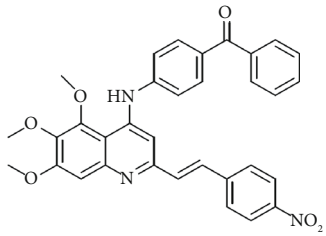
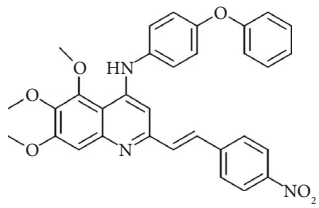
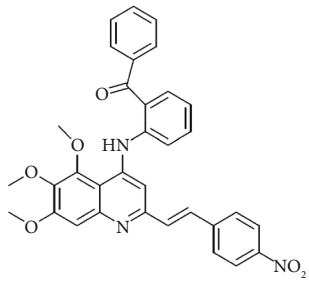
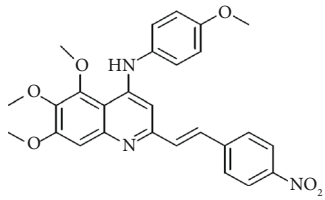
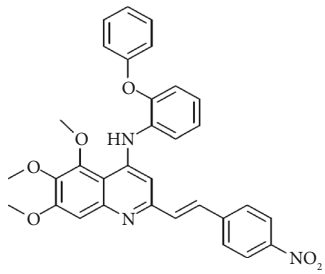
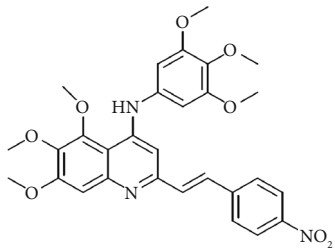
Ligand	Structure	Experimental activity (pIC_{50})	Predicted activity (pIC_{50})	Residual activity	Fitness
19		5.423	5.20	0.223	2.80
20		5.038	4.83	0.203	2.78
21		5.015	5.08	-0.065	2.73
22		6.414	5.82	0.594	2.89
23		4.734	4.79	-0.056	2.70
24		5.498	5.58	-0.082	2.87

Table 4: Continued.

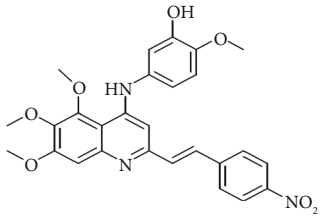
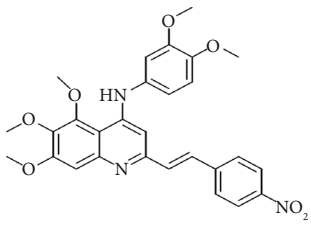
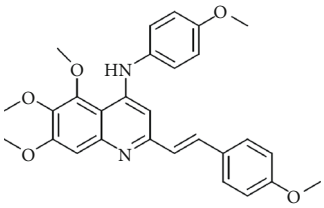
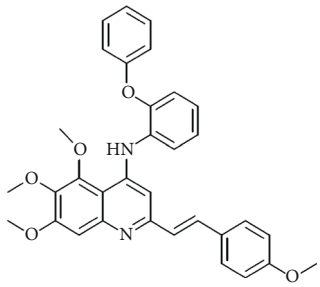
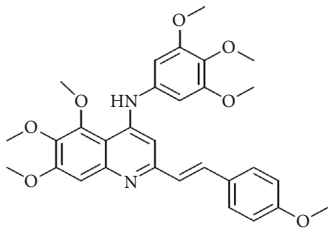
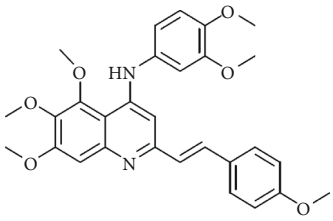
Ligand	Structure	Experimental activity (pIC_{50})	Predicted activity (pIC_{50})	Residual activity	Fitness
25		5.129	5.27	-0.141	2.87
26		5.840	5.58	0.26	2.91
27[t]		5.356	5.72	-0.364	2.90
28[t]		5.056	4.93	0.126	2.65
29[t]		5.376	5.34	0.036	2.88
30		5.460	5.53	-0.07	2.85

Table 4: Continued.

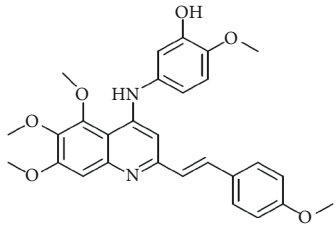
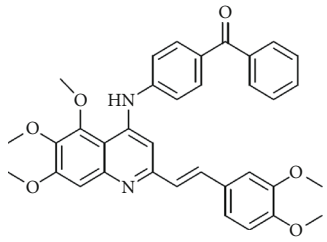
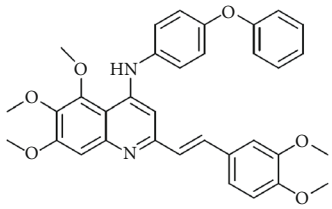
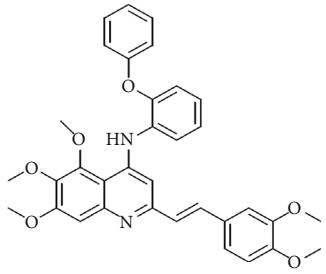
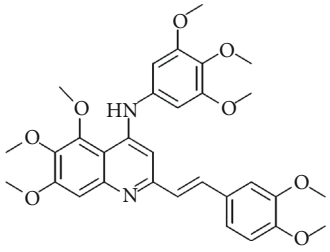
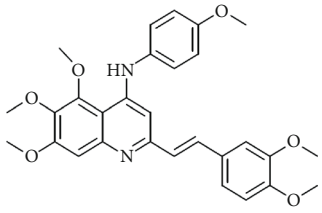
Ligand	Structure	Experimental activity (pIC_{50})	Predicted activity (pIC_{50})	Residual activity	Fitness
31		5.612	5.50	0.112	2.87
32		4.885	4.79	0.095	2.82
33[t]		5.393	5.31	0.083	2.81
34		5.149	4.95	0.199	2.73
35[t]		6.297	5.63	0.667	2.90
36		5.986	5.44	0.546	2.96

Table 4: Continued.

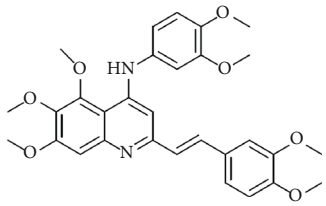
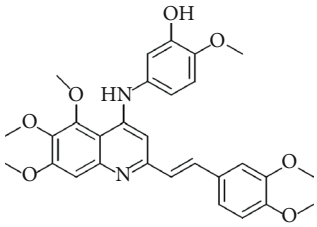
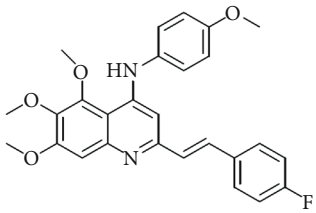
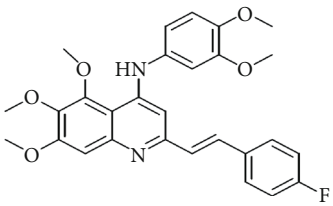
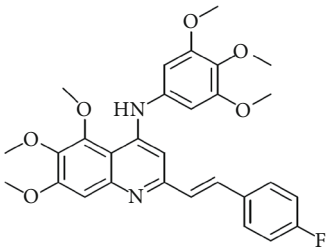
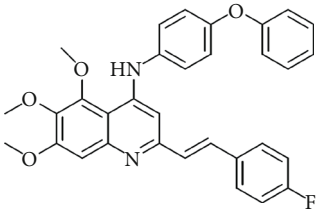
Ligand	Structure	Experimental activity (pIC_{50})	Predicted activity (pIC_{50})	Residual activity	Fitness
37[t]		5.768	5.57	0.198	2.91
38[t]		5.640	5.43	0.21	3
39[t]		5.492	5.62	-0.128	2.91
40		5.895	5.62	0.275	2.91
41		5.963	5.83	0.133	2.83
42		4.930	5.40	-0.47	2.81

Table 4: Continued.

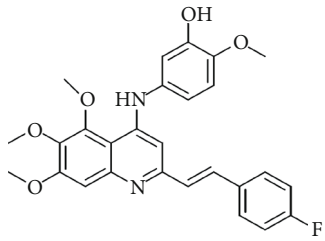
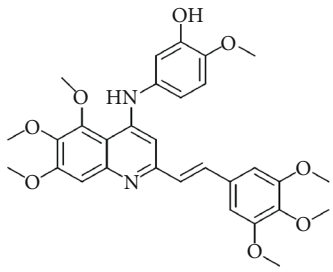
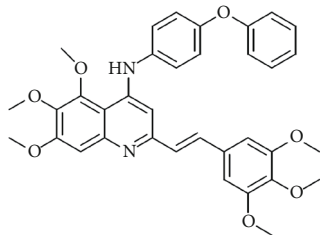
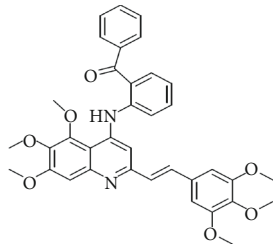
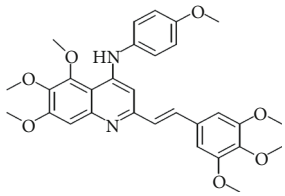
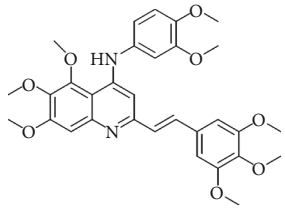
Ligand	Structure	Experimental activity (pIC_{50})	Predicted activity (pIC_{50})	Residual activity	Fitness
43		5.931	5.83	0.101	2.87
44		5.708	5.89	-0.182	2.87
45		4.594	4.78	-0.186	2.77
46		5.504	5.44	0.064	2.67
47		5.403	5.66	-0.257	2.91
48		5.539	5.55	-0.011	2.91

Table 4: Continued.

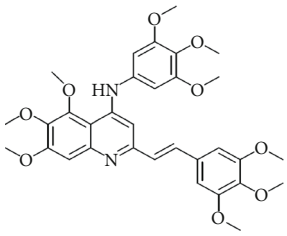
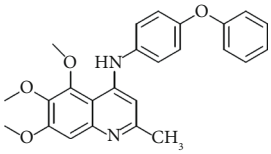
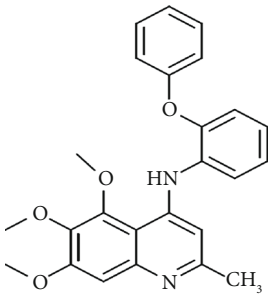
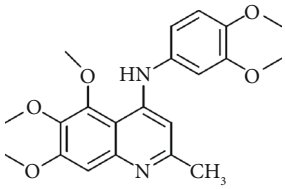
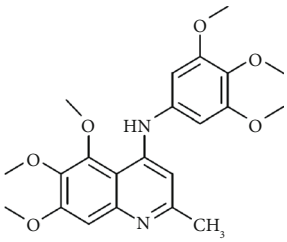
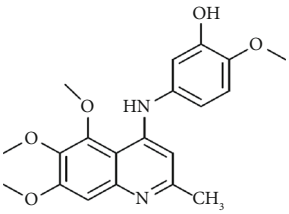
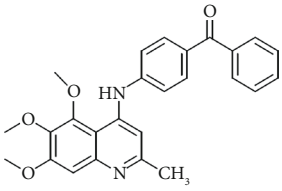
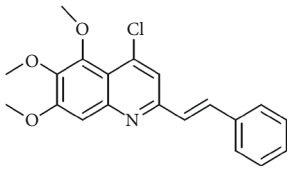
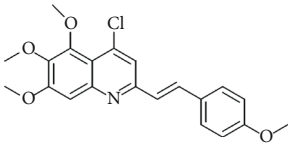
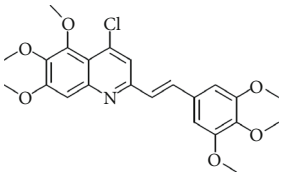
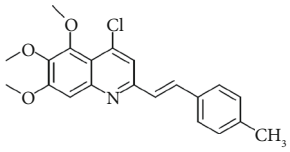
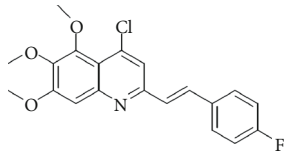
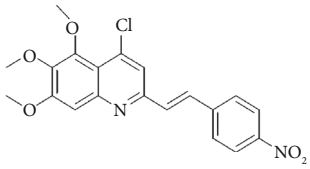
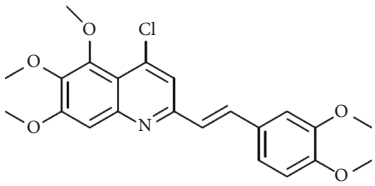
Ligand	Structure	Experimental activity (pIC_{50})	Predicted activity (pIC_{50})	Residual activity	Fitness
49		5.824	5.77	0.054	2.89
50		5.074	4.77	0.304	2.21
51		4.919	5.05	-0.131	2.14
52		4.234	4.41	-0.176	2.30
53		4.095	4.14	-0.045	2.26
54		4.459	4.44	0.019	.27
55		4.342	4.13	0.212	2.19

Table 4: Continued.

Ligand	Structure	Experimental activity (pIC_{50})	Predicted activity (pIC_{50})	Residual activity	Fitness
56		4.516	3.90	0.616	2.67
57		4.399	3.96	0.439	2.69
58		4.319	4.27	0.049	2.71
59		3.830	3.89	-0.06	2.68
60		3.734	3.95	-0.216	2.69
61		3.771	4.02	-0.249	2.69
62		3.127	3.77	-0.643	2.74

[t]Test.

The distance and angles between different sites of the model AAARRR.1061 are shown in Table 2 and Table S1 (supplementary data), respectively. The alignment of active and inactive compounds with the generated common pharmacophore was shown in Figures 2(a) and 2(b). PLS statistical parameters of the model AAARRR.1061 were shown in Table 3. Structure of compounds, experimental and predicted inhibitory activities (pIC_{50} values), residual values, and fitness score of all the ligands were reported in Table 4.

2.3. Model Validation. The last step of developing the QSAR model was model validation [32, 33]. The developed pharmacophore hypothesis was validated using potent approaches like the Y-Randomization test and ROC-AUC analysis [34]. The statistical parameters, including the squared correlation coefficient (R^2), cross-validation (leave one out) Q^2, the standard deviation of regression (SD), Pearson's correlation coefficient (Pearson's R), statistical significance (P), root mean square error (RMSE), and variance ratio (F), were shown in Table 3 [35].

2.3.1. Y-Randomization Test. To ensure the validity of our QSAR model, the Y-Randomization technique was performed. The dependent variable vectors were randomly shuffled, and a new QSAR model was developed. The procedure was repeated several times, and the new QSAR models were developed for each random. The resulting R^2 and Q^2 values were compared to that of the original model [36, 37].

2.3.2. ROC-AUC Analysis. To evaluate our hypothesis, receiver operating characteristic (ROC) curve analysis was also performed using the MedCalc statistical software (http://www.medcalc.org). In ROC analysis, the ability of the obtained pharmacophore model was indicated with the area under the curve (AUC) to distinguish a list of compounds as active or inactive compounds in terms of two parameters, sensitivity, and specificity [24].

2.4. Drug-Likeness Filtration and Virtual Screening. The IBScreen database containing 211432 compounds was selected. The drug-likeness behavior of the database compounds was predicted by using QikProp version 4.3 (Schrödinger). The compounds were employed for the calculation of pharmacokinetic parameters by QikProp v4.3. Physiochemical descriptors and pharmaceutically relevant properties of compounds were evaluated to analyze druggable properties [27, 38]. Lipinski's rule of five was used to filter the compounds with drug-like properties [39]. To identify the best match molecules, the AAARRR.1061 hypothesis was applied for screening the IBScreen database with drug-like properties. Finally, we selected the compounds with the pIC_{50} value of more than 4 which were considered as the most active compounds, and then, further screening of these compounds is done by ADMET (absorption, distribution, metabolism, elimination, and toxicity) properties using QikProp version 4.3. The compounds complied with Lipinski's rule, with good predicted activity and good ADMET properties, were selected for molecular docking studies [24, 32].

2.5. Molecular Docking. The docking study was performed using Glide module in Schrodinger suite 2016-1. The crystal structure of tubulin in complex with colchicine (PDB Code: 4O2B) was obtained from protein data bank Brookhaven (*Protein Data Bank (PDB)* at *Brookhaven* National Laboratory). The protein was prepared using the protein preparation wizard in Schrodinger [40]. All the water molecules were deleted, *hydrogen atoms* were added, and energy minimization was performed using the OPLS_2005 force field [26]. The active site was defined with a radius of 15 Å around the ligand present in the crystal structure. The grid box was generated at a centroid of the active site. The compounds were docked into the catalytic domain of tubulin protein (PDB Code: 4O2B) using Grid-based Ligand Docking [41]. The best docked structures were identified using Dock score and Glide energy (Figure S1: supplementary data). The compound STOCK2S-23597 with the lowest docking energy was selected for further studies (Figure 3).

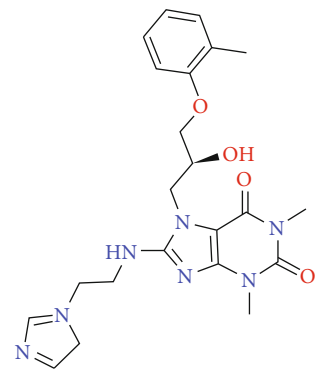


Figure 3: Compound STOCK2S-23597 with the highest docking score.

3. Results and Discussion

3.1. Pharmacophore and 3D-QSAR Models. For the development of the pharmacophore model, the Phase (v4.3) module of Schrodinger 2016-1 was used [28]. All the 62 compounds were randomly divided into a training set (50 compounds), to identify the common pharmacophore hypothesis, and a test set (12 compounds). A maximum of six sites was chosen to find the optimum combination of features to the most active compounds in the data set. In total, 279 common pharmacophore hypotheses were generated using different combinations of variants, such as AAARRR, AAAHRR, AAHRRR, AHHRRR, AHHHRR, and AAHHRR with survival scores ranging between 3.48 and 3.87. Six of the best hypotheses were shown in Table 1.

The best-fitted model AAARRR.1061 consists of three hydrogen bond acceptors (A) and three aromatic ring (R) features, and regression scores of this (AAARRR.1061) pharmacophore hypothesis were further analyzed by PLS in the Phase module using 6 PLS factors (Table 3).

In the regression model, R^2 was used to describe the fitness of data, the correlation coefficient (Q^2) was used to check the external predictability, and the significance of the model was measured by the Fisher ratio (F) [31, 36, 37]. Thus, the best QSAR model was chosen based on maximum survival score, good statistical value, good predictive power, and lowest relative conformational energy. Scatter plots for experimental and predicted activities of ligands showed a significant linear correlation of training and test set compounds (Figures 4(a) and 4(b)).

The reliability and predictability of the common pharmacophore model, AAARRR.1061, based on active compounds were determined. The PLS of four was selected as the best model (Table 3). The relevance of the model was displayed by the regression coefficient ($R^2 = 0.865$) of the training and cross-validation coefficient ($Q^2 = 0.718$). The stability of the generated models ranged from 0.454 to 0.941. The F value was found to be 72.3 and a P value of 5.278e-019. Moreover, the greater degree of confidence in the model is indicated by Pearson R of 0.778. Standard deviation (SD) value of 0.260 and root mean square error (RMSE) of 0.334 showed the ability of the generated model for prediction of unknown compound activity in the test set.

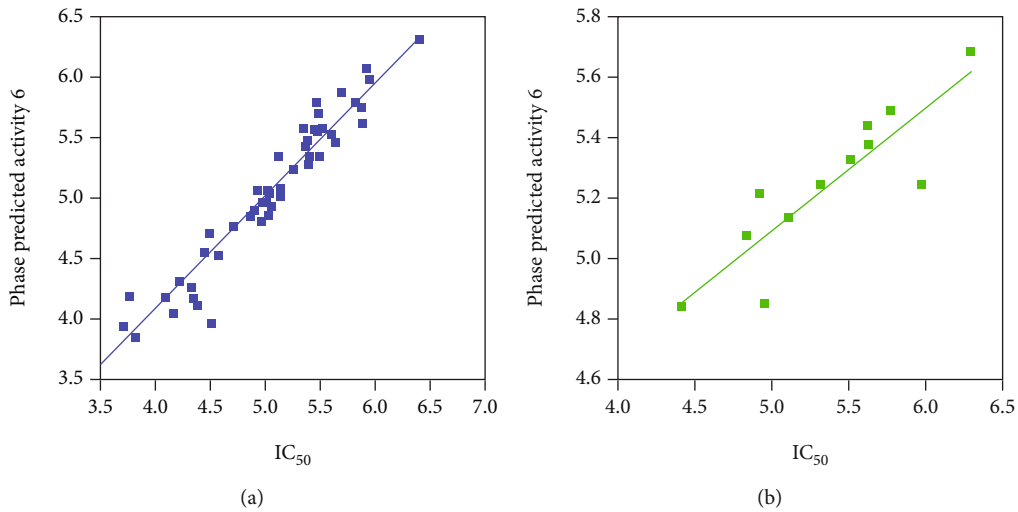


Figure 4: (a) Scatter plot of training set with best-fit line $y = 0.93x + 0.36$ ($R^2 = 0.93$) and (b) test set compounds with best-fit line $y = 0.41x + 3.07$ ($R^2 = 0.76$).

These results demonstrated that not only the pharmacophore model AAARRR.1061 could efficiently estimate the cytotoxic activity of the training set, but it also was fitted for the test set. Hence, AAARRR.1061 was validated to be a reliable pharmacophore mode to recognize cytotoxic agents with the ability to *inhibit* the *tubulin* polymerization, and then, it will be used to screen the database.

3.2. Model Validation

3.2.1. Y-Randomization Test. Validation of the model was performed by applying Y-Randomization. We built ten random and repeated all procedures to develop a model (Table S2). All the R^2 and Q^2 of the generated models from random were lower (less than 0.26) compared to the original model. This indicated that our model was not generated by chance.

3.2.2. ROC-AUC Analysis. Additional validation of the common pharmacophore model, AAARRR.1061, was performed using the AUC of the ROC curve. The ROC curve obtained for the validation showed an excellent AUC value of 0.916 (Figure 5), indicating that the model differentiated the active compounds from the inactive ones efficiently ($P < 0.001$). The sensitivity, specificity, and accuracy of the model were 74.67, 86.09, and 80.03%, respectively.

3.3. 3D-QSAR Contour Map Analysis. To analyze 3D-QSAR results, the model was superimposed on the most active ligand (compound 22) and the least active ligand (compound 62). The generated contour plots (Figures 6(a)–6(f)) showed the correlation of structural properties of compounds including electron-withdrawing, hydrophobic, and H-bond donor properties concerning their activity. Blue cubes indicated favorable regions while red cubes indicated unfavorable regions for biological activity [42, 43].

The hydrogen-bond donor nature was compared for the most active compound 22 (Figure 6(a)) and the least active compound 62 (Figure 6(b)). In Figure 6(a), blue cubes were

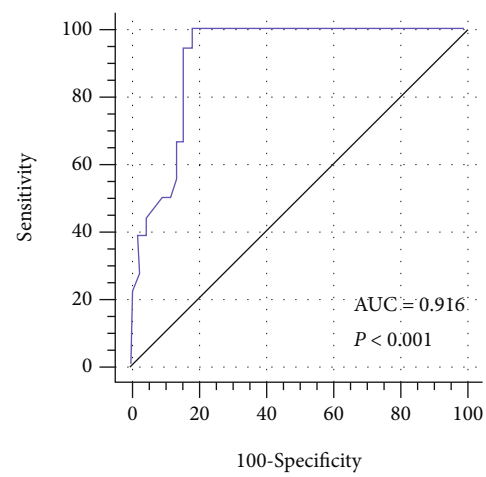


Figure 5: ROC curve obtained by AAARRR.1061 model against random curve.

observed at regions lied over the amine group present at position 4 of the quinoline ring. On the other hand, in the least active compound 62 without an amino group at the same steric position (Figure 6(b)), no blue cube was observed in the same region. Therefore, the presence of N-aryl with hydrogen donor amine group was vital for the cytotoxicity and tubulin inhibitory activity. This assumption was further supported by the low activity of compounds 65-71, which do not have N-aryl at position 4 of the quinoline ring.

Figures 6(c) and 6(d) showed the favorable and unfavorable hydrophobic features for the most active compound and least active compound.

Figure 6(c) revealed that the blue cubes were generated around the hydrophobic arylstyryl at position 2 and N-aryl at position 4 of the quinoline core were essential for anticancer activity.

In Figure 6(d), red cubes were generated at position 4 of the quinoline core of the least active compound. In this

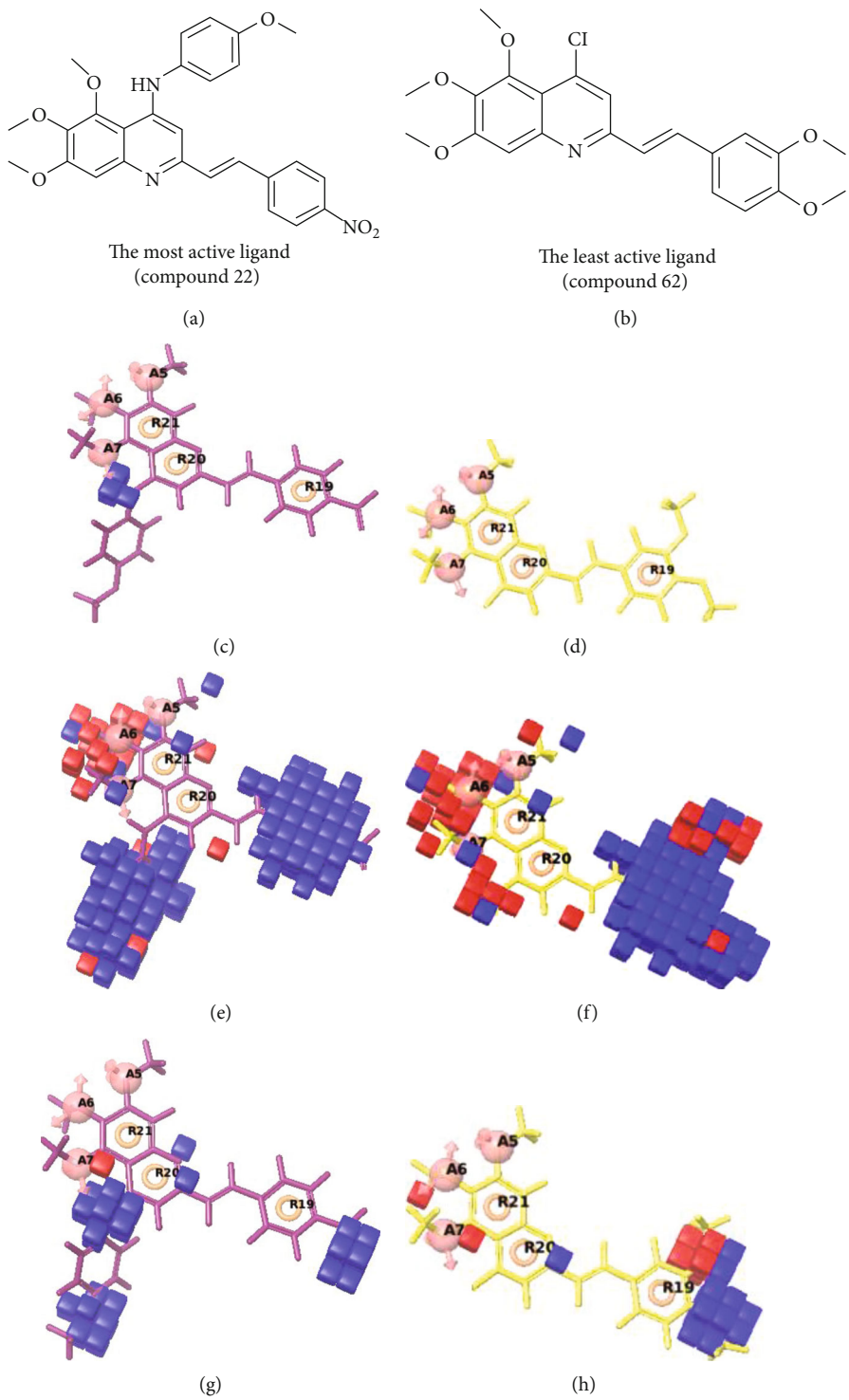


Figure 6: Pictorial representation of the contour maps generated in the context of (c, d) hydrogen bond donor, (e, f) hydrophobic features, and (g, h) electron-withdrawing groups with (a) most active compound 22 and (b) least active compound 62 using the QSAR model.

compound, a chloro substituent was present at this region instead of the hydrophobic N-aryl group. Thus, the results revealed that red colored unfavorable regions at these positions could be responsible for the decrease of activity. This was also confirmed by less activity of compounds 65-71 possessing chloro group at position 4 of the quinoline ring.

In Figure 6(e), blue cubes were observed at the para position of N-aryl indicating the preference of electron-withdrawing groups at this position (the presence of an electronegative atom, such as oxygen or nitrogen, was desirable because of the *inductive electron-withdrawing effect*). Also, blue cubes were observed at the para position of the styryl

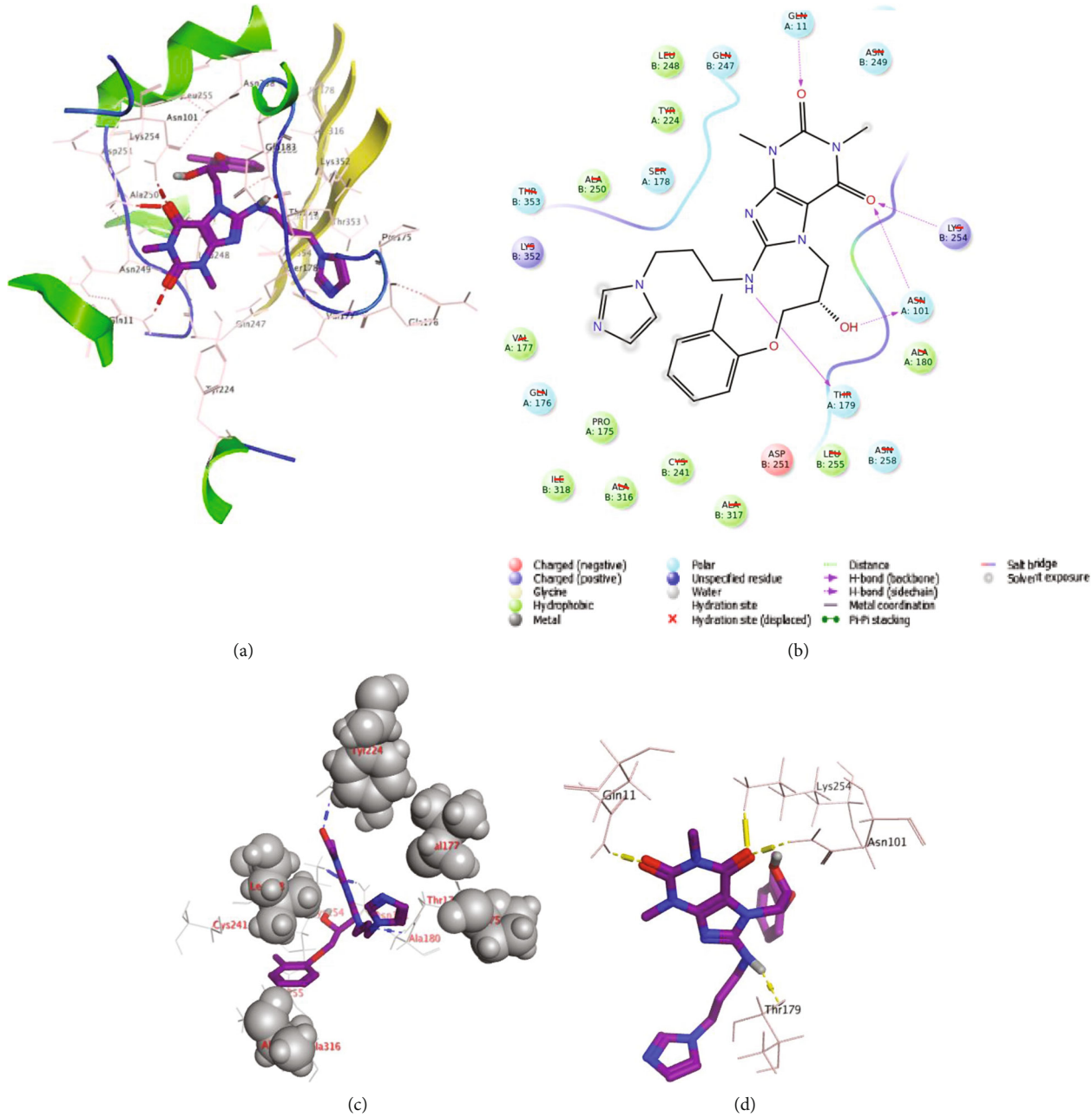


Figure 7: (a) Binding mode, (b) 2D-ligand interaction diagram, (c) hydrophobic interactions, and (d) hydrogen interactions of compound STOCK2S-23597 in the catalytic pocket of 4O2B.

group at position 2 of the quinoline core of the most active compound possessing electron-withdrawing group (NO_2). It seems that the presence of an electron-withdrawing group at styryl moiety increased the anticancer activity. The high activity of compounds 23-30 and 45-49 possessing NO_2 and F groups at styryl moiety supports this finding.

3.4. Virtual Screening. First, the IBScreen database containing 211432 compounds was selected. Lipinski's rule of five was used to scrutinize these compounds. The applied filter gave a total of 183084 compounds. The validated 3D-QSAR pharmacophore model AAARRR.1061 was used as a 3D structural query for retrieving potent compounds from the IBScreen database [24]. A thousand hits were obtained in this step for a match to hypothesis AAARRR.1061.

Compounds with a pIC_{50} value of more than 4 were selected as the active compounds, and hence, 34 compounds were obtained which were further subjected to filtering by applying ADMET properties. The result showed that 10 hits were retrieved with good ADMET properties. The descriptors used in ADMET prediction and criteria for hit selection were shown in Table S3.

Structures of the top 10 best-fit molecules were shown in figure S1 (supplementary data), and molecular properties and ADMET properties were depicted in Table S3 and S4, respectively.

3.5. Molecular Docking. Above 10 hits obtained from the virtual screening process were subjected to molecular docking studies. A tubulin complex with colchicine was chosen as the target protein for molecular docking using the Glide in Schrodinger 2016. Molecular docking studies were performed to predict the binding conformation of the compounds. The top 10 compounds retrieved by ADMET properties were docked into the binding site of tubulin. Binding interactions and binding energies of these 10 molecules with tubulin were shown in Table S5. Dock scores for the screened compounds ranged between -10.948 (STOCK2S-23597) and -5.991 (STOCK2S-05500).

Protein-ligand interactions of compound STOCK2S-23597 with the highest docking score were further analyzed (Figures 7(a)–7(d)). Molecular docking analysis of this compound showed hydrogen bond interactions with four residues of Glnα (chain A) 11 (bond length = 2.63 Å), Lysβ (chain B) 254 (bond length = 2.48 Å), Asnα101 (bond length = 2.33 Å), and Thrα 179 (bond length = 2.05 Å). Compound STOCK2S-23597 exhibited hydrophobic interactions with the key residues necessary for binding of inhibitors including Leuβ 248, Tyrα 224, Alaβ 250, Valα 177, Ileβ 318, Cysβ 241, Alaβ 316, Alaβ 317, Alaα 180, and Leuβ 255.

This compound also showed polar interactions with amino acid residues such as Asnβ 249, Serα 178, Thrβ 353, Glnα 176, and Asnβ 258.

The binding mode of the most active compound (22) is exhibited in the figure S2. In this compound nitrogen atom of the aniline ring interacted with Thrα 179, the methoxy group, interacted with Lysβ 254, and the nitro group formed a hydrogen bond with Argα 221. These results are consistent with the finding of 3D-QSAR study which showed nitrogen atom of the aniline ring, and the nitro group has key roles in the activity of compound 22.

4. Conclusion

In this study, 279 pharmacophore models were generated based on a series of quinolines as anticancer agents and tubulin inhibitors. A six-point pharmacophore model (AAARRR.1061) was identified as the best model which consisted of three hydrogen bond acceptors (A) and three aromatic ring (R) features and was then validated by Y-Randomization test and ROC-AUC analysis. The model showed a high correlation coefficient ($R^2 = 0.865$), cross-validation coefficient ($Q^2 = 0.718$), F value (72.3), and a P value of 5.278e-019 at 6 component PLS level. The contour maps obtained for the model with the most active and the least active compounds confirmed that the presence of N-aryl with hydrogen donor amine group at position 4 of the quinoline ring, the arylstyryl hydrophobic group at position 2, and N-aryl at position 4 of the quinoline core and the presence electron-withdrawing group at the para position of arylstyryl group were vital for the cytotoxicity and tubulin inhibitory activity.

AAARRR.1061 was used as a 3D query to screen the IBScreen database, and we obtained 1000 compounds. Compounds with a pIC_{50} value of more than 4 (34 compounds) were selected as the most active compounds. After applying ADMET properties, 10 compounds were selected for further docking studies. Ultimately, compound STOCK2S-23597 with the highest docking score (-10.948 kcal/mol) was selected as a potent tubulin inhibitor. It formed four hydrogen bonds and hydrophobic interactions with tubulin active site residues.

The obtained results suggested that the proposed 3D-QSAR model and hits obtained on virtual screening of the database have provided new chemical starting points for design and development of novel tubulin targeting agents.

Data Availability

Data are available upon a reasonable request from the corresponding authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We thank Research Deputy of Mashhad University of Medical Sciences for financial support of this study (as part of thesis of Salimeh Mirzaei).

Supplementary Materials

Figure S1: structures of top 10 best-fit molecules from the IBScreen database. Figure S2: 2D-ligand interaction diagram of compound 22 in the catalytic pocket of 4O2B. Table S1: intersite angles between the pharmacophoric sites of AAARRR.1061. Table S2: R^2 and Q^2 values after several Y-Randomisation test. Table S3: the molecular property descriptors used in ADMET prediction. Table S4: the descriptors used in ADMET prediction. Table S5: binding interactions of best-fit IBScreen database compounds. *(Supplementary Materials)*

References

[1] R. L. Siegel, K. D. Miller, S. A. Fedewa et al., "Colorectal cancer statistics, 2017," *CA: a Cancer Journal for Clinicians*, vol. 67, no. 3, pp. 177–193, 2017.

[2] F. Jafari, H. Baghayi, P. Lavaee et al., "Design, synthesis and biological evaluation of novel benzo-and tetrahydrobenzo-[h] quinoline derivatives as potential DNA-intercalating antitumor agents," *European Journal of Medicinal Chemistry*, vol. 164, pp. 292–303, 2019.

[3] A. Kamal, P. Srikanth, M. Vishnuvardhan et al., "Combretastatin linked 1,3,4-oxadiazole conjugates as a Potent Tubulin Polymerization inhibitors," *Bioorganic Chemistry*, vol. 65, pp. 126–136, 2016.

[4] I. Mignot, L. Pecqueur, A. Dorléans et al., "Design and characterization of modular scaffolds for tubulin assembly," *Journal of Biological Chemistry*, vol. 287, no. 37, pp. 31085–31094, 2012.

[5] S. Mirzaei, F. Eisvand, F. Hadizadeh, F. Mosaffa, and R. Ghodsi, "Design, synthesis, and biological evaluation of novel 5, 6, 7-trimethoxy quinolines as potential anticancer agents and tubulin polymerization inhibitors," *Iranian Journal of Basic Medical Sciences*, vol. 23, no. 12, pp. 1527–1537, 2020.

[6] S. H. Abbas, G. E.-D. A. Abuo-Rahma, M. Abdel-Aziz, O. M. Aly, E. A. Beshr, and A. M. Gamal-Eldeen, "Synthesis, cytotoxic activity, and tubulin polymerization inhibitory activity of new pyrrol-2(3*H*)-ones and pyridazin-3(2*H*)-ones," *Bioorganic Chemistry*, vol. 66, pp. 46–62, 2016.

[7] A. Kamal, C. R. Reddy, M. Vishnuvardhan et al., "Synthesis and biological evaluation of cinnamido linked benzophenone hybrids as tubulin polymerization inhibitors and apoptosis inducing agents," *Bioorganic & Medicinal Chemistry Letters*, vol. 24, no. 10, pp. 2309–2314, 2014.

[8] S. Mirzaei, F. Eisvand, F. Hadizadeh, F. Mosaffa, A. Ghasemi, and R. Ghodsi, "Design, synthesis and biological evaluation of novel 5,6,7-trimethoxy-*N*-aryl-2-styrylquinolin-4-amines as potential anticancer agents and tubulin polymerization inhibitors," *Bioorganic Chemistry*, vol. 98, article 103711, 2020.

[9] R. O. Carlson, "New tubulin targeting agents currently in clinical development," *Expert Opinion on Investigational Drugs*, vol. 17, no. 5, pp. 707–722, 2008.

[10] S. Mirzaei, F. Hadizadeh, F. Eisvand, F. Mosaffa, and R. Ghodsi, "Synthesis, structure-activity relationship and molecular docking studies of novel quinoline-chalcone hybrids as potential anticancer agents and tubulin inhibitors," *Journal of Molecular Structure*, vol. 1202, p. 127310, 2020.

[11] S. Sengupta and S. A. Thomas, "Drug target interaction of tubulin-binding drugs in cancer therapy," *Expert Review of Anticancer Therapy*, vol. 6, no. 10, pp. 1433–1447, 2006.

[12] F. S. Behbahani, J. Tabeshpour, S. Mirzaei et al., "Synthesis and biological evaluation of novel benzo [c] acridine-diones as potential anticancer agents and tubulin polymerization inhibitors," *Archiv der Pharmazie*, vol. 352, no. 6, article 1800307, 2019.

[13] N. Shobeiri, M. Rashedi, F. Mosaffa et al., "Synthesis and biological evaluation of quinoline analogues of flavones as potential anticancer agents and tubulin polymerization inhibitors," *European Journal of Medicinal Chemistry*, vol. 114, pp. 14–23, 2016.

[14] M. A. Jordan and L. Wilson, "Microtubules as a target for anticancer drugs," *Nature Reviews Cancer*, vol. 4, no. 4, pp. 253–265, 2004.

[15] N. Mahindroo, J.-P. Liou, J.-Y. Chang, and H.-P. Hsieh, "Antitubulin agents for the treatment of cancer–a medicinal chemistry update," *Expert Opinion on Therapeutic Patents*, vol. 16, no. 5, pp. 647–691, 2006.

[16] C. Dumontet and M. A. Jordan, "Microtubule-binding agents: a dynamic field of cancer therapeutics," *Nature Reviews Drug Discovery*, vol. 9, no. 10, pp. 790–803, 2010.

[17] C. Dumontet and B. I. Sikic, "Mechanisms of action of and resistance to antitubulin agents: microtubule dynamics, drug transport, and cell death," *Journal of Clinical Oncology*, vol. 17, no. 3, pp. 1061–1070, 1999.

[18] E. Karimikia, J. Behravan, A. Zarghi, M. Ghandadi, S. O. Malayeri, and R. Ghodsi, "Colchicine-like β-acetamidoketones as inhibitors of microtubule polymerization: design, synthesis and biological evaluation of in vitro anticancer activity," *Iranian Journal of Basic Medical Sciences*, vol. 22, p. 1138, 2019.

[19] S. Mirzaei, M. Qayumov, F. Gangi, J. Behravan, and R. Ghodsi, "Synthesis and biological evaluation of oxazinonaphthalene-3-one derivatives as potential anticancer agents and tubulin inhibitors," *Iranian Journal of Basic Medical Sciences*, vol. 23, p. 1388, 2020.

[20] X. Lin, X. Li, and X. Lin, "A review on applications of computational methods in drug screening and design," *Molecules*, vol. 25, no. 6, p. 1375, 2020.

[21] N. Zhou, Y. Xu, X. Liu et al., "Combinatorial pharmacophore-based 3D-QSAR analysis and virtual screening of FGFR1 inhibitors," *International Journal of Molecular Sciences*, vol. 16, no. 12, pp. 13407–13426, 2015.

[22] F. Fan, D. Toledo Warshaviak, H. K. Hamadeh, and R. T. Dunn, "The integration of pharmacophore-based 3D QSAR modeling and virtual screening in safety profiling: a case study to identify antagonistic activities against adenosine receptor, A2A, using 1,897 known drugs," *PLoS One*, vol. 14, no. 1, article e0204378, 2019.

[23] Y.-H. Chen, Z.-S. Yang, C.-C. Wen et al., "Evaluation of the structure–activity relationship of flavonoids as antioxidants and toxicants of zebrafish larvae," *Food Chemistry*, vol. 134, no. 2, pp. 717–724, 2012.

[24] D. R. Gade, A. Makkapati, R. B. Yarlagadda, G. J. Peters, B. Sastry, and V. V. S. Rajendra Prasad, "Elucidation of chemosensitization effect of acridones in cancer cell lines: combined pharmacophore modeling, 3D QSAR, and molecular dynamics studies," *Computational Biology and Chemistry*, vol. 74, pp. 63–75, 2018.

[25] T. Abdizadeh, R. Ghodsi, and F. Hadizadeh, "3D-QSAR (CoMFA, CoMSIA) and molecular docking studies on histone deacetylase 1 selective inhibitors," *Recent Patents on Anti-Cancer Drug Discovery*, vol. 12, no. 4, pp. 365–383, 2017.

[26] D. Shivakumar, J. Williams, Y. Wu, W. Damm, J. Shelley, and W. Sherman, "Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field," *Journal of Chemical Theory and Computation*, vol. 6, no. 5, pp. 1509–1519, 2010.

[27] M. F. Khan, G. Verma, W. Akhtar et al., "Pharmacophore modeling, 3D-QSAR, docking study and ADME prediction of acyl 1, 3, 4-thiadiazole amides and sulfonamides as antitubulin agents," *Arabian Journal of Chemistry*, vol. 12, no. 8, pp. 5000–5018, 2019.

[28] S. L. Dixon, A. M. Smondyrev, E. H. Knoll, S. N. Rao, D. E. Shaw, and R. A. Friesner, "PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results," *Journal of computer-aided molecular design*, vol. 20, no. 10-11, pp. 647–671, 2006.

[29] M. D. Miller, R. P. Sheridan, and S. K. Kearsley, "SQ: a program for rapidly producing pharmacophorically relevent molecular superpositions," *Journal of Medicinal Chemistry*, vol. 42, no. 9, pp. 1505–1514, 1999.

[30] G. M. Sastry, S. L. Dixon, and W. Sherman, "Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring," *Journal of Chemical Information and Modeling*, vol. 51, no. 10, pp. 2455–2466, 2011.

[31] T. Sindhu and P. Srinivasan, "Pharmacophore modeling, 3D-QSAR and molecular docking studies of benzimidazole derivatives as potential FXR agonists," *Journal of Receptors and Signal Transduction*, vol. 34, no. 4, pp. 241–253, 2014.

[32] S. G. Bhansali and V. M. Kulkarni, "Pharmacophore generation, atom-based 3D-QSAR, docking, and virtual screening studies of p38-α mitogen activated protein kinase inhibitors: pyridopyridazin-6-ones (part 2)," *Research and Reports in Medicinal Chemistry*, vol. 4, pp. 1–21, 2013.

[33] E. Oskoueian, N. Abdullah, R. Hendra, and E. Karimi, "Bioactive compounds, antioxidant, xanthine oxidase inhibitory, tyrosinase inhibitory and anti-inflammatory activities of selected agro-industrial by-products," *International Journal of Molecular Sciences*, vol. 12, no. 12, pp. 8610–8625, 2011.

[34] Y. Jiang and H. Gao, "Pharmacophore-based drug design for the identification of novel butyrylcholinesterase inhibitors against Alzheimer's disease," *Phytomedicine*, vol. 54, pp. 278–290, 2019.

[35] S. L. Dixon, A. M. Smondyrev, and S. N. Rao, "PHASE: a novel approach to pharmacophore modeling and 3D database searching," *Chemical Biology & Drug Design*, vol. 67, no. 5, pp. 370–372, 2006.

[36] M. Shen, C. Béguin, A. Golbraikh, J. P. Stables, H. Kohn, and A. Tropsha, "Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds," *Journal of Medicinal Chemistry*, vol. 47, no. 9, pp. 2356–2364, 2004.

[37] A. Tropsha, P. Gramatica, and V. K. Gombar, "The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models," *QSAR & Combinatorial Science*, vol. 22, no. 1, pp. 69–77, 2003.

[38] S. Jana, A. Ganeshpurkar, and S. K. Singh, "Multiple 3D-QSAR modeling, e-pharmacophore, molecular docking, and in vitro study to explore novel AChE inhibitors," *RSC Advances*, vol. 8, no. 69, pp. 39477–39495, 2018.

[39] S. Sakkiah and K. W. Lee, "Pharmacophore-based virtual screening and density functional theory approach to identifying novel butyrylcholinesterase inhibitors," *Acta Pharmacologica Sinica*, vol. 33, no. 7, pp. 964–978, 2012.

[40] G. Madhavi Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, and W. Sherman, "Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments," *Journal of Computer-Aided Molecular Design*, vol. 27, no. 3, pp. 221–234, 2013.

[41] R. A. Friesner, R. B. Murphy, M. P. Repasky et al., "Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein– ligand complexes," *Journal of Medicinal Chemistry*, vol. 49, no. 21, pp. 6177–6196, 2006.

[42] Y. Luo, K.-M. Qiu, X. Lu, K. Liu, J. Fu, and H.-L. Zhu, "Synthesis, biological evaluation, and molecular modeling of cinnamic acyl sulfonamide derivatives as novel antitubulin agents," *Bioorganic & Medicinal Chemistry*, vol. 19, no. 16, pp. 4730–4738, 2011.

[43] X.-H. Yang, Q. Wen, T.-T. Zhao et al., "Synthesis, biological evaluation, and molecular docking studies of cinnamic acyl 1, 3, 4-thiadiazole amide derivatives as novel antitubulin agents," *Bioorganic & Medicinal Chemistry*, vol. 20, no. 3, pp. 1181–1187, 2012.

Hindawi
BioMed Research International
Volume 2021, Article ID 3805748, 7 pages
https://doi.org/10.1155/2021/3805748

Research Article

Implementing PSO-ELM Model to Approximate Trolox Equivalent Antioxidant Capacity as One of the Most Important Biological Properties of Food

Marischa Elveny,[1] **Ravil Akhmadeev**,[2] **Mina Dinari**,[3] **Walid Kamal Abdelbasset**,[4,5] **Dmitry O. Bokov**,[6,7] **and Mohammad Mahdi Molla Jafari**[8]

[1]*DS & CI Research Group, Universitas Sumatera Utara, Medan, Indonesia*
[2]*Department of Accounting and Taxation, Plekhanov Russian University of Economics (PRUE), Stremyanny Lane 36, 117997 Moscow, Russia*
[3]*Department of Law, Faculty of Economics and Social Sciences, Shahid Chamran University of Ahwaz, Ahvaz, Iran*
[4]*Department of Health and Rehabilitation Sciences, College of Applied Medical Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia*
[5]*Department of Physical Therapy, Kasr Al-Aini Hospital, Cairo University, Giza, Egypt*
[6]*Institute of Pharmacy, Sechenov First Moscow State Medical University, 8 Trubetskaya St., Bldg. 2, Moscow 119991, Russia*
[7]*Laboratory of Food Chemistry, Federal Research Center of Nutrition, Biotechnology and Food Safety, 2/14 Ustyinsky Pr., Moscow 109240, Russia*
[8]*Department of Petroleum Engineering, Ahwaz Faculty of Petroleum Engineering, Petroleum University of Technology (PUT), Ahvaz, Iran*

Correspondence should be addressed to Marischa Elveny; marischaelveny@usu.ac.id
and Mohammad Mahdi Molla Jafari; mohammad.molajafari@afp.put.ac.ir

Received 1 July 2021; Revised 10 July 2021; Accepted 16 July 2021; Published 3 August 2021

Academic Editor: Alireza Baghban

In this paper, the Trolox equivalent antioxidant capacity (TEAC) is estimated through a robust machine-learning algorithm known as the Particle Swarm Optimization-based Extreme Learning Machine (PSO-ELM) model. For this purpose, a large dataset from previously published reports was gathered. Various analyses were performed to evaluate the proposed model. The results of the statistical analysis showed that this model can predict the actual values with high accuracy, so that the calculated R^2 and RMSE values were equal to 0.973 and 3.56, respectively. Sensitivity analysis was also performed on the effective input parameters. The leverage technique was also performed to check the accuracy of real data, and the results showed that the majority of data are reliable. This simple yet accurate model can be very powerful in predicting the Trolox equivalent antioxidant capacity values and can be a good alternative to laboratory data.

1. Introduction

Many antioxidant compounds may be found in vegetable materials [1, 2]. To find possible sources of natural antioxidants for use in edible products, many plants have been researched and several substances have been identified [3, 4].

Biologists and physicians are also interested in antioxidants because of their use for guarding the human organs against the harm of reactive oxygen species (ROS) [5, 6]. Because of the alleged tight relationship between oxidative stress and illness, antioxidants are thought to be preventive agents against similar illnesses [7, 8].

There is a negative relationship between consuming fruits and vegetables, as the greatest sources of antioxidants, and the cancer risk, so that the risk is reduced by 30-50% [9, 10]. The antioxidant potential of a wide range of substances

is investigated by many researchers; the substances include different vegetables and fruits, drinks, green teas and black ones, plant parts (their leaves, the roots, and their bark), kernel, shell, and also industrial wastes and secondary products apart from the ones derived from plants [11]. The "antioxidant capacity" refers to the certain amount of free radicals being removed by a laboratory solution, without considering the properties of any antioxidant of the mix [12, 13].

The antioxidant capacity of the Trolox equivalent of nutritional excerpts is then measured by taking into account the aggregate activity of all antioxidants of the excerpts, together with their chain-breaking, cleansing, and chelating impacts, hence offering an inclusive factor instead of the measurable antioxidants' simple sum [14]. So we can identify the antioxidants' familiar and unfamiliar capacities, and their synergistic relationship, providing a method for identifying a diverse variety of nutrition for antioxidant characteristics [15, 16]. There are many kinds of research for measuring the antioxidant capacity of Trolox equivalent of nutritional products [17–21].

The basis of each technique is producing various free radicals via a range of methods, followed by measuring a variety of endpoints at a defining moment or above the limit [22]. A spectrophotometric test of the antioxidant capacity of the Trolox equivalent is a frequent technique, which has relied on the antioxidants' relative capacities that exist in nutritional products to remove the radical cation of $ABTS^{+}$ (2,2′-azinobis-(3-ethyl-benzothiazoline-6-sulfonic acid)), comparing to the capacity of the antioxidant of 6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid (Trolox) standard amounts [17, 23]. We previously evaluated the substances of physiologically active chemicals and the cruciferous seeds' Trolox equivalent antioxidant capacity during sprouting, and the findings on ascorbic acid levels were previously reported [24].

The present paper is aimed at developing a PSO-ELM model for forecasting the antioxidant capacity of the Trolox equivalent of various sprouting cruciferous seeds dependent upon the entire phenolic composites, inositol hexaphosphate, glucosinolates, soluble proteins, ascorbic acid, and the entire tocopherol content (as an add-up of α-T, β-T, γ-T, and δ-T labeled as T_{tot}) and the antioxidant capacity of the Trolox equivalent of sprouted cruciferous seeds, as established experimentally. After the model construction stage, various analyses are used to evaluate its accuracy. Sensitivity analysis is also used to determine the effect of each of the input parameters on the target values.

2. Description of Modeling

2.1. PSO. One of the techniques for stochastic optimization is PSO which was presented by Eberhart and Kennedy [25]. The application and manual for this procedure are introduced in [26–28]. An abstract of this method is summarized in six steps which will be mentioned as follows [29].

$$v_k^i(t+1) = wv_k^i(t) + c_1.\text{rand}\,()\left(p_k^i(t) - x_k^i(t)\right) + c_2.\text{rand}\,()\left(g_k^i(t) - x_k^i(t)\right), \quad (1)$$

$$x_k^i(t+1) = x_k^i(t) + v_k^i(t+1) 1 \le i \le N, \quad 1 \le K \le D. \quad (2)$$

Step 1: a majority of stochastic solutions is formed as the searching space. Assume that the searching space has two parameters: dimension (D) and particle number (N). Each possible solution is dedicated to two attributes: position and velocity of the ith particle in iteration k. These particles are then "flown" through the search space of possible solutions which are indicated as follows. *Step 2*: measure the fitness between each particle in the swarm. *Step 3*: for each iteration, evaluate the particle's fitness with the best fitness acquired in the previous ones. If this value is better than the best acquired previous ones, replace the amount and location of the previous one with the current value and location, respectively. *Step 4*: evaluate some fragments together and update the finest place with the best fitness (g_k^i). *Step 5*: the pace of each fragment is increased towards its (g_k^i) and (p_k^i). This speed or acceleration is valued by an accidental term. *Step 6*: start again from step 2 based on favorable factors until a new convergence is reached.

2.2. ELM. An ELM can be introduced as a least square-based single hidden layer (HL) feed-forward neural network (SLFN) for two problems: regression and classification. Huang et al., for the design of an ELM, apply the kernel function instead of the HL with a huge amount of nodes [30]. Hung et al. and Pal and Deswal [31] both suggested techniques; the abstract of these methods is as follows.

Hidden neurons (H), ELM for the training set (N), and activation function $f(x)$ can be described as follows:

$$e_j = \sum_{i=1}^{H} \alpha_i f\left(w_i, c_i, x_j\right) j = 1 \cdots N. \quad (3)$$

α_i and ω_i are H-output layers and weight vectors (WVs) of the connecting input HLs (input Ws), respectively. x_j indicates input variables. C_i indicates the H bias for the ith H neuron, and e_j is the output of ELM for multiple data points (j). The process of generating input Ws is random and is based on consecutive distribution. Via a linear function, the output and result of Ws are calculated which are as follows:

$$\beta = A\dagger Y. \quad (4)$$

A shows the output of the HL matrix (equation (5)), A † indicates the inverse of A when using the Moore-Penrose method, and Y represents the values that ELM tries to reach. The compact and simplified form of equation (4) is $A\alpha = Y$. A is the matrix of HL in the neural network (NN) and Y is the vectors of the output variable. The three matrixes, A, α, and Y, can be represented as follows:

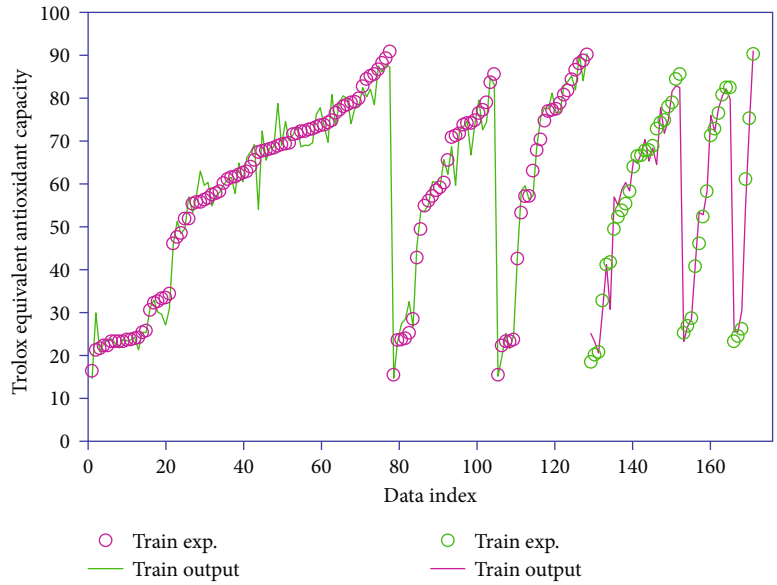


Figure 1: Simultaneous viewing of real and corresponding modeled data.

$$A = \begin{bmatrix} h(x_i) \\ \vdots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} f(w_1, c_1, x_1) & \cdots & f(w_H, c_H, x_1) \\ \vdots & \cdots & \vdots \\ f(w_1, c_1, x_j) & \cdots & f(w_H, c_H, x_j) \end{bmatrix}, \alpha = \begin{bmatrix} \alpha_1^T \\ \vdots \\ \alpha_H^T \end{bmatrix}, Y = \begin{bmatrix} y_1^T \\ \vdots \\ y_N^T \end{bmatrix}, \tag{5}$$

in which $h(x)$ represents the feature mapping of the HL. Matrix A strongly assigns the result of the ELM algorithm. One of the traditional solutions is using NNs in the HLs in which this type of solution uses a gradient descent algorithm as represented in [31]. In order to solve ELM, the kernel function is applied, and on the other hand, for solving the kernel matrix, feature mapping is used as can be seen in the following [30]:

$$k(x_i, x_j) = h(x_i).h(x_j). \tag{6}$$

In this article, we attempt to comprehend the efficiency of the kernel on the Mr prediction via KELM application and integrating ELM method with PSO for designing a fresh model for predicting Mr. In the point of view of the learning rate, predictive performance, and generalization capability, ELM has better performance in comparison with normal NNs. ELM via the Moore-Penrose generalized reverse method distinguishes the Ws of the output and input layers and produces some random values for H biases and input Ws [32, 33]. The general structure of a NN known as SLFN (single HL feed-forward NN) included the output and input layer neurons, m and n, respectively, and also HL neurons. For instance, suppose $\{X_i, Y_i\}$ is a training dataset, then it can be understood that the input dataset is $X_i = [X_{i1}, X_{i2}, \cdots, X_{\text{in}}]$ and the output dataset is $Y_i = [Y_{i1}, Y_{i2}, \cdots, Y_{im}]$ for $i = 1, 2, \cdots, n$. m is the number of training samples.

3. Data Gathering

There are 172 data points in the database utilized in the present study [34], with two train (129 data points (about 75%)) and test (43 data points (about 25%)) datasets for training and testing the efficiency of proposed models, respectively. To boost the efficiency of the study models, the data points were normalized between −1 and +1.

4. Results and Discussion

The efficiency analysis needs to be carried out for evaluating the capability of the model. Accordingly, various statistical analyses were carried out between the actual values and the model outputs, including standard deviations (STD), mean relative errors (MRE), root mean square error (RMSE), mean squared error (MSE), and R-squared (R^2) to evaluate the capability of the study model [35–38].

Figure 1 represents the actual values versus model outputs for the output data at the train and test phases. The target is accurately estimated by the study model with a decent agreement between the real information and model yields, highlighting their capability in output prediction.

Also, the model results from regression analysis represented in Figure 2 at the train and test phases. Based on related literature, the R^2 value is an eminent statistic indicating the model output-actual value relationship. The basic objective was to conduct a comparative analysis between model yields and real values. The accuracy of the model's accuracy is improved when the fitted line approaches the bisector line [39, 40]. A remarkable linear correlation is achieved between the model outputs and actual values for $R^2 = 1$, which gets weaker when the R^2 value approaches to

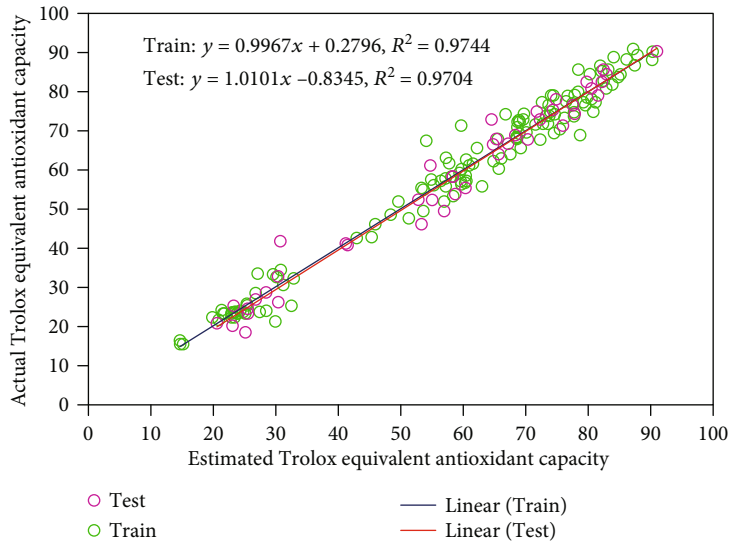


Figure 2: Regression analysis to evaluate the accuracy of the proposed model.

Table 1: Determining the values of different statistical parameters for the model in different phases.

Phase	R^2	MRE (%)	MSE	RMSE	STD
Train	0.974	5.06	12.37	3.52	2.38
Test	0.970	5.90	13.65	3.70	2.60
Total	0.973	5.27	12.69	3.56	2.43

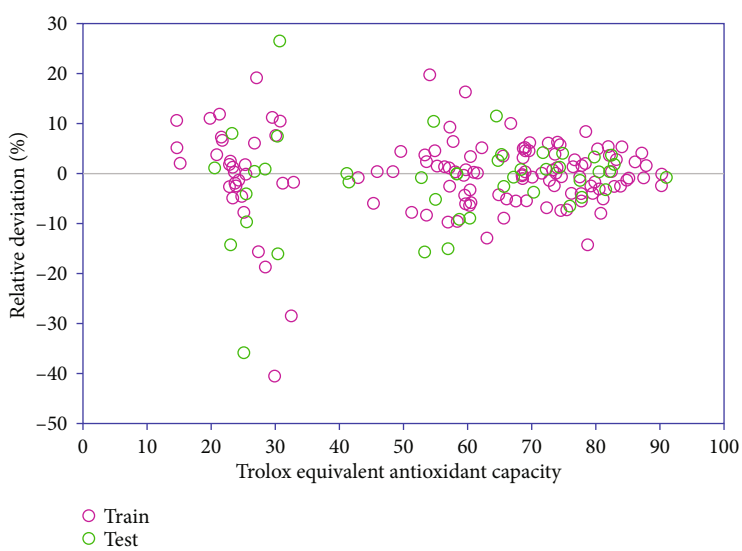


Figure 3: Determining relative deviation values to evaluate the accuracy of the model in predicting the target data.

zero [41, 42]. Accuracy is represented by a close-fitting of data points around the 45° line for the prediction models. As shown in this figure, this model shows a high ability to predict target values in different phases.

Table 1 presents the results from statistical analyses of the study model based on the RMSE, MSE, STD, MRE, and R^2 parameters [43, 44].

In a study with similar input data, Buciński et al. predicted TAEC values using the artificial neural network (ANN) method [34]. Their model showed an accuracy of $R^2 = 0.931$ in estimating the output data in the testing phase, which was weaker than the model proposed in this paper.

Furthermore, Figure 3 represents the absolute relative deviation between the actual values and model yields of output anticipated utilizing the examination model.

William's plot was utilized for determining the outliers of the model [45, 46]. Figure 4 represents the standardized residuals versus hat values. This figure clearly shows three

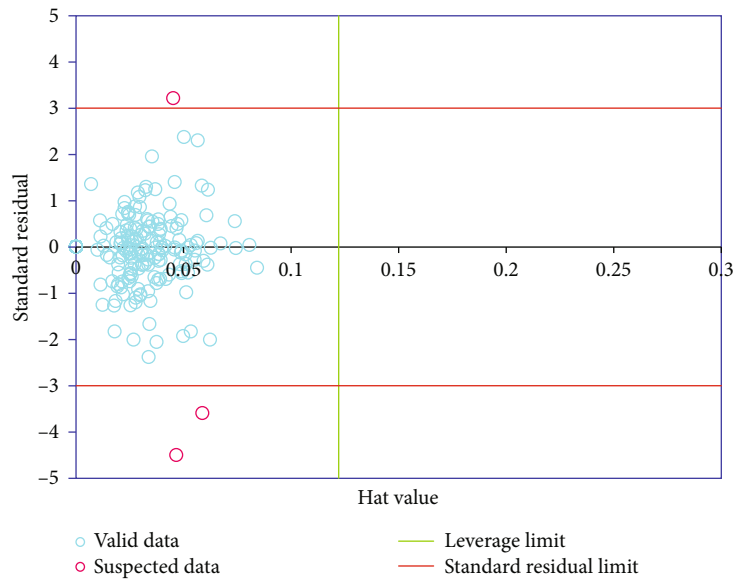


Figure 4: Detection of suspicious points using William's plot.

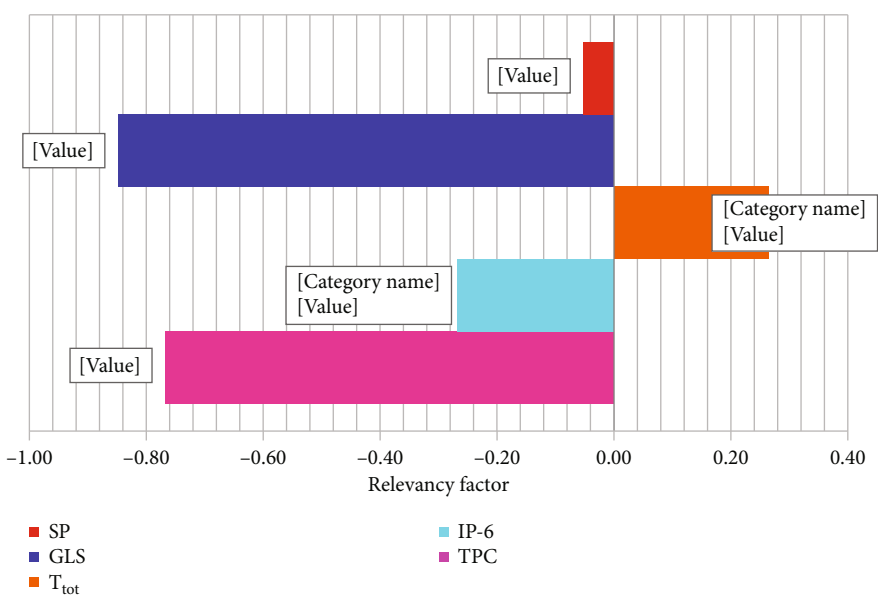


Figure 5: Sensitivity analysis on various input parameters of the model.

limited boundaries: leverage limit, upper limit, and down suspected limit [47]. Outliers are data with higher standardized residual values > 3 or <−3, and the data with hat > hat ∗ (referred to as the warning leverage value) are beyond the applicability domain of the study model [48]. As can be seen from this figure, among all the data points, only three suspicious points are seen.

Finally, sensitivity analysis was used to determine the effect of different input parameters on the target parameter. More details about this analysis are given elsewhere [49, 50]. According to Figure 5, it was found that T_{tot} has the most direct effect on the target parameter, which corresponds to the relevancy factor (r) equal to +0.26, while other input parameters showed an inverse effect on the target parameter so that GLS showed the most negative effect with r equal to -0.85.

5. Conclusion

This study was aimed at seeing how well a statistical learning-based model could predict the antioxidant capacity of cruciferous sprouts. To this end, the PSO was implemented in the ELM model. When it came to setting the tuning parameters, the PSO algorithm showed good performance. Estimates were found to be quite accurate when compared to actual data points. The efficiency of the proposed techniques was verified by an excellent agreement achieved between the model outputs and the actual values in assessing the model

during the train and test phases, as demonstrated by results from statistical analyses. The models' accuracy was confirmed as predicted by a comparison which was made between the proposed models' outcomes and another reported correlation. The proposed strategy to predict the antioxidant capacity of cruciferous sprouts is user-friendly so that they can be considered a useful tool for researchers, particularly in related fields, unlike the sophisticated mathematical techniques developed for this output prediction.

Data Availability

The data used to support the findings of this study are provided within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. Anwar, G. Hussain, and I. Mustafa, "Antioxidants from natural sources," *Antioxidants in Foods and its Applications*, 2018.

[2] S. C. Lourenço, M. Moldão-Martins, and V. D. Alves, "Antioxidants of natural plant origins: from sources to food industry applications," *Molecules*, vol. 24, no. 22, p. 4132, 2019.

[3] A. E. Hagerman, K. M. Riedl, G. A. Jones et al., "High molecular weight plant polyphenolics (tannins) as biological antioxidants," *Journal of Agricultural and Food Chemistry*, vol. 46, no. 5, pp. 1887–1892, 1998.

[4] İ. Gulcin, "Antioxidants and antioxidant methods: an updated overview," *Archives of Toxicology*, vol. 94, no. 3, pp. 651–715, 2020.

[5] N. J. Temple, "Antioxidants and disease: more questions than answers," *Nutrition Research*, vol. 20, no. 3, pp. 449–459, 2000.

[6] H. Zielinski, "Low molecular weight in the cereal grains-a review," *Polish Journal of Food and Nutrition Sciences (Poland)*, vol. 52, pp. 3–6, 2002.

[7] S. Asgary, A. Rastqar, and M. Keshvari, "Functional food and cardiovascular disease prevention and treatment: a review," *Journal of the American College of Nutrition*, vol. 37, no. 5, pp. 429–455, 2018.

[8] J. W. Ballway and B.-J. Song, "Translational approaches with antioxidant phytochemicals against alcohol-mediated oxidative stress, gut dysbiosis, intestinal barrier dysfunction and fatty liver disease," *Antioxidants*, vol. 10, no. 3, p. 384, 2021.

[9] G. Block, B. Patterson, and A. Subar, "Fruit, vegetables, and cancer prevention: a review of the epidemiological evidence," *Nutrition and Cancer*, vol. 18, no. 1, pp. 1–29, 1992.

[10] K. A. Steinmetz and J. D. Potter, "Vegetables, fruit, and cancer prevention: a review," *Journal of the American Dietetic Association*, vol. 96, no. 10, pp. 1027–1039, 1996.

[11] N. Chhikara, R. Kaur, S. Jaglan, P. Sharma, Y. Gat, and A. Panghal, "Bioactive compounds and pharmacological and food applications of Syzygium cumini–a review," *Food & Function*, vol. 9, no. 12, pp. 6096–6115, 2018.

[12] A. Ghiselli, M. Serafini, F. Natella, and C. Scaccini, "Total antioxidant capacity as a tool to assess redox status: critical view and experimental data," *Free Radical Biology and Medicine*, vol. 29, no. 11, pp. 1106–1114, 2000.

[13] H. Zielinski, "Peroxyl radical-trapping capacity of germinated legume seeds," *Food/Nahrung*, vol. 46, no. 2, pp. 100–104, 2002.

[14] S. Ortega-Requena, S. Rebouillat, and F. Pla, "Paving the highway to sustainable, value adding open-innovation integrating bigger-data challenges: three examples from bio-ingredients to robust durable applications of electrochemical impacts," *Journal of Biomaterials and Nanobiotechnology*, vol. 9, no. 2, pp. 117–188, 2018.

[15] G. F. Biggi, *Investigating the molecular genetic basis of antioxidants in Lactuca sativa for the enhancement of its nutritional qualities*, University of Southampton, 2010.

[16] K. O. Chu, *The study of feasibility of green tea treatment on fetus: from chemistry to treatment*, The Chinese University of Hong Kong, Hong Kong, 2005.

[17] N. Miller and C. Rice-Evans, "Spectrophotometric determination of antioxidant activity," *Redox Report*, vol. 2, no. 3, pp. 161–171, 1996.

[18] C. Rice-Evans and N. J. Miller, "Antioxidants the case for fruit and vegetables in the diet," *British Food Journal*, vol. 97, no. 9, pp. 35–40, 1995.

[19] N. Salah, N. J. Miller, G. Paganga, L. Tijburg, G. P. Bolwell, and C. Riceevans, "Polyphenolic Flavanols as Scavengers of Aqueous Phase Radicals and as Chain- Breaking Antioxidants," *Archives of Biochemistry and Biophysics*, vol. 322, no. 2, pp. 339–346, 1995.

[20] H. Wang, G. Cao, and R. L. Prior, "Total antioxidant capacity of fruits," *Journal of Agricultural and Food Chemistry*, vol. 44, no. 3, pp. 701–705, 1996.

[21] H. Zieliński and H. Kozłowska, "Antioxidant activity and total phenolics in selected cereal grains and their different morphological fractions," *Journal of Agricultural and Food Chemistry*, vol. 48, no. 6, pp. 2008–2016, 2000.

[22] R. Apak, M. Özyürek, K. Güçlü, and E. Çapanoğlu, "Antioxidant activity/capacity measurement. 1. Classification, physicochemical principles, mechanisms, and electron transfer (ET)-based assays," *Journal of agricultural and food chemistry*, vol. 64, no. 5, pp. 997–1027, 2016.

[23] K. Guclu, M. Altun, M. Ozyurek, E. Saliha, and R. A. Karademir, "Antioxidant capacity of fresh, sun- and sulphited-dried Malatya apricot (Prunus armeniaca) assayed by CUPRAC, ABTS/TEAC and folin methods," *International Journal of Food Science & Technology*, vol. 41, no. s1, pp. 76–85, 2006.

[24] H. Zielinski, A. Bucinski, and H. Kozlowska, "Monitoring of the vitamin C content in germinating cruciferae seeds by HPLC," *Polish Journal of Food and Nutrition Sciences*, vol. 11, pp. 142–146, 2002.

[25] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory In MHS'95," in *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, Ieee, 1995.

[26] P. Wilson and H. A. Mantooth, *Model-based engineering for complex electronic systems*, Newnes, 2013.

[27] A. M. Sharaf and A. A. Elgammal, *Novel AI-based soft computing applications in motor drives*, in Power Electronics Handbook, Elsevier, 2018.

[28] F. Han, H.-F. Yao, and Q.-H. Ling, "An improved evolutionary extreme learning machine based on particle swarm optimization," *Neurocomputing*, vol. 116, pp. 87–93, 2013.

[29] H. Guo, B. Li, W. Li, F. Qiao, X. Rong, and Y. Li, "Local coupled extreme learning machine based on particle swarm optimization," *Algorithms*, vol. 11, no. 11, p. 174, 2018.

[30] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2011.

[31] M. Pal and S. Deswal, "Extreme learning machine based modeling of resilient modulus of subgrade soils," *Geotechnical and Geological Engineering*, vol. 32, no. 2, pp. 287–296, 2014.

[32] J. Cao, Z. Lin, and G.-B. Huang, "Self-adaptive evolutionary extreme learning machine," *Neural Processing Letters*, vol. 36, no. 3, pp. 285–305, 2012.

[33] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[34] A. Buciński, H. Zieliński, and H. Kozłowska, "Artificial neural networks for prediction of antioxidant capacity of cruciferous sprouts," *Trends in Food Science & Technology*, vol. 15, no. 3-4, pp. 161–169, 2004.

[35] M. Mahdaviara, A. Rostami, F. Keivanimehr, and K. Shahbazi, "Accurate determination of permeability in carbonate reservoirs using Gaussian process regression," *Journal of Petroleum Science and Engineering*, vol. 196, p. 107807, 2021.

[36] E. Khamehchi and A. Bemani, "Prediction of pressure in different two-phase flow conditions: machine learning applications," *Measurement*, vol. 173, p. 108665, 2021.

[37] X. Zhou, F. Zhou, and M. Naseri, "An insight into the estimation of frost thermal conductivity on parallel surface channels using kernel based GPR strategy," *Scientific Reports*, vol. 11, no. 1, 2021.

[38] M. R. Kaloop, A. Bardhan, N. Kardani, P. Samui, J. W. Hu, and A. Ramzy, "Novel application of adaptive swarm intelligence techniques coupled with adaptive network-based fuzzy inference system in predicting photovoltaic power," *Renewable and Sustainable Energy Reviews*, vol. 148, p. 111315, 2021.

[39] A. Hemmati-Sarapardeh, E. Mohagheghian, M. Fathinasab, and A. H. Mohammadi, "Determination of minimum miscibility pressure in N_2-crude oil system: A robust compositional model," *Fuel*, vol. 182, pp. 402–410, 2016.

[40] H. Mokarizadeh, S. Atashrouz, H. Mirshekar, A. Hemmati-Sarapardeh, and A. M. Pour, "Comparison of LSSVM model results with artificial neural network model for determination of the solubility of SO_2 in ionic liquids," *Journal of Molecular Liquids*, vol. 304, p. 112771, 2020.

[41] A. Ghanbari, M. N. Kardani, A. M. Goodarzi, M. J. Lariche, and A. Baghban, "Neural computing approach for estimation of natural gas dew point temperature in glycol dehydration plant," *International Journal of Ambient Energy*, vol. 41, no. 7, pp. 775–782, 2020.

[42] N. Kardani, A. Bardhan, D. Kim, P. Samui, and A. Zhou, "Modelling the energy performance of residential buildings using advanced computational frameworks based on RVM, GMDH, ANFIS-BBO and ANFIS-IPSO," *Journal of Building Engineering*, vol. 35, 2021.

[43] M. Ahmadi, M. Ghazvini, A. Baghban et al., "Soft computing approaches for thermal conductivity estimation of CNT/water nanofluid," *Revue des Composites et des Matériaux Avancés*, vol. 29, no. 2, 2019.

[44] X. Zhou, F. Zhou, and M. Naseri, "An insight into the estimation of frost thermal conductivity on parallel surface channels using kernel based GPR strategy," *Scientific Reports*, vol. 11, no. 1, 2021.

[45] N. Nabipour, A. Mosavi, A. Baghban, S. Shamshirband, and I. Felde, "Extreme learning machine-based model for solubility estimation of hydrocarbon gases in electrolyte solutions," *PRO*, vol. 8, no. 1, 2020.

[46] A. Baghban, J. Sasanipour, F. Pourfayaz et al., "Towards experimental and modeling study of heat transfer performance of $_{\text{water-}}$ SiO2nanofluid in quadrangular cross-section channels," *Engineering Applications of Computational Fluid Mechanics*, vol. 13, no. 1, pp. 453–469, 2019.

[47] S. R. Moosavi, B. Vaferi, and D. A. Wood, "Auto-characterization of naturally fractured reservoirs drilled by horizontal well using multi-output least squares support vector regression," *Arabian Journal of Geosciences*, vol. 14, no. 7, 2021.

[48] R. Razavi, A. Bemani, A. Baghban, and A. H. Mohammadi, "Modeling ofCO2absorption capabilities of amino acid solutions using a computational scheme," *Environmental Progress & Sustainable Energy*, vol. 39, no. 6, 2020.

[49] M. H. Ahmadi, A. Baghban, M. Sadeghzadeh, M. Hadipoor, and M. Ghazvini, "Evolving connectionist approaches to compute thermal conductivity of TiO2/water nanofluid," *Physica A: Statistical Mechanics and its Applications*, vol. 540, p. 122489, 2020.

[50] A. Bemani, A. Baghban, S. Shamshirband, A. Mosavi, P. Csiba, and A. R. Varkonyi-Koczy, "Applying ANN, ANFIS, and LSSVM models for estimation of acid solvent solubility in supercritical CO_2," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1175–1204, 2020.

Hindawi
BioMed Research International
Volume 2021, Article ID 5519436, 7 pages
https://doi.org/10.1155/2021/5519436

Research Article

Ability of Procalcitonin and C-Reactive Protein for Discriminating between Bacterial and Enteroviral Meningitis in Children Using Decision Tree

Dmitriy Babenko,[1,2] **Aliya Seidullayeva**,[3,4] **Dinagul Bayesheva**,[3,4] **Bayan Turdalina**,[3,4] **Baurzhan Omarkulov**,[1] **Aigul Almabayeva**,[5] **Marina Zhanaliyeva**,[5] **Almagul Kushugulova**,[6] **and Samat Kozhakhmetov**[2,6]

[1]*Karagandy Medical University, Karagandy, Kazakhstan*
[2]*Innovative Center ArtScience, Nur-Sultan, Kazakhstan*
[3]*Department of Paediatric Infectious Diseases, Astana Medical University, Nur-Sultan City, Kazakhstan*
[4]*Department of Paediatric Infectious Diseases, №3 Multidisciplinary City Children's Hospital, Nur-Sultan City, Kazakhstan*
[5]*NSC Astana Medical University, Nur-Sultan City, Kazakhstan*
[6]*Centre for Life Sciences, National Laboratory Astana, Nazarbayev University, Nur-Sultan City, Kazakhstan*

Correspondence should be addressed to Samat Kozhakhmetov; sskozhakhmetov@gmail.com

Received 22 January 2021; Accepted 24 July 2021; Published 3 August 2021

Academic Editor: Alireza Baghban

Bacterial meningitis (BM) is a public health burden in developing countries, including Central Asia. This disease is characterized by a high mortality rate and serious neurological complications. Delay with the start of adequate therapy is associated with an increase in mortality for patients with acute bacterial meningitis. Cerebrospinal fluid culture, as a gold standard in bacterial meningitis diagnosis, is time-consuming with modest sensitivity, and this is unsuitable for timely decision-making. It has been shown that bacterial meningitis differentiation from viral meningitis could be done through different parameters such as clinical signs and symptoms, laboratory values, such as PCR, including blood and cerebrospinal fluid (CSF) analysis. In this study, we proposed the method for distinguishing the bacterial form of meningitis from enteroviral one. The method is based on the machine learning process deriving making decision rules. The proposed fast-and-frugal trees (FFTree) decision tree approach showed an ability to determine procalcitonin and C-reactive protein (CRP) with cut-off values for distinguishing between bacterial and enteroviral meningitis (EVM) in children. Such a method demonstrated 100% sensitivity, 96% specificity, and 98% accuracy in the differentiation of all cases of bacterial meningitis in this study. These findings and proposed method may be useful for clinicians to facilitate the decision-making process and optimize the diagnostics of meningitis.

1. Introduction

Meningitis is a life-threatening inflammatory disease of the brain and spinal cord, mostly caused by bacterial, viral, and fungal infection [1–3]. Meningococcal infection has been a big threat to the globe and exists as a sporadic, hypersporadic, and epidemic disease. In 2012, an estimated 1.2 million cases of meningococcal infection per year were reported, with ~135,000 deaths worldwide [4]. The average annual incidence of meningococcal infection in Kazakhstan for the last decade is 0.83/100 000 with a peak in 2015 (2.42/100 000) [5].

Bacterial meningitis as a more serious form of meningitis is caused by pyogenic bacteria, such as *S. pneumoniae*, *N. meningitidis*, and *H. influenzae* [6]. Viruses are the most common cause of aseptic meningitis, primarily enteroviruses, together with numerous nonviral and noninfectious disorders [7, 8].

Although bacterial meningitis has a lower incidence rate than viral/aseptic meningitis [9, 10], prompt correct diagnosis

Table 1: Demographic data of the studied population with meningitis.

	[ALL] $N = 269$	EVM $N = 146$	BM $N = 123$	p.overall
Age (in months)	48.0 [17.0; 96.0]	82.5 [40.0; 124]	23.0 [9.00; 50.0]	<0.001
Age group:				<0.001
≤1 y	52 (19.3%)	11 (7.53%)	41 (33.3%)	
>1-3 y	61 (22.7%)	23 (15.8%)	38 (30.9%)	
>3-5 y	46 (17.1%)	27 (18.5%)	19 (15.4%)	
>5-7 y	33 (12.3%)	17 (11.6%)	16 (13.0%)	
>7-10 y	36 (13.4%)	30 (20.5%)	6 (4.88%)	
>10 y	41 (15.2%)	38 (26.0%)	3 (2.44%)	
Gender:				0.622
Female	117 (43.5%)	66 (45.2%)	51 (41.5%)	
Male	152 (56.5%)	80 (54.8%)	72 (58.5%)	

and adequate treatment are necessary due to its hazardous nature [11]. Delay in the start of proper therapy introduces the potential for increased morbidity and mortality if the patient does indeed have acute bacterial meningitis [12].

Diagnosis of bacterial meningitis is based on a positive culture of cerebrospinal fluid (or detecting etiological agent by polymerase chain reaction—PCR), along with typical clinical symptoms (fever, headache, and neck stiffness). CSF culture is highly specific but lacks sensitivity, especially when antimicrobials have been given as well as the time needed until results appear [13]. In this case, PCR analysis can play a diagnostic role, but as the direct culture of cerebrospinal fluid, it takes some time. It should also be noted that not every clinic has the appropriate equipment and capabilities for conducting PCR analysis in CSF, especially in developing countries [14].

Distinguishing bacterial meningitis is often difficult [15] and therefore highly accurate decision support tools are necessary to guide decision making and limit unnecessary hospital admissions and prolonged antibiotic use.

Our study is aimed at assessing the role of clinical presentations, serum, and CSF profiles to distinguish BM and EVM in children.

2. Materials and Methods

2.1. Subjects. Recruiting patients for the study was carried out in the Department of Reanimation and Intensive Care, Infection Department No. 1, Multidisciplinary City Children's Hospital No. 3, Nur-Sultan City (Kazakhstan). The study covers the period between 2017 and 2019.

Inclusion criteria were as follows: children from 1 month to 17 years old, both sexes, presence of bacterial antigen, bacterial or viral nucleic acids identified in CSF in blood serum, results of a positive culture study for pathogens, and the presence of clinical signs of meningitis.

Exclusion criteria were as follows: children diagnosed with tuberculous meningitis, benign and malignant brain tumors, and children over 17 years old. The study did not include samples with meningitis of nonenteroviral etiology and combined forms of meningitis, bacteremia (meningococcemia).

The study has been approved by the Local Ethics Committee of the National Laboratory Astana (NLA) at Nazarbayev University (by 22nd of September 2017. Approval No. 20).

A total of 269 patients were recruited and divided into 6 groups from 1 month to 10 years and more (Table 1).

2.2. Physical Examination. Clinical symptoms such as temperature, vomiting, impaired consciousness, headache, pallor of the skin, rash, tension, and bulging of the fontanel in children under one-year-old, stiff neck muscles, Lesage, Brudzinski, and Kernig symptoms were determined.

2.3. Laboratory Examination. The number of white blood cells (WBC), neutrophils, and level of protein in the CSF were determined using the Cobas Integra 400 plus analyzer (Roche, EU). The level of glucose in CSF was determined using the ABL800 Flex Analyzer (Radiometer Medical ApS, Denmark). The levels of haemoglobin, erythrocyte sedimentation rate (ESR), white blood cells, count, neutrophils, and platelets in blood were determined using the hematologic analyzer Sysmex XP-300 (Sysmex). The levels of CRP and procalcitonin were analysed using fluorescence immunechromatographic system Finecare FIA Meter (Guangzhou Wondfo Biotech Co. Ltd., China.).

Samples of CSF or blood were placed on the surface of the culture medium for cultivation and identification on the "chocolate" agar (based on trypticase soy agar with the addition of defibrinated ram blood). Then, the samples were incubated at a temperature of 37.0°C in an atmosphere of 5.0% CO_2 for 24-48 hours. The presence of etiological agents of viral meningitis was determined by commercial PCR kits according to the manufacturer's protocol.

2.4. Statistical Analysis. Statistical analysis was carried out using SigmaPlot 11.0 software (Systat Software Inc., USA) with the following conditions:

(1) *Quantitative Data with a Normal Distribution*. Student t-test for two groups or analysis of variance (ANOVA), when the number of groups is more than two

Table 2: Underlying and associated conditions in bacterial and enteroviral meningitis groups in the studied population.

	[ALL] $N = 269$	EVM $N = 146$	BM $N = 123$	p.overall
Temperature (°C):	38.5 [37.8; 39.0]	37.9 [37.4; 38.5]	39.0 [38.6; 39.5]	<0.001
Vomiting:	262 (97.4%)	144 (98.6%)	118 (95.9%)	0.252
Headache:	146 (54.3%)	44 (30.1%)	102 (82.9%)	<0.001
Bulging fontanelle (for ≤18 months olds):	71 out 72 (98.6%)	19 out 20 (95%)	52 out 52 (100%)	0.278
Neck rigidity:	266 (98.9%)	146 (100%)	120 (97.6%)	0.094
Kernig's sign:	168 (62.5%)	67 (45.9%)	101 (82.1%)	<0.001
Brudzinski's sign:	61 (22.7%)	6 (4.11%)	55 (44.7%)	<0.001
Loss of consciousness:	15 (5.58%)	0 (0.00%)	15 (12.2%)	<0.001
Drowsiness:	36 (13.4%)	5 (3.42%)	31 (25.2%)	<0.001
Spasms:	27 (10.0%)	0 (0.00%)	27 (22.0%)	<0.001

Table 3: Laboratory findings of blood and cerebrospinal fluid in the studied population.

	[ALL] $N = 269$	EVM $N = 146$	BM $N = 123$	p.overall
Glucose in CSF: [2.3-3.9] mmol/L	2.80 [1.60; 3.80]	3.70 [3.10; 4.80]	1.50 [0.57; 1.90]	<0.001
Haemoglobin in blood: [110-140] g/L	119 (16.7)	125 (15.4)	111 (15.2)	<0.001
Protein in CSF: [0.12-0.45] g/L	0.50 [0.20; 1.30]	0.20 [0.10; 0.40]	1.50 [0.60; 2.10]	<0.001
CRP in blood: [≤10] mg/L	22.8 [4.80; 110]	5.00 [2.12; 12.0]	118 [43.5; 196]	<0.001
ESR in blood: [0-10] mm/H	15.0 [10.0; 22.0]	12.0 [7.00; 17.0]	18.0 [14.5; 30.0]	<0.001
Procalcitonin in blood: [≤0.05] ng/mL	0.05 [0.02; 3.30]	0.02 [0.01; 0.03]	3.50 [2.15; 5.15]	<0.001
WBC in CSF: $[\leq 30] * 10^9/L$	380 [110; 1300]	126 [78.2; 233]	1455 [815; 4250]	<0.001
WBC in blood: $[4.5 - 10.5] * 10^9/L$	13.1 [8.80; 18.0]	9.25 [7.60; 12.2]	18.0 [15.0; 23.2]	<0.001
Neutrophils in CSF: $[\leq 10] * 10^6/L$	72.0 [19.0; 90.0]	21.5 [10.0; 61.5]	88.0 [75.0; 90.0]	<0.001
Neutrophils in blood: [42-72%]	79.8 [69.9; 87.1]	74.0 [58.2; 84.0]	85.0 [76.3; 90.0]	<0.001
Platelets in blood: $[180 - 320] * 10^9/L$	233 [195; 314]	228 [190; 288]	256 [209; 321]	0.020

(2) *Quantitative Data with Abnormal Distribution*. Non-parametric Kruskal-Wallis test and Mann–Whitney test for independent groups; and Wilcoxon Matched Pairs Test for dependent groups

(3) *Categorical Data*. Chi-square or Fisher's tests if necessary (when the expected frequency is less than 5 in one of the cells)

Shapiro-Wilk test was employed for the evaluation of the distribution of data (normality test). $p < 0.05$ was considered statistically significant for all analyses.

2.5. Modelling. Fast-and-frugal trees (FFTs) as a supervised learning algorithm described in Phillips, Neth, Woike, and Gaissmaier [16] and implemented in FFTrees R package was used to predict a binary criterion BM and EVM. Before the machine training, we split the entire dataset into training (80%) and testing (20%) subsets. An optimal cut-off point was calculated according to the highest accuracy (minimal false-negative and false-positive results). The area (AUC, area under curve) under the receiver operating characteristic curve (ROC) was used to check the prognostic value of a particular parameter. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated for the given cut-off values for predicting bacterial meningitis.

3. Results

A total of 269 children (117 females and 152 males) were included in this study. The median and IQR age of the participants were 48.0 [17.0-96.0] months old. Children with EVM, that had a median age of 82.5 [40.0; 124] months old, were older compared with the BM group (median age 23.0 [9.00; 50.0]; $p < 0.001$) (Table 1). The highest rate of meningitis was found out among children aged up to 1-year old in a group with BM. In contrast, the rate of EVM was relatively low (7.53%) in a group up to 1-year children, while in groups > 1 year, EVM incidences were higher (on average18.48%) and reached 26% in the group > 10 years.

Among the studied patients, bacterial meningitis represented 45.7% (123 patients) compared to 54.3% (146 patients) nonbacterial (enteroviral) meningitis. Among bacterial meningitis, 95 were due to *N. meningitidis*, 25 cases *S. pneumoniae*,

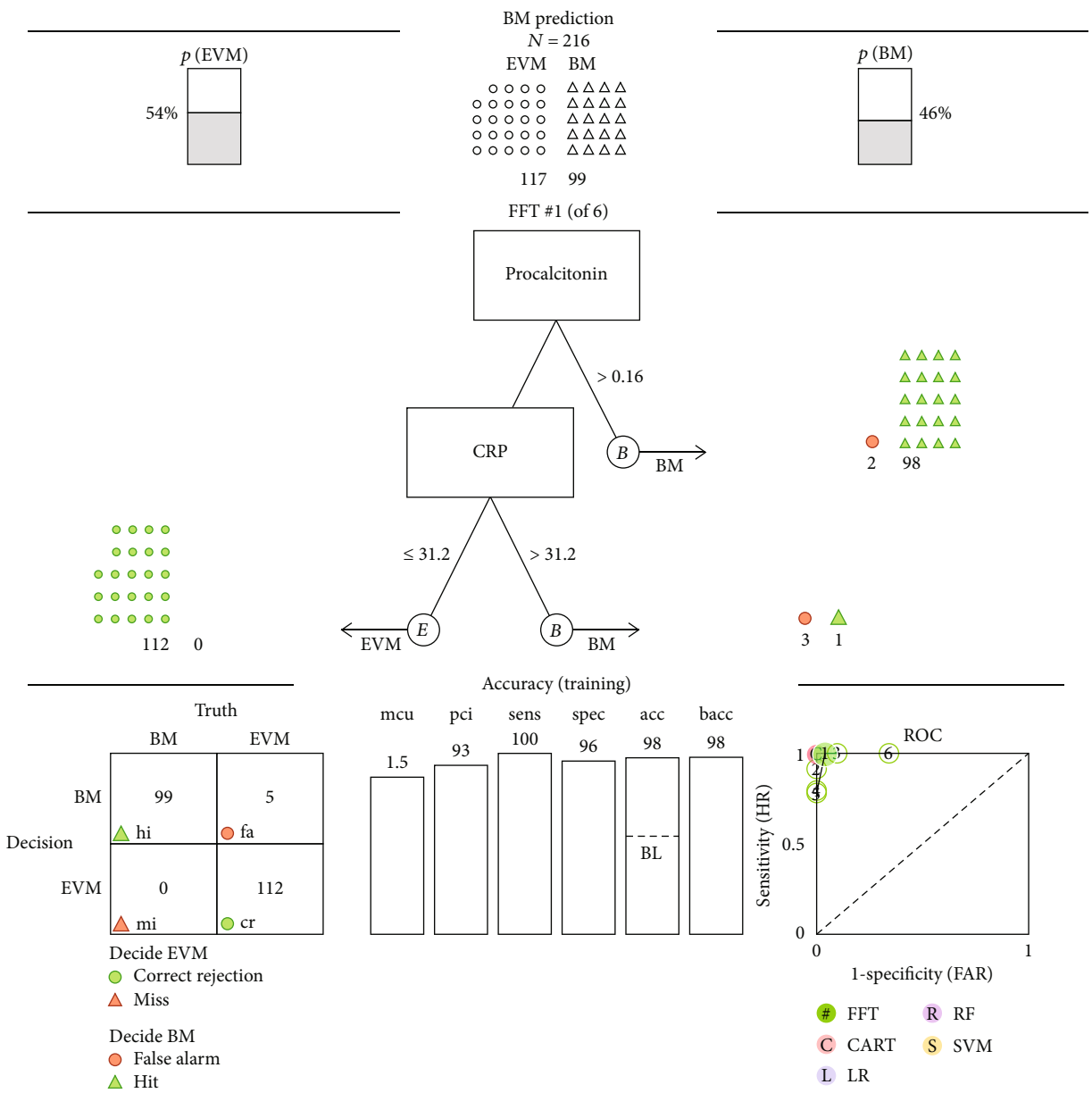


Figure 1: A fast-and-frugal tree (FFT) for classifying patients as either with BM or with EVM based on up to two parameters. It should be interpreted as if a patient's procalcitonin value is more than 0.16 ng/mL, classify as BM. If not, check the CRP value. If this is less or equal 31.2 mg/L, classify as EVM, otherwise, classify as BM.

2 cases *S. agalactiae*, and 1 case *S. aureus*. Among 146 viral/aseptic cases, all 146 cases were caused by enterovirus.

The most common presenting symptom in children with BM was the high temperature (39°C vs. 37.9°C, $p < 0.001$) followed by headache (82.9%) and Kernig's sign (82.1%). The occurrence of Brudzinski's sign (44.7%), drowsiness (25.2%), spasms (22%), and loss of consciousness (12.2%) was significantly higher in the BM group compared with the EVM group ($p < 0.001$). Vomiting, bulging fontanelle, and neck rigidity had no difference among these groups. The clinical features of BM and EVM groups are presented in Table 2.

Table 3 represents the comparison of the laboratory results of blood and CSF between the two groups. Blood and CSF laboratory testing data showed significantly increased levels of proteins, CRP, procalcitonin, WBC, neutrophils, and platelets in the blood of children with BM ($p < 0.001$). WBC and neutrophils in CFS were also significantly higher in the BM group ($p < 0.001$). The blood glucose level was 1.50 [0.57; 1.90] mmol/L in the BM group, which is significantly lower than that for the EVM group (3.70 [3.10; 4.80] mmol/L; $p < 0.001$). Haemoglobin was also significantly lower in the BM group in comparison with the EVM group ($p < 0.001$).

Fast-and-frugal trees (FFTs) algorithm was performed for providing efficient and accurate decisions in the prediction of bacterial meningitis. All data, such as demographic variables (gender and age), clinical data, and laboratory results, were included in training FFTs to get the best-trained algorithm based on the highest sensitivity. The results

Table 4: Validation of clinical and laboratory parameters in bacterial meningitis prediction.

Parameters	Threshold	Direction	*N*	Sens	Spec	PPV	NPV	Acc
Procalcitonin in blood (ng/mL)	0.16	>	216	0.990	0.983	0.980	0.991	0.986
CRP in blood (mg/L)	31.2	>	216	0.939	0.974	0.969	0.950	0.958
Glucose in CSF (mmol/L)	2.2	<=	216	0.848	0.949	0.933	0.881	0.903
Neutrophils in CSF ($*10^6$/L)	64	>	216	0.980	0.726	0.752	0.977	0.843
WBC in CSF ($*10^9$/L)	513	>	216	0.818	0.872	0.844	0.850	0.847
Temperature (°C)	38.4	>	216	0.909	0.752	0.756	0.907	0.824
Protein in CSF (g/L)	0.3	>	216	0.949	0.709	0.734	0.943	0.819
WBC in blood ($*10^9$/L)	12.5	>	216	0.879	0.769	0.763	0.882	0.819
Headache	Yes	=	216	0.838	0.701	0.703	0.837	0.764
Age (months)	85	≤	216	0.949	0.470	0.603	0.917	0.690
Kernig's sign	Positive	=	216	0.838	0.538	0.606	0.797	0.676
Brudzinski's sign	Positive	=	216	0.424	0.949	0.875	0.661	0.708
Neutrophils in blood (%)	73	>	216	0.879	0.453	0.576	0.815	0.648
ESR in blood (mm/H)	11	>	216	0.859	0.462	0.574	0.794	0.644
Haemoglobin in blood (g/L)	113	≤	216	0.576	0.744	0.655	0.674	0.667
Spasms	Single	=	216	0.212	1.000	1.000	0.600	0.639
Drowsiness	Yes	=	216	0.232	0.957	0.821	0.596	0.625
Platelets in blood ($*10^9$/L)	252	>	216	0.545	0.607	0.540	0.612	0.579
Consciousness	Nonnormal	=	216	0.111	1.000	1.000	0.571	0.593
Gender	M	=	216	0.566	0.444	0.463	0.547	0.500

Abbreviations: sens: sensitivity; spec: specificity; ppv: positive predictive value; npv: negative predictive value; acc: accuracy.

of the best FFT with sensitivity, specificity, accuracy, and ROC curve are presented in Figure 1.

Two parameters, procalcitonin and C-reactive protein, appeared to have a good predictive value in bacterial meningitis. ROC curve was plotted with these two parameters having the best performing curve for diagnosing bacterial meningitis with a sensitivity of 100%, a specificity of 96%, and an accuracy of 98%.

4. Discussion

Meningococcal infection is one of the most severe infectious diseases of childhood [11]. The highest burden of disease is in Africa and Asia. However, the epidemics can occur in any part of the world. According to WHO reports, Asia has had some major epidemics of meningococcal disease in the last 30 years [17–20]. According to official statistical data, the peak of meningitis incidence in Kazakhstan was noted in 2015 (2.4 per 100 000). In subsequent years, there was a trend towards a decrease in 3.6 times by 2016, 7 times by 2017, and 4.6 times by 2018 compared to 2015 [5].

The early diagnostics of meningitis and differentiation of bacterial forms from aseptic (viral) ones plays a crucial role in the effective treatment of children. The analysis of CSF culture, as a gold standard in bacterial meningitis diagnostics, is a time-consuming process with modest sensitivity (70–85%). Moreover, in the case of antibiotic pretreatment, the sensitivity of CSF culture decreases by 20% [21]. In this regard, the CFS culture test is unsuitable for timely decision-making and effective diagnostics. The classic approach for the differentiation between bacterial and viral meningitis is based on the assessment of clinical signs and symptoms and laboratory tests (blood and CSF analysis) [11, 22, 23].

In contrast to the standard method, our study is focused on searching a novel method of meningitis diagnostics and differentiation. The proposed method is based on the validation of the factors such as demographic variables, clinical, and routine diagnostic tests, with the most discriminating power and sensitivity in differentiating bacterial meningitis from nonbacterial (enteroviral) meningitis.

Our trained FFTree model determined two parameters, procalcitonin and CRP in blood, based on which the differentiation of BM from EVM is effective. Definition of the decision tree is "if procalcitonin > 0.16 ng/mL, decide BM, If CRP < = 31.2 mg/L, decide EVM, otherwise, decide BM". This trained FFT model demonstrated an extremely high sensitivity (100% vs. 100%) and NPV (100% vs. 100%), and very high specificity (95.7% vs. 93%), PPV (95% vs. 92.3%), and accuracy (97.7% vs. 96.2%) on both trained and test datasets, respectively. The parameters of the predictive ability of each indicator are shown in Table 4.

The results demonstrated that the top three parameters (procalcitonin, CRP, and glucose level) were of the highest accuracy (98.6%, 95.8%, and 90.3%, respectively) in discriminating between bacterial and enteroviral meningitis. Previous studies also showed the same results with variation in cut-off values to distinguish bacterial and viral meningitis [24–31]. At the same time, the authors emphasized the importance of the heterogeneity of populations, techniques, and approaches of decision-making threshold for BM diagnosis markers.

In this paper, we addressed the task of distinguishing bacterial from viral meningitis in children through a machine learning-based approach deriving making decision rules. The proposed FFTree decision tree approach showed an ability to determine procalcitonin and CRP in blood with cut-off values for distinguishing between bacterial and enteroviral meningitis in children. It should be noted that the proposed method uses a minimally invasive procedure for taking material for diagnosis. Also, the method demonstrated 100% sensitivity, 96% specificity, and 98% accuracy in differentiation of all cases of bacterial meningitis in this study. These findings and proposed method may be useful for clinicians to facilitate the decision-making process and optimize the diagnostics of meningitis.

Data Availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

Each author declares that he/she has no commercial associations (e.g., consultancies, stock ownership, equity interest, and patent/licensing arrangement) that might pose a conflict of interest in connection with the submitted article.

Authors' Contributions

Dmitriy Babenko and Samat Kozhakhmetov conceived and designed the experiments. Samat Kozhakhmetov, Aliya Seidullayeva, Dinagul Bayesheva, Bayan Turdalina, and Baurzhan Omarkulov performed the experiments. Dmitriy Babenko, Samat Kozhakhmetov, and Almagul Kushugulova analysed the data. Aigul Almabayeva and Marina Zhanaliyeva contributed reagents/materials/analysis tools. Dmitriy Babenko, Almagul Kushugulova, and Samat Kozhakhmetov wrote the paper.

Acknowledgments

This study was supported by the grant of Science Committee, Ministry of Education and Science of the Republic of Kazakhstan: "Health programming, evolution of the infants' microbiome" (AP09259975) and "Development of Early Diagnostics and Preventive Measures of Hearing Impairment after Suffering from Bacterial Meningitis in Children" (AP05135091).

References

[1] T. J. Runde and J. W. Hafner, *Meningitis, Bacterial*, StatPearls, Treasure Island, FL, USA, 2020.

[2] R. M. Cantu and J. M Das, *Viral Meningitis*, StatPearls, Treasure Island, FL, USA, 2020.

[3] P. Pagliano, S. Esposito, T. Ascione, and A. M. Spera, "Burden of fungal meningitis," *Future Microbiology*, vol. 15, no. 7, pp. 469–472, 2020.

[4] N. G. Rouphael and D. S. Stephens, "Neisseria meningitidis: biology, microbiology, and epidemiology," *Methods in Molecular Biology*, vol. 799, pp. 1–20, 2012.

[5] A. Seidullayeva, D. Bayesheva, G. Zhaxylykova, B. Turdalina, A. Mukasheva, and S. Kozhakhmetov, "Sensorineural hearing loss after suffering bacterial meningitis in children," *Revista Latinoamericana de Hipertension*, vol. 14, no. 4, pp. 275–280, 2019.

[6] K. S. Adriani, D. van de Beek, M. C. Brouwer, L. Spanjaard, and J. de Gans, "Community-acquired recurrent bacterial meningitis in adults," *Clinical Infectious Diseases*, vol. 45, no. 5, pp. e46–e51, 2007.

[7] A. S. Shyshov, M. V. Bazarova, I. A. Blank et al., "Enterovirus infections and meningitis in children," *Zhurnal Nevrologii i Psikhiatrii Imeni S.S. Korsakova*, vol. 116, no. 2, pp. 9–15, 2016.

[8] C. W. Holmes, S. S. F. Koo, H. Osman et al., "Predominance of enterovirus B and echovirus 30 as cause of viral meningitis in a UK population," *Journal of Clinical Virology*, vol. 81, pp. 90–93, 2016.

[9] L. E. Nigrovic, N. Kuppermann, C. G. Macias et al., "Clinical prediction rule for identifying children with cerebrospinal fluid pleocytosis at very low risk of bacterial meningitis," *JAMA*, vol. 297, no. 1, pp. 52–60, 2007.

[10] F. Dubos, B. Korczowski, D. A. Aygun et al., "Serum procalcitonin level and other biological markers to distinguish between bacterial and aseptic meningitis in children: a European multicenter case cohort study," *Archives of Pediatrics & Adolescent Medicine*, vol. 162, no. 12, pp. 1157–1163, 2008.

[11] X. Saez-Llorens and G. H. McCracken Jr., "Bacterial meningitis in children," *T Lancet*, vol. 361, no. 9375, pp. 2139–2148, 2003.

[12] D. van de Beek, M. C. Brouwer, G. E. Thwaites, and A. R. Tunkel, "Advances in treatment of bacterial meningitis," *Lancet*, vol. 380, no. 9854, pp. 1693–1702, 2012.

[13] L. E. Nigrovic, R. Malley, C. G. Macias et al., "Effect of antibiotic pretreatment on cerebrospinal fluid profiles of children with bacterial meningitis," *Pediatrics*, vol. 122, no. 4, pp. 726–730, 2008.

[14] E. Ö. Başpınar, S. Dayan, M. Bekçibaşı et al., "Comparison of culture and PCR methods in the diagnosis of bacterial meningitis," *Brazilian Journal of Microbiology*, vol. 48, no. 2, pp. 232–236, 2017.

[15] L. E. Nigrovic, N. Kuppermann, and R. Malley, "Development and validation of a multivariable predictive model to distinguish bacterial from aseptic meningitis in children in the post-Haemophilus influenzae era," *Pediatrics*, vol. 110, no. 4, pp. 712–719, 2002.

[16] N. D. Phillips, H. Neth, J. K. Woike, and W. Gaissmaier, "FFTrees: a toolbox to create, visualize, and evaluate fast-and-frugal decision trees," *Judgment and Decision making*, vol. 12, no. 4, pp. 344–368, 2017.

[17] R. Steffen, A. Baños, and C. deBernardis, "Vaccination priorities," *International Journal of Antimicrobial Agents*, vol. 21, no. 2, pp. 175–180, 2003.

[18] Y. Chen, F. Li, M. Zhu, L. Liu, and Y. P. Luo, "Outcome and factors of patients with nosocomial meningitis by multidrug-resistant gram-negative bacteria in a tertiary hospital in China: a retrospective study," *British Journal of Neurosurgery*, vol. 34, no. 3, pp. 324–328, 2020.

[19] H. Lee, Y. Seo, K. H. Kim, K. Lee, and K. W. Choe, "Prevalence and serogroup changes of _Neisseria meningitidis_ in South Korea, 2010 -2016," *Scientific Reports*, vol. 8, no. 1, p. 5292, 2018.

[20] C. H. Wang, T. L. Lin, C. H. Muo et al., "Increase of meningitis risk in stroke patients in Taiwan," *Frontiers in Neurology*, vol. 9, p. 116, 2018.

[21] M. C. Brouwer, G. E. Thwaites, A. R. Tunkel, and D. van de Beek, "Dilemmas in the diagnosis of acute community-acquired bacterial meningitis," *Lancet*, vol. 380, no. 9854, pp. 1684–1692, 2012.

[22] M. Dastych, J. Gottwaldova, and Z. Cermakova, "Calprotectin and lactoferrin in the cerebrospinal fluid; biomarkers utilisable for differential diagnostics of bacterial and aseptic meningitis?," *Clinical Chemistry and Laboratory Medicine*, vol. 53, no. 4, pp. 599–603, 2015.

[23] J. W. Lee, C. I. Park, H. I. Kim et al., "The usefulness of serum delta neutrophil index for differentiating bacterial and viral meningitis in the emergency department," *Clinical and Experimental Emergency Medicine*, vol. 3, no. 2, pp. 95–99, 2016.

[24] M. Assicot, C. Bohuon, D. Gendrel, J. Raymond, H. Carsin, and J. Guilbaud, "High serum procalcitonin concentrations in patients with sepsis and infection," *The Lancet*, vol. 341, no. 8844, pp. 515–518, 1993.

[25] D. Gendrel, J. Raymond, M. Assicot et al., "Measurement of procalcitonin levels in children with bacterial or viral meningitis," *Clinical Infectious Diseases*, vol. 24, no. 6, pp. 1240–1242, 1997.

[26] A. Viallon, F. Zeni, C. Lambert et al., "High sensitivity and specificity of serum procalcitonin levels in adults with bacterial meningitis," *Clinical Infectious Diseases*, vol. 28, no. 6, 1999.

[27] S. Schwarz, M. Bertram, S. Schwab, K. Andrassy, and W. Hacke, "Serum procalcitonin levels in bacterial and abacterial meningitis," *Critical Care Medicine*, vol. 28, no. 6, pp. 1828–1832, 2000.

[28] O. Hoffmann, U. Reuter, F. Masuhr, M. Holtkamp, N. Kassim, and J. R. Weber, "Low sensitivity of serum procalcitonin in bacterial meningitis in adults," *Scandinavian Journal of Infectious Diseases*, vol. 33, no. 3, pp. 215–218, 2001.

[29] P. Ray, G. Badarou-Acossi, A. Viallon et al., "Accuracy of the cerebrospinal fluid results to differentiate bacterial from non bacterial meningitis, in case of negative gram-stained smear," *The American Journal of Emergency Medicine*, vol. 25, no. 2, pp. 179–184, 2007.

[30] H. Onal, Z. Onal, M. Ozdil, and S. Alhaj, "A new parameter in the differential diagnosis of bacterial and viral meningitis," *Neurosciences (Riyadh)*, vol. 13, no. 1, pp. 91-92, 2008.

[31] H. Tamune, H. Takeya, W. Suzuki et al., "Cerebrospinal fluid/-blood glucose ratio as an indicator for bacterial meningitis," *The American Journal of Emergency Medicine*, vol. 32, no. 3, pp. 263–266, 2014.

Hindawi
BioMed Research International
Volume 2021, Article ID 7332776, 7 pages
https://doi.org/10.1155/2021/7332776

Research Article

Estimation of Isentropic Compressibility of Biodiesel Using ELM Strategy: Application in Biofuel Production Processes

Marischa Elveny,[1] Meysam Hosseini,[2] Tzu-Chia Chen,[3] Adedoyin Isola Lawal,[4,5,6,7,8] and S. M. Alizadeh[9]

[1]*Data Science & Computational Intelligence Research Group, Universitas Sumatera Utara, Medan, Indonesia*
[2]*Department of Mathematics, Campus of Bijar, University of Kurdistan, Sanandaj, Kurdistan, Iran*
[3]*CAIC, DPU, Bangkok, Thailand*
[4]*Dept. of Accounting and Finance, Landmark University, Omu-Aran, Nigeria*
[5]*Sustainable Development Goal 17 (Partnership for the Goals) Research Cluster, Landmark University, Nigeria*
[6]*SDG 8 (Decent Work and Economic Growth) Research Cluster, Landmark University, Nigeria*
[7]*SDG1 (Zero Hunger) Research Cluster, Landmark University, Nigeria*
[8]*SDG6 (Clean Energy) Research Cluster, Landmark University, Nigeria*
[9]*Petroleum Engineering Department, Australian College of Kuwait, West Mishref, Kuwait*

Correspondence should be addressed to Marischa Elveny; marischaelveny@usu.ac.id, Meysam Hosseini; me.hosseini@uok.ac.ir, and Adedoyin Isola Lawal; lawal.adedoyin@lmu.edu.ng

Received 5 June 2021; Revised 23 June 2021; Accepted 1 July 2021; Published 13 July 2021

Academic Editor: Alireza Baghban

Isentropic compressibility is one of the significant properties of biofuel. On the other hand, the complexity related to the experimental procedure makes the detection process of this parameter time-consuming and hard. Thus, we propose a new Machine Learning (ML) method based on Extreme Learning Machine (ELM) to model this important value. A real database containing 483 actual datasets is compared with the outputs predicted by the ELM model. The results of this comparison show that this ML method, with a mean relative error of 0.19 and R^2 values of 1, has a great performance in calculations related to the biodiesel field. In addition, sensitivity analysis exhibits that the most efficient parameter of input variables is the normal melting point to determine isentropic compressibility.

1. Introduction

The suitability of oils and fats is often determined by their physicochemical properties. The fact that the terms "oil" and "fat" are used interchangeably in many languages indicates that the liquid or solid state of products at room temperature is considered crucial in distinguishing these two classes of goods [1]. From a technical point of view, to design rational lipidic materials, a thorough understanding of the rheological behavior, molecular structure, crystallization, and melting characteristics of oils and fats is needed [2]. It is well understood that the properties of oils and fats are principally determined by their triacylglycerol (TAG) structure, which includes the degree of unsaturation and the carbon chain length of the fatty acid molecules in the TAG component [3–5].

Besides their importance in food processing, animal fats and vegetable oils have been viewed as important sustainable resources for biodiesel production due to the impending depletion of fossil fuels [6–8]. The wide range of source oil FA profile is linked to many important biodiesel parameters including density, pour point, cloud point, cold filter plug-

ging point, and viscosity. As a result, animal or plant/seed sources utilized in biofuel production play an important role in biodiesel quality [9–11].

The laboratory assessment of physicochemical characteristics of oils and fats, like viscosity, density, composition, crystallization, and melting, necessitates the use of a variety of analytical tools, including differential scanning calorimeter, nuclear magnetic resonance, high-performance gas or liquid chromatography, and spectrometer analyzers [12–14]. Nevertheless, given the variety of feedstock, FA profiles, and lipids, this huge need for analytical instruments can make the physical characterization of these items an expensive and delaying process [15–17]. Since it is not possible to collect data on properties under all possible conditions, accurate methods for predicting them can be very useful for the design of products and processes [18]. In predictive modeling, the physicochemical phenomena-based models can be more complete and less constrained compared to simple polynomial or linear-fitted equations [19–22].

In general, the construction of models with a physics base begins with an understanding of the mathematics of the phenomena under investigation and then continues with doing simplifications to obtain a realistic model that presents a reasonable explanation for the phenomenon [23]. This analytical technique of modeling has commonly been applied in many fields, as shown by many literature references [24–26]. Although modeling approaches have primarily been utilized to explain thermodynamic features of fat and oil melting and crystallization, they have also been used in the production processes of biodiesel, process optimizations, and quantitative determination of biodiesel properties and compare them with thermodynamic characteristics [27, 28].

Given the above, a piece of detailed knowledge about the constraints and potentials of estimating modeling used in the measurement of physicochemical properties of biodiesel fuels is needed to fully exploit the opportunities presented by this method for the production of novel fat-based products and the processes of biodiesel production. In this paper, for the first time, the ELM algorithm is used to model and predict isentropic compressibility, one of the important properties of biodiesel. In this research work, after stating how to collect and use experimental data, the modeling method of this theorem is stated and in Results and Discussion, various analyses are used to evaluate the accuracy of this method.

2. Actual Data Collection

The database, including 483 data, related to this study was gathered from the literature [29]. In the following, we develop thorough the precise methods to estimate the isentropic compressibility. Also, variables were selected on the basis of existing data (including pressure, temperature, molecular weight, and melting point) due to having a predictive tool to estimate output values. It is noted that this database is divided into 120 testing data and 363 training data by chance. Then, after implementation of data, they are normalized as follows:

$$X_N = 2\frac{x - x_{\min}}{x_{\max} - x_{\min}} - 1. \tag{1}$$

3. Extreme Learning Machine (ELM)

ELM is invented by Huang et al. that has a structure like a single-layer feed-forward Neural Network (NN). But they differ from each other because of lacking bias of the output neuron (ON) [30, 31]. In an ELM algorithm, every input layer neuron is linked to all of the neurons in the hidden layer (HL). So, all of the neurons in the HL can have values related to their own bias, and the activation function of the output layer has a linear form, while the activation function of the HL is in the form of a piecewise continuous function [32]. Unlike other algorithms such as the back-propagation algorithm or conjugate gradient descent, ELM uses another way to find bias and weight [33]. In this way, ELM uses an algorithm to determine weights and biases of input layer neurons and those of HL neurons, randomly. So, it is assumed that an ELM algorithm has "i" input neurons and "k" training cases with "j" HL neurons where the HL activation can be defined as follows:

$$H_{jk} = g\left(\sum\left(w_{ji}x_{ik}\right) + B_j\right), \tag{2}$$

in which H_{jk} is the activation matrix of j_{th} HL neuron for the k_{th} training case, g is the nonlinear activation function, B_j is the bias of j_{th} HL neuron, x_{ik} is the i_{th} input neuron for k_{th} training neuron, and w_{ji} is the weight between i_{th} input neuron and j_{th} HL neuron.

The $i \times j$-dimension H matrix shows all HL neurons activated for all training cases [34].

By fitting least-squares on targets in training, we can compute the weight values between the HL neurons and ON. This implementation is linear and performed by Eqs. (3)–(5) as follows [35, 36]:

$$H_{k\times j}\beta_{j\times 1} = T_{k\times 1}, \tag{3}$$

$$\beta = \left(\beta_1 \cdots \beta_j\right)_{j\times 1}, \tag{4}$$

$$T = \left(T_1 \cdots T_k\right)_{k\times 1}, \tag{5}$$

where T and β are the target vector of training cases and the weight vector of hidden neurons and ON. Rather than these equations, we can multiply the Moore-Penrose pseudoinverse matrix, H', by T. This computation seems like the least-squares multilinear regression is implementing [37].

$$\beta = H'T, \tag{6}$$

in which, H' is the inverse matrix, known as Moore-Penrose pseudoinverse of H. So, after doing these calculations, the network training is completed [38, 39].

In total, this training process has only two essential steps: (1) determination of random biases and weights for hidden

Table 1: Various statistical analyses according to ELM algorithm.

Model	Dataset	R^2	MRE (%)	MSE	RMSE	STD
Isentropic compressibility (1/Gpa)	Train	1.000	0.18	0.0000020	0.0014306	0.0010710
	Test	1.000	0.21	0.0000032	0.0017773	0.0013781
	Total	1.000	0.19	0.0000023	0.0015249	0.0011567

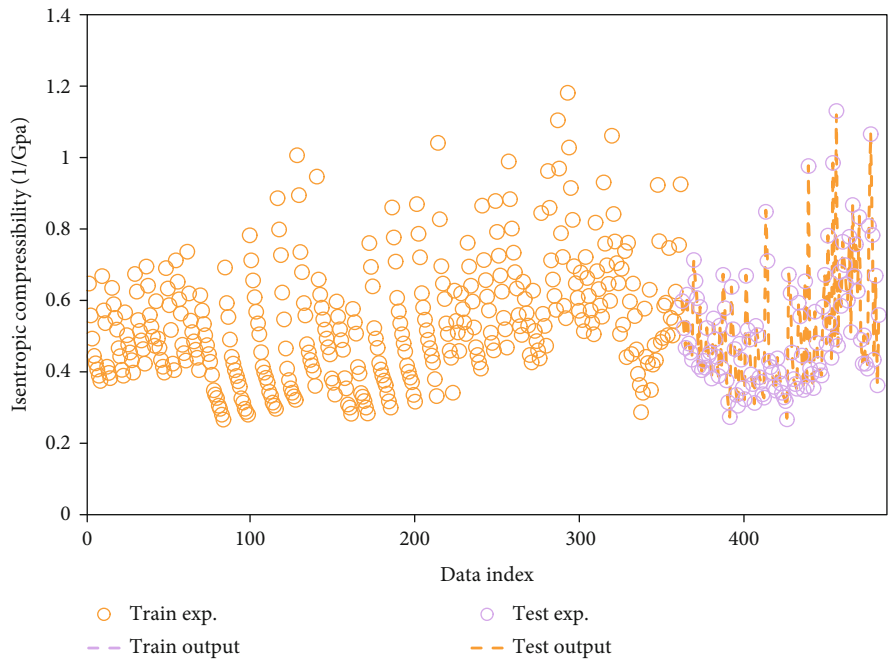


Figure 1: Simultaneous and visual comparison of real and corresponding modeled data in test and training phases.

neurons to find the HL output and (2) estimation of output weights using Moore-Penrose pseudoinverse of H. As mentioned before, the training process of the ELM algorithm is implemented by the determination of H' for the HL. This process is much faster than other training methods such as Levenberg-Marquardt. This method does not utilize the approach of nonlinear optimization, and it is only dependent on a closed-form solution [40, 41].

4. Results and Discussion

In the following, some formulations are used to estimate various types of statistical indices for the ELM algorithm.

$$\text{Mean squared error (MSE)} = \frac{1}{N}\sum_{i=1}^{N}\left(x_I^{\text{actual}} - x_I^{\text{predicted}}\right)^2, \quad (7)$$

$$\text{Mean relative error (MRE)} = \frac{100}{N}\sum_{i=1}^{N}\left(\frac{X_I^{\text{actual}} - X_I^{\text{predicted}}}{X_I^{\text{actual}}}\right), \quad (8)$$

$$R-\text{squared}\left(R^2\right) = 1 - \frac{\sum_{i=1}^{N}\left(x_I^{\text{actual}} - x_I^{\text{predicted}}\right)^2}{\sum_{i=1}^{N}\left(x_I^{\text{actual}} - \overline{x_I^{\text{actual}}}\right)^2}, \quad (9)$$

$$\text{Standard deviations (STD)} = \left(\frac{1}{N-1}\sum_{i=1}^{N}(\text{error} - \overline{\text{error}})\right)^{0.5}, \quad (10)$$

$$\text{Root mean square error (RMSE)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(x_I^{\text{actual}} - x_I^{\text{predicted}}\right)^2\right)}. \quad (11)$$

Table 1 shows the various statistical analyses for the evaluation of the ELM model in predicting actual output values. As can be seen from this table, this model shows a high ability to predict output values. By comparing the coefficient of determination values obtained from this model with similar work done by Abooali et al. [29], the better performance of the proposed model can be concluded. They used SGB and GP models to predict isentropic compressibility values, and their models were able to estimate this parameter with coefficients of determinations 0.99993 and 0.99608, respectively.

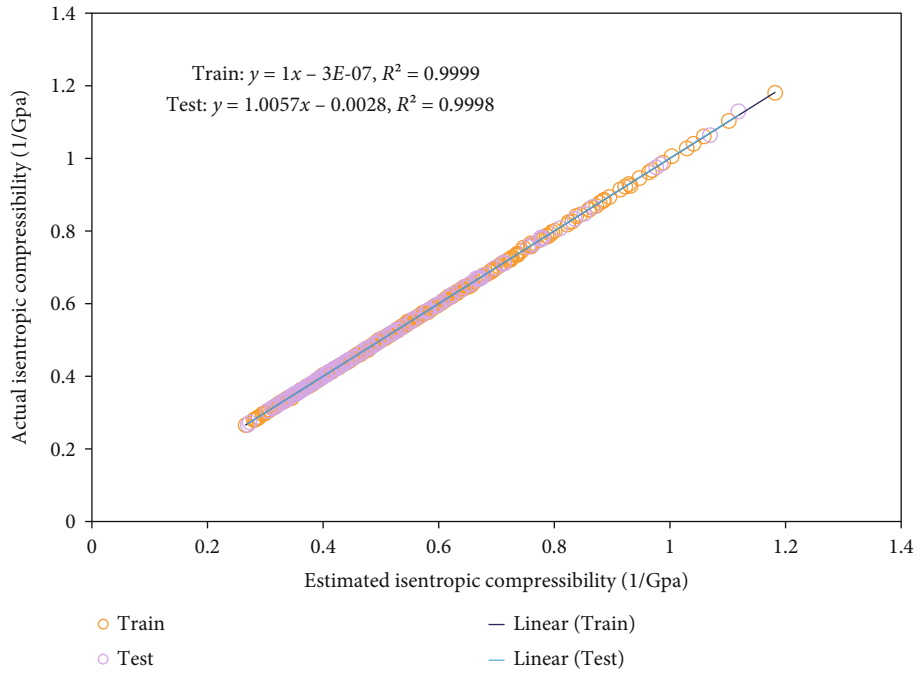


Figure 2: Cross plot analysis on the model to determine its accuracy in predicting actual values.

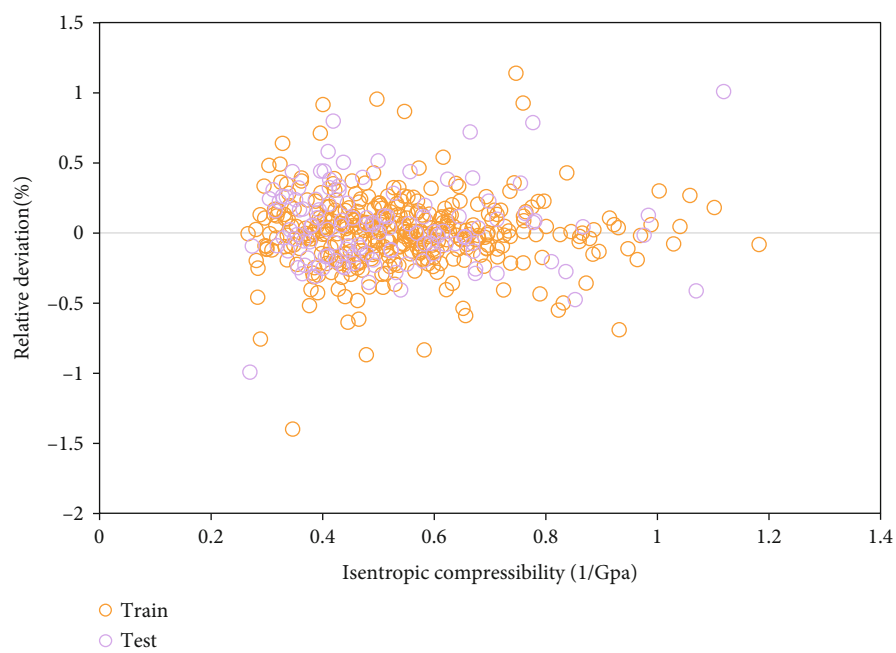


Figure 3: Relative deviation analysis on the model to determine its accuracy.

The comparison between modeled outputs and actual data is performed visually in Figure 1 that creates a helpful viewpoint about the precision of the ELM algorithm to make a prediction from the isentropic compressibility. Also, Figure 2 shows the cross plot of the regression of actual and modeled values of target data using the ELM algorithm. This prediction has an excellent agreement with real data for the model. As can be seen in this figure, the coefficients of determination related to the training and testing phases are 0.9999 and 0.9998, respectively.

According to actual data, the assessment is done by the relative deviations of generated outcomes of the ELM algorithm that is shown in Figure 3. So, it is concluded that relative errors of this algorithm, resulted in the prediction of target values, are close to zero. Also, all of the relative error values of the ELM algorithm are less than 1.5%, which verifies its power.

It is noted that the database utilized for the preparation of the proposed model can affect the accuracy and reliability of this model [42]. One of the important steps to propose

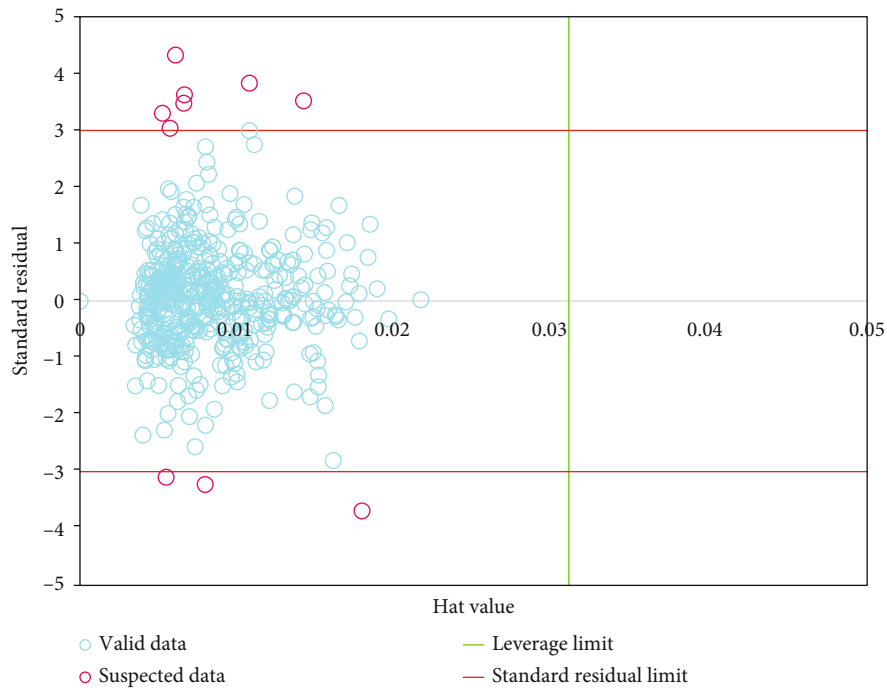


Figure 4: Leverage analysis to identify suspicious data.

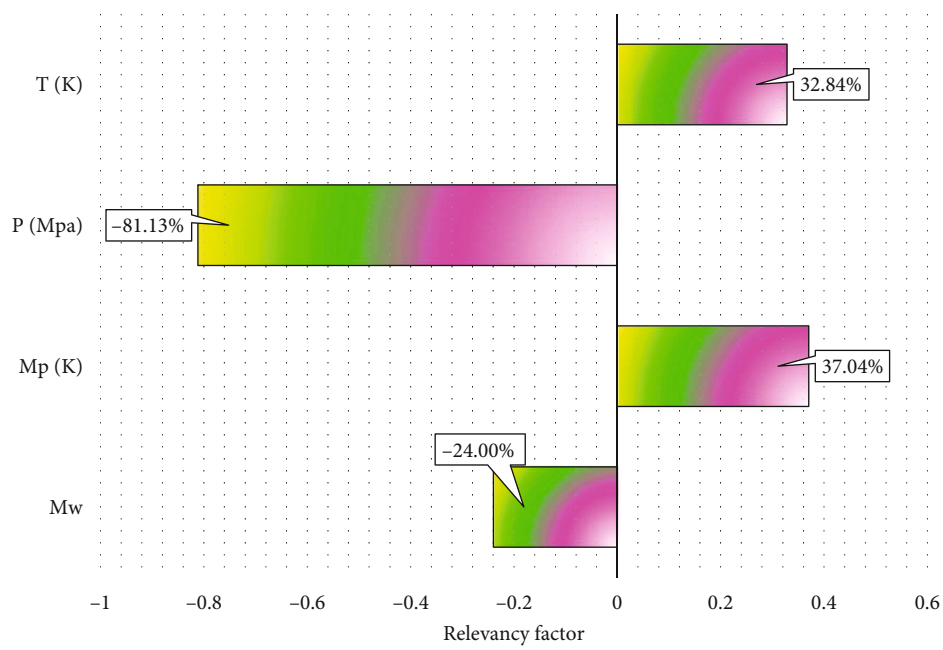


Figure 5: Sensitivity analysis on effective input data.

models with great accuracy is finding and removing suspected data. These data are the points that behave differently from others. So, to detect these kinds of data, Leverage analysis is employed. Also, to define the standardized residuals versus hat values, William's plot is used to recognize the outlying points. To determine the hat value on the basis of diagonal elements, the hat matrix is given as follows [43]:

$$H = A\left(A^T A\right)^{-1} A^T, \tag{12}$$

where A is a $a \times b$-dimension matrix which a and b are the number of the model parameter as well as training points, respectively. The squared limited area, known as reliable area, is enclosed by cut-off and warning leverage values of vertical and horizontal axes, respectively. The warning leverage values are calculated as follows:

$$H^* = \frac{3(b+1)}{a}. \tag{13}$$

Also, the cut-off value can be +3 and -3. The outlier detection of the ELM has been shown in Figure 4. As you see, the main number of the target data is placed within the nonsuspected/reliable area.

Then, sensitivity analysis is used to evaluate the dependence of the output values upon input parameters by a relevancy factor (r) in the range of +1 to -1 [44, 45]:

$$r = \frac{\sum_{i=1}^{n}\left(x_{K,i} - \bar{x}_k\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{K,i} - \bar{x}_k\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}, \tag{14}$$

where $X_{k.i}$ and Y_i are the input and output, as well as $\overline{X_k}$ and $\bar{Y}$ are mean values of input and outputs. Here, the higher absolute value of r shows the higher impact of arguing variable on the isentropic compressibility. Furthermore, negative and positive r values are related to this variable. The effects of the temperature, pressure, molecular weight, and melting point on the isentropic compressibility have been shown in Figure 5. It is notably shown that the temperature and melting point, with r values of 32.84% and 37.04%, are the most efficient parameters for the isentropic compressibility determination. Also, the pressure and molecular weight, with r values of -81.13% and -24%, are the least efficient variables for the isentropic compressibility, and these variables have a reverse relationship with the output value due to having negative r values.

5. Conclusion

In this study, we attempted at closing our aim by predicting the isentropic compressibility with the help of various affecting parameters based on a precise and new technique of the ELM. Thus, a comprehensive database is used for training and testing this algorithm. Afterward, the mathematical and graphical modeling was done, and it is shown that this algorithm can predict the isentropic compressibility with high accuracy of $R^2 = 1.000$, RMSE = 0.0015249, STD = 0.0011567, MRE = 0.19, and MSE = 0.0000023. It is shown that this model has a great ability to learn the behavior of the isentropic compressibility. In addition, it shows a great performance of the testing phase for unknown data points. Also, graphical comparisons demonstrated that the predicted data cover the real data with high accuracy for this algorithm. Last but not least, a comprehensive sensitivity analysis can be used to identify the effects of input variables on the determination of the isentropic compressibility. Temperature and melting point are considered as the most efficient parameters for finding the output values. These findings show that this study can help engineers to simulate and track this parameter in biodiesel. Previous implemented studies need too many parameters which may not be accessible, but our model requires the least number of parameters and predicts the output more precisely.

Data Availability

The data used to support the findings of this study are provided within the paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] R. Timms, "Physical properties of oils and mixtures of oils," *Journal of the American Oil Chemists' Society*, vol. 62, no. 2, pp. 241–249, 1985.

[2] A. G. Marangoni, N. Acevedo, F. Maleky et al., "Structure and functionality of edible fats," *Soft Matter*, vol. 8, no. 5, pp. 1275–1300, 2012.

[3] H. W. Lawson, *Food Oils and Fats: Technology, Utilization and Nutrition*, Springer Science & Business Media, 2013.

[4] A. J. Folayan, P. A. L. Anawe, A. E. Aladejare, and A. O. Ayeni, "Experimental investigation of the effect of fatty acids configuration, chain length, branching and degree of unsaturation on biodiesel fuel properties obtained from lauric oils, high-oleic and high-linoleic vegetable oil biomass," *Energy Reports*, vol. 5, pp. 793–806, 2019.

[5] L. E. Jamieson, A. Li, K. Faulds, and D. Graham, "Ratiometric analysis using Raman spectroscopy as a powerful predictor of structural properties of fatty acids," *Royal Society Open Science*, vol. 5, no. 12, p. 181483, 2018.

[6] N. Z. Zyaykina, V. Van Hoed, W. De Greyt, and R. Verhé, "The use of alternative lipid resources for bioenergy," *Lipid Technology*, vol. 21, no. 8-9, pp. 182–185, 2009.

[7] M. Ayoub and A. Z. Abdullah, "Critical review on the current scenario and significance of crude glycerol resulting from biodiesel industry towards more sustainable renewable energy industry," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 5, pp. 2671–2686, 2012.

[8] P. S. Nigam and A. Singh, "Production of liquid biofuels from renewable resources," *Progress in Energy and Combustion Science*, vol. 37, no. 1, pp. 52–68, 2011.

[9] C. Echim, J. Maes, and W. De Greyt, "Improvement of cold filter plugging point of biodiesel from alternative feedstocks," *Fuel*, vol. 93, pp. 642–648, 2012.

[10] K. Poyadji, M. Stylianou, A. Agapiou, C. Kallis, and N. Kokkinos, "Determination of quality properties of low-grade biodiesel and its heating oil blends," *Environments*, vol. 5, no. 9, p. 96, 2018.

[11] N. Isioma, Y. Muhammad, O.'. D. Sylvester, D. Innocent, and O. Linus, "Cold flow properties and kinematic viscosity of biodiesel," *Universal Journal of Chemistry*, vol. 1, no. 4, pp. 135–141, 2013.

[12] D. Firestone, *AOCS Official Methods and Recommended Practices of the AOCS*, AOCS, Urbana, IL, USA, 2009.

[13] J. S. de Ropp and M. J. McCarthy, "Nuclear Magnetic Resonance in the Analysis of Foodstuffs and Plant Materials," *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation*, 2006.

[14] A. I. Blake, E. D. Co, and A. G. Marangoni, "Structure and physical properties of plant wax crystal networks and their relationship to oil binding capacity," *Journal of the American Oil Chemists' Society*, vol. 91, no. 6, pp. 885–903, 2014.

[15] A. F. Talebi, S. K. Mohtashami, M. Tabatabaei et al., "Fatty acids profiling: a selective criterion for screening microalgae strains for biodiesel production," *Algal Research*, vol. 2, no. 3, pp. 258–267, 2013.

[16] A. E. Atabani, A. S. Silitonga, I. A. Badruddin, T. M. I. Mahlia, H. H. Masjuki, and S. Mekhilef, "A comprehensive review on biodiesel as an alternative energy resource and its characteristics," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 4, pp. 2070–2093, 2012.

[17] I. A. Adeoti and K. Hawboldt, "A review of lipid extraction from fish processing by-product for use as a biofuel," *Biomass and Bioenergy*, vol. 63, pp. 330–340, 2014.

[18] E. Pereira, A. J. Meirelles, and G. J. Maximo, "Predictive models for physical properties of fats, oils, and biodiesel fuels," *Fluid Phase Equilibria*, vol. 508, p. 112440, 2020.

[19] A. Erdal Tümer, S. Edebali, and Ş. Gülcü, "Modeling of removal of chromium (VI) from aqueous solutions using artificial neural network," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 39, no. 1, pp. 163–175, 2020.

[20] R. Maachou, A. Lefkir, A. Bermad, and S. Abdelaziz, "Energy consumption modeling in activated sludge process using coupling PCA-ANFIS approach," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 38, no. 6, pp. 261–273, 2019.

[21] S. Anbazhagan, V. Thiruvengatam, and K. Kulanthai, "Adaptive neuro-fuzzy inference system and artificial neural network modeling for the adsorption of methylene blue by novel adsorbent in a fixed-bed column method," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 39, no. 6, pp. 75–93, 2020.

[22] A. N. Seghat, M. Mousavi, J. Sargolzaei, and M. Khoshnoudi, "A neuro-fuzzy model for a dynamic prediction of milk ultrafiltration flux and resistance," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 26, no. 2, pp. 53–61, 2007.

[23] A. K. Datta, "Status of physics-based models in the design of food products, processes, and equipment," *Comprehensive Reviews in Food Science and Food Safety*, vol. 7, no. 1, pp. 121–129, 2008.

[24] B. Nicolaie, P. Verboven, and N. Scheerlinck, *The modelling of heat and mass transfer*, Food Process Modelling, Woodhead Publishing Limited and CRC Press LLC, Cambridge, UK and Boca Raton, FL, USA, 2001.

[25] M. M. Farid, *Mathematical Modeling of Food Processing*, CRC Press, 2010.

[26] F. Erdogdu, F. Sarghini, and F. Marra, "Mathematical modeling for virtualization in food processing," *Food Engineering Reviews*, vol. 9, no. 4, pp. 295–313, 2017.

[27] C. Himawan, V. Starov, and A. Stapley, "Thermodynamic and kinetic aspects of fat crystallization," *Advances in Colloid and Interface Science*, vol. 122, no. 1-3, pp. 3–33, 2006.

[28] B. Sajjadi, A. A. A. Raman, and H. Arandiyan, "A comprehensive review on properties of edible and non-edible vegetable oil-based biodiesel: composition, specifications and prediction models," *Renewable and Sustainable Energy Reviews*, vol. 63, pp. 62–92, 2016.

[29] D. Abooali, R. Soleimani, and S. Gholamreza-Ravi, "Characterization of physico-chemical properties of biodiesel components using smart data mining approaches," *Fuel*, vol. 266, p. 117075, 2020.

[30] J. Zhang and W. Ding, "Prediction of air pollutants concentration based on an extreme learning machine: the case of Hong Kong," *International Journal of Environmental Research and Public Health*, vol. 14, no. 2, p. 114, 2017.

[31] S. M. Salaken, A. Khosravi, T. Nguyen, and S. Nahavandi, "Extreme learning machine based transfer learning algorithms: a survey," *Neurocomputing*, vol. 267, pp. 516–524, 2017.

[32] Y. Park and H. S. Yang, "Convolutional neural network based on an extreme learning machine for image classification," *Neurocomputing*, vol. 339, pp. 66–76, 2019.

[33] E. J. Sadgrove, G. Falzon, D. Miron, and D. Lamb, "Fast object detection in pastoral landscapes using a colour feature extreme learning machine," *Computers and Electronics in Agriculture*, vol. 139, pp. 204–212, 2017.

[34] K. Javed, R. Gouriveau, and N. Zerhouni, "A new multivariate approach for prognostics based on extreme learning machine and fuzzy clustering," *IEEE transactions on cybernetics*, vol. 45, no. 12, pp. 2626–2639, 2015.

[35] R. C. Deo, P. Samui, and D. Kim, "Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models," *Stochastic Environmental Research and Risk Assessment*, vol. 30, no. 6, pp. 1769–1784, 2016.

[36] A. Akusok, K. M. Bjork, Y. Miche, and A. Lendasse, "High-performance extreme learning machines: a complete toolbox for big data applications," *IEEE Access*, vol. 3, pp. 1011–1025, 2015.

[37] H. Bonakdari and I. Ebtehaj, "A comparative study of extreme learning machines and support vector machines in prediction of sediment transport in open channels," *International Journal of Engineering*, vol. 29, no. 11, pp. 1499–1506, 2016.

[38] E. Khamehchi and A. Bemani, "Prediction of pressure in different two-phase flow conditions: machine learning applications," *Measurement*, vol. 173, p. 108665, 2021.

[39] S. Kariminia, S. Shamshirband, S. Motamedi, R. Hashim, and C. Roy, "A systematic extreme learning machine approach to analyze visitors' thermal comfort at a public urban space," *Renewable and Sustainable Energy Reviews*, vol. 58, pp. 751–760, 2016.

[40] Y. Bengio, *Learning Deep Architectures for AI*, Now Publishers Inc, 2009.

[41] C. R. Rao and S. K. Mitra, "Further contributions to the theory of generalized inverse of matrices and its applications," *Sankhyā: The Indian Journal of Statistics, Series A*, vol. 33, pp. 289–300, 1971, http://repository.ias.ac.in/96504/.

[42] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, vol. 589, John wiley & sons, 2005.

[43] M. Hosseinzadeh and A. Hemmati-Sarapardeh, "Toward a predictive model for estimating viscosity of ternary mixtures containing ionic liquids," *Journal of Molecular Liquids*, vol. 200, pp. 340–348, 2014.

[44] N. Kardani, A. Zhou, M. Nazem, and S. L. Shen, "Estimation of bearing capacity of piles in cohesionless soil using optimised machine learning approaches," *Geotechnical and Geological Engineering*, vol. 38, no. 2, pp. 2271–2291, 2020.

[45] M. N. Kardani and A. Baghban, "Utilization of LSSVM strategy to predict water content of sweet natural gas," *Petroleum Science and Technology*, vol. 35, no. 8, pp. 761–767, 2017.

Hindawi
BioMed Research International
Volume 2021, Article ID 4814888, 9 pages
https://doi.org/10.1155/2021/4814888

Research Article

On the Investigation of Effective Factors on Higher Heating Value of Biodiesel: Robust Modeling and Data Assessments

Shicheng Wang,[1] **Wei Li**,[1,2] **and Issam Alruyemi**[3]

[1]*School of Economics and Management, Southwest Petroleum University, Chengdu, Sichuan 610500, China*
[2]*Petroleum Engineering School, Southwest Petroleum University, Chengdu, Sichuan 610500, China*
[3]*Fouman Faculty of Engineering, College of Engineering, University of Tehran, Fouman, Iran*

Correspondence should be addressed to Shicheng Wang; shicheng_wang@hotmail.com
and Issam Alruyemi; essamkhudur@ut.ac.ir

Received 15 June 2021; Revised 22 June 2021; Accepted 1 July 2021; Published 12 July 2021

Academic Editor: Alireza Baghban

Higher heating value (HHV) is one of the properties of biomass fuels which is essential in investigating their special characteristics and potentialities. In this paper, various techniques based on Gaussian process regression (GPR) were utilized to assess this value for biomass fuels, including several kernel functions, i.e., exponential, Matern, rational quadratic, and squared exponential. An extensive databank was collected from literature. The findings were compared, and the results indicated that Exponential-based model was more accurate, with the coefficient of regression (R^2) of 0.961 and the mean relative error (% MRE) of 3.11 for total data. Compared to former models presented by previous researchers, the model proposed in this study showed a higher ability to predict output values. With various analyses, it can be concluded that the proposed method has a high rate of efficiency in assessing the HHV of various biomass.

1. Introduction

The use of fossil fuels has problems and disadvantages such as environmental pollution, asphaltene deposition, and limited resources [1–3]. There have been growing attempts at reducing the use of conventional fossil fuels and finding suitable replacements to use in a world with an ever-increasing population and industrial expansion, a compromised environment, and steadily depleting energy sources. Among these alternatives, biomass has become of particular interest due to its carbon neutrality and ease in being processed (e.g., chemically, thermally, and biochemically) to produce energy [4]. In recent years, coal-fired power stations have turned to use biomass to replace part of their fuel. This way, without needing to change any of their equipment, they can lower their use of coal and thus contribute to environmental and economic prosperity [5, 6].

Characteristics of biomass fuel, before being incorporated as a renewable source of energy, must be fully identified. Among these characteristics, the higher heating value is fundamental for allocating the feedstock for specific uses. The conventional method of measuring the HHV for liquid and solid fuel is adiabatic oxygen bomb calorimetry, which is, however, expensive and inefficient [7]. There are two methods of finding correlations for HHV: ultimate and proximate. The former is capable of identifying the composition of the fuel and its elements [8] but is more expensive than the latter method and cannot function without specific prior experiments. This has led to the widespread use of the proximate method of analysis [9]. This method works by first determining the changes in the enthalpy of products and reactants of a specific type of fuel. The procedure is not complicated but it takes a long time and requires equipment that might sometimes be unavailable. As a result, calculations are

made by other empirical methods using the data from proximate or elemental analysis. Proximate analysis, which is simpler and faster, has more widespread use for measuring the HHV. From the gathered data, fixed carbon (FC), volatile matter (VM), and ash are the factors incorporated in calculations [10, 11].

In previous years, the use of artificial intelligence (AI) methods has many applications in various fields, and researchers have investigated complete and close analyses to develop empirical methods (which mostly involve linear and nonlinear models) to reliably approximate the HHV of different types of biomass fuel [12]. Despite the efforts to estimate biomass, the complications associated with its structure make understanding the relationship between HHV and the data from proximate or ultimate analysis problematic. As a result, attention has recently been turned to artificial intelligence and its high potential to solve complicated problems. Mesroghli et al. utilized ANN models to assess the HHV of coal [13]. Ghugare et al. assessed the HHV of solid biomass fuel utilizing MPL-ANN and GA-based models and used ultimate analysis to find correlation [14]. Another attempt at estimating HHV of biomass was undertaken by Hosseinpour et al. [15] using iterative neural network-adapted partial least squares. The data gathered by the proximate analysis were incorporated into an ANFIS model by Akkaya to estimate the heating value (HV) of biomass [16]. Uzun et al. experimented with various ANN structures to estimate the HHV of biomass [17]. Finally, Estiati et al. utilized ANN together with a few linear models [18].

The present study involves expanding models of estimating the HHV for biomass fuels to replace the ultimate analysis with the proximate analysis, which is both cheaper and faster. Innovative models are introduced based on Gaussian process regression modeling including four kernel functions, i.e., exponential, Matern, rational quadratic, and squared exponential. To design the models, the data regarding the HHV of various biomass were gathered from 382 studies. A comparison is drawn of these models with those studied and published in the past. The new models were further studied for their efficacy and usefulness in six types of biomass fuel.

2. Materials and Methods

2.1. Data Collection. The independent variables of volatile matter (VM), ash (A), and fixed carbon (FC) content on dry basis are the inputs in the present study. The output is the data regarding the HHV of biomass. Here, the aim is to find the most practical y or function f for the input data x_1, x_2, x_3, i.e., FC, VM, and A, and y or function f indicates the HHV of biomass fuels.

The data from 382 proximate analyses regarding biomass and their HHVs were gathered from open literature. The data collected have been reported elsewhere [19]. The data regarding HHV were categorized into the following six groups:

(1) Byproducts of fruits

(2) Agri-wastes

(3) Wood chips and/or tree species

(4) Grasses, leaves, and fibrous materials

(5) Other waste materials

(6) Briquettes, charcoals, and pellets

Learning from literature, 30% of the data were randomly set apart as a test set to prevent overtraining [20]. Designing and training the nonlinear regression and AI models were performed using the remaining 70%. The test dataset helped examine the precision of the results and generalize the newly proposed models.

2.2. Gaussian Process Regression. To establish Gaussian process regression (GPR), it is required to select random training dataset $L = \{x_{L,i}, Y_{L,i}\}_{i=1}^{n}$ and testing dataset $T = \{x_{T,i}, Y_{T,i}\}_{i=1}^{n}$ from a particular distribution. The training dataset is employed to set the tuning parameters of the model [21, 22]. The testing dataset, which includes the excluded observations of the previous stage, is utilized to perform the approximate justification of the extended model. Also, x is the input variable, while y denotes the target variable. They are impacted by noise. The general form of GPR modeling is formulated as [22]:

$$y_{L,i} = f(x_{L,i}) + \varepsilon_{L,i},\ i = 1.2.3 \cdots \cdots n, \tag{1}$$

in which x_L is the independent variable of the learning dataset, y_L is the learning dataset target, and $\varepsilon \sim N(0, \sigma^2_{\text{noise}} I_n)$ represents the observation noise of an independent Gaussian distribution (where σ^2_{noise} stand for the noise variance, while I_{noise} represent the unit array variance). Then, the measured targets are connected to the function $f(x)$ by using a Gaussian noise model [23, 24]. It is worth mentioning that f values are assumed to be random variables in the GP. Likewise,

$$y_{T,i} = f(x_{T,i}) + \varepsilon_{T,i}, \quad i = 1.2.3 \cdots \cdots .. n, \tag{2}$$

in which x_T is the testing dataset independent variable, while y_T is the testing dataset target. Also, $f(x)$ is a latent parameter and has a GP distribution with a mean of $m(x)$ and covariance of $k(x, x')$ [23].

$$f(x_{T,i}) \sim Gp\left(m(x), K\left(x.x'\right)\right). \tag{3}$$

To specify the mean function $m(x)$, one can utilize an explicit basis function, even though it would lead to a complex specification of a fixed $m(x)$. To simplify the calculations, one can let $m(x)$ be zero [25–27]:

$$f(x_{L,i}) \sim Gp\left(0, K\left(x.x'\right)\right). \tag{4}$$

One can combine Equation (1) and Equation (4) to obtain

the prior distribution of y [25]:

$$y \sim N\left(0, K\left(x.x'\right) + \sigma^2_{\text{noise}} I_n\right). \tag{5}$$

The above equations could be collected as [27]:

$$\begin{bmatrix} \overrightarrow{f_l} \\ \overrightarrow{f_T} \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(x_L.x_L) & K(x_L.x_T) \\ K(x_T.x_L) & K(x_L.x_T) \end{bmatrix}\right),$$
$$\begin{bmatrix} \overrightarrow{\varepsilon_l} \\ \overrightarrow{\varepsilon_T} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma^2_{\text{noise}} I_n & 0 \\ 0 & \sigma^2_{\text{noise}} I_n \end{bmatrix}\right). \tag{6}$$

These equations can be summed up into a Gaussian formulation as [21]:

$$\begin{bmatrix} \overrightarrow{y_l} \\ \overrightarrow{y_T} \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(x_L - x_L) + \sigma^2_{\text{noise}} I_n & K(x_L.x_T) \\ K(x_T.x_L) & K\left(x_L.x_y\right) + \sigma^2_{\text{noise}} I_n \end{bmatrix}\right). \tag{7}$$

Then, the Gaussian conditioning rule could be applied to find the posterior distribution of y_T, [27]:

$$\left(\overrightarrow{y_T} \middle| \overrightarrow{y_L}\right) \sim N\left(\mu_r, \sum r\right), \tag{8}$$

where the mean value and covariance are written as:

$$\mu_T = m\left(\overrightarrow{y_T}\right) = K(x_T.x_L)\left(K(x_L.x_L) + \sigma^2_{\text{noise}} I_n\right)^{-1} \overrightarrow{y_L},$$
$$\sum r = K(x_T.x_T)K(x_T.x_T) + \sigma^2_{\text{noise}} I_n - K(x_T.x_L)\left(K(x_L.x_L) + \sigma^2_{\text{noise}} I_n\right)^{-1} K(x_T.x_L). \tag{9}$$

The theoretical GPR modeling concept is implemented. It is possible to predict the testing dataset outputs through the independent variable and training dataset [28]. These formulations are supportive of the claim that the mean function and covariance could provide a complete GP description through the introduction of the Gaussian distribution. It is important to select a Kernel function (i.e., a strong covariance function) in the training phase. The Kernel matrix has a symmetric, invertible matrix. This contributes to GPR model robustness in target prediction. To identify the optimal Kernel function, the present study manipulated four common Kernel functions, namely, (1) rational quadratic, (2) exponential, (3) squared exponential, and (4) Matern functions, to perform the learning process. The rational quadratic covariance function is defined as:

$$K_{RQ}\left(x.x'\right) = \sigma^2 \left(1 + \frac{x - x'2}{2\alpha\ell^2}\right)^{-\alpha}, \tag{10}$$

in which σ denotes the amplitude, σ^2 is the variance, ℓ represents the length scale, and $\alpha > 0$ is the scale mixture that ascertains the change weights at both small and large scales. The exponential covariance function is formulated as:

$$K_E\left(x.x'\right) = \sigma^2 \exp - \left(\frac{x - x'}{\ell}\right). \tag{11}$$

The squared exponential covariance function is expressed as:

$$K_{SE}\left(x.x'\right) = \sigma^2 \exp\left(-\frac{x - x'}{\ell^2}\right). \tag{12}$$

Finally, the Matern covariance function is represented as:

$$K_M\left(x.x'\right) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\text{v})}\left(\sqrt{2\nu}\frac{x - x'}{\ell}\right) K_\nu\left(\sqrt{2\nu}\frac{x - x'}{\ell}\right), \tag{13}$$

where Γ is the gamma function, K_ν represents the modified Bessel function, and ℓ and ν are positive variables. In fact, the exponential covariance function and squared exponential covariance function are two particular forms of the Matern covariance function. Setting ν to 0.5 converts the Matern covariance function into the exponential covariance function. Also, the Matern covariance function transforms into the squared exponential covariance function at a ν approaching infinity. In light of its additional parameter (i.e., ν) as a larger degree of freedom, the Matern covariance function could make more accurate estimates as compared to the exponential and squared exponential covariance functions.

3. Results and Discussion

3.1. Analysis of Validity and Reliability. For the accuracy and reliability evaluation of the developed GPR models in the higher heating value prediction of biodiesels, the present study performed a multivariable statistical test. This work coupled some typical statistical measures and some graphical depictions.

3.2. Statistical Variables. For the performance evaluation of the proposed models, the present study exploited the mean square error (MSE), the mean of relative error (MRE), standard deviation (STD), root mean square error (RMSE), and coefficient of determination (R^2).

Table 1: A comparison of the models in MRE, RMSE, MSE, STD, and R-squared for the training, testing, and total data under various Kernel functions.

Model	Phase	R^2	MRE (%)	MSE	RMSE	STD
GPR (exponential)	Train	0.973	2.70	0.45	0.67	0.46
	Test	0.907	4.53	1.03	1.02	0.63
	Total	0.961	3.11	0.58	1.02	0.52
GPR (Matern)	Train	0.951	3.57	0.78	0.88	0.59
	Test	0.932	4.07	0.80	0.90	0.52
	Total	0.944	3.82	0.83	0.90	0.59
GPR (squared exponential)	Train	0.941	3.40	0.72	0.85	0.58
	Test	0.947	4.91	1.18	1.09	0.69
	Total	0.940	3.99	0.89	1.09	0.62
GPR (rational quadratic)	Train	0.946	3.94	0.92	0.96	0.64
	Test	0.928	3.65	0.59	0.77	0.40
	Total	0.943	3.86	0.84	0.77	0.59

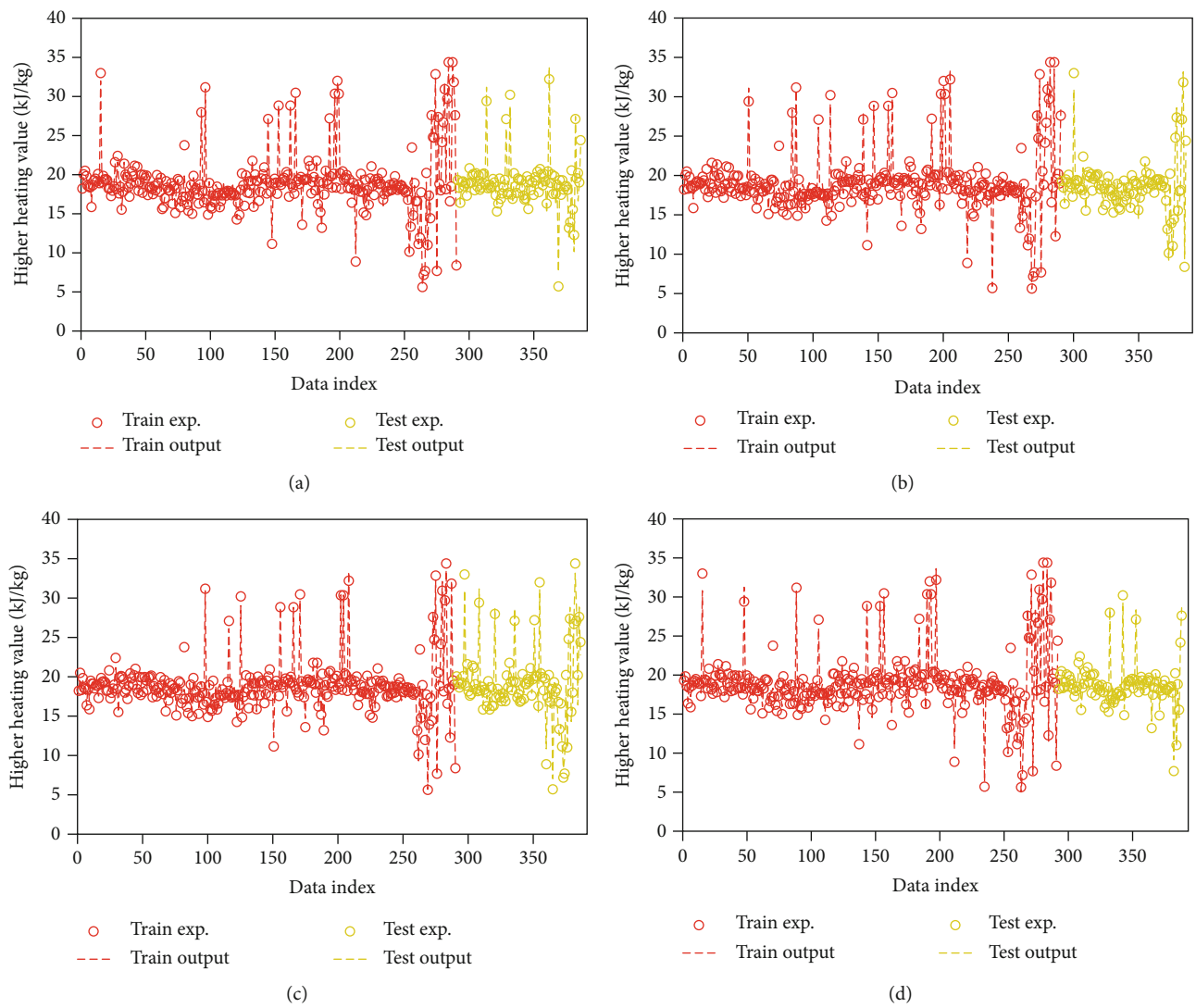


Figure 1: A point-by-point comparison of the modeled estimates to experimental quantities under the (a) exponential, (b) Matern, (c) squared exponential, and (d) rational quadratic kernel functions.

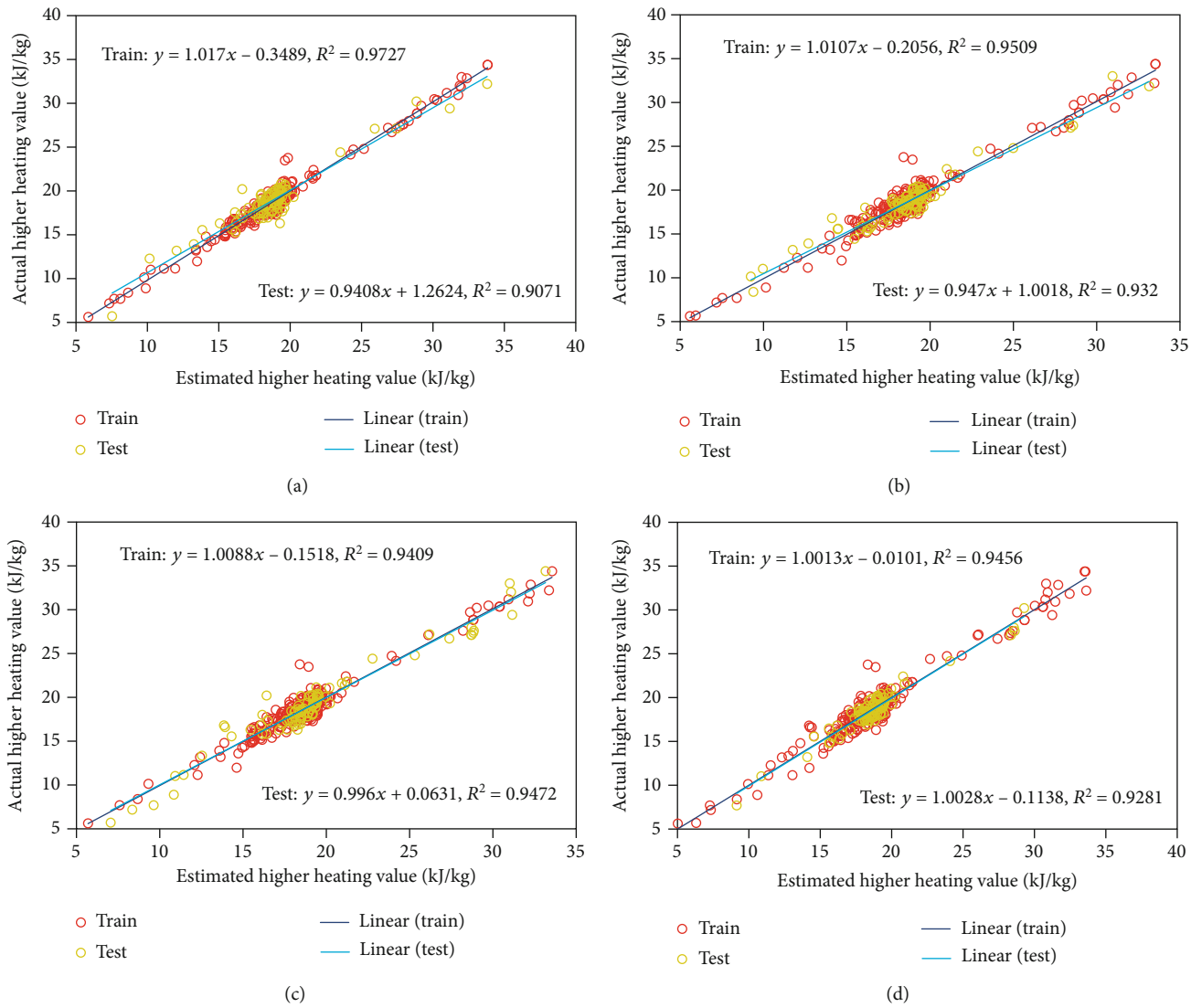


Figure 2: A cross plot comparison of the modeled estimates to the experimental data under the (a) exponential, (b) Matern, (c) squared exponential, and (d) rational quadratic kernel functions.

$$\begin{aligned}
\mathrm{MRE} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_{\mathrm{exp.},i} - y_{\mathrm{pred.},i}\right|}{y_{\mathrm{pred.},i}},\\
\mathrm{MSE} &= \frac{1}{n}\sum_{i=1}^{n}\left(y_{\mathrm{exp}.i} - y_{\mathrm{pred}.i}\right)^2,\\
\mathrm{RMSE} &= \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{\mathrm{exp}.i} - y_{\mathrm{pred}.i}\right)^2},\\
\mathrm{STD} &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_{\mathrm{exp}.i} - y_{\mathrm{pred}.i}}{y_{\mathrm{exp}.i}}\right)^2},\\
R^2 &= 1 - \frac{\sum_{i=1}^{n}\left(y_{\mathrm{pred}.i} - y_{\mathrm{exp}.i}\right)^2}{\sum_{i=1}^{n}\left(y_{\mathrm{pred}.i} - \overline{y_{\mathrm{exp}.i}}\right)^2}.
\end{aligned} \tag{14}$$

The statistical parameters related to the mentioned models are calculated and given in Table 1. Dashti and his colleagues used different models to predict the HHV data [19]. The input and output data used in our paper are similar to their work. The most powerful model they presented was the GARBF model, which has ability to estimate the target values with R^2 and MSE equal to 0.9500 and 0.7401, respectively. However, according to the values obtained in Table 1 of our paper, the GPR (exponential) model has the ability to estimate these values with an accuracy of 0.961 and 0.58, respectively.

3.3. Point-by-Point Agreement Plot. Figure 1 compares the HHV estimates of the GPR models to the measured values, in which "Data Index" represents the sample number, "Train Exp." Represents the experimental training set, "Train Output" stands for the training set estimate, "Test Exp." denotes the experimental testing dataset, and "Test Output" represents the testing dataset estimates. According to this figure, most estimates are in good agreement with the experimental data points in all the models. Also, the exponential approach

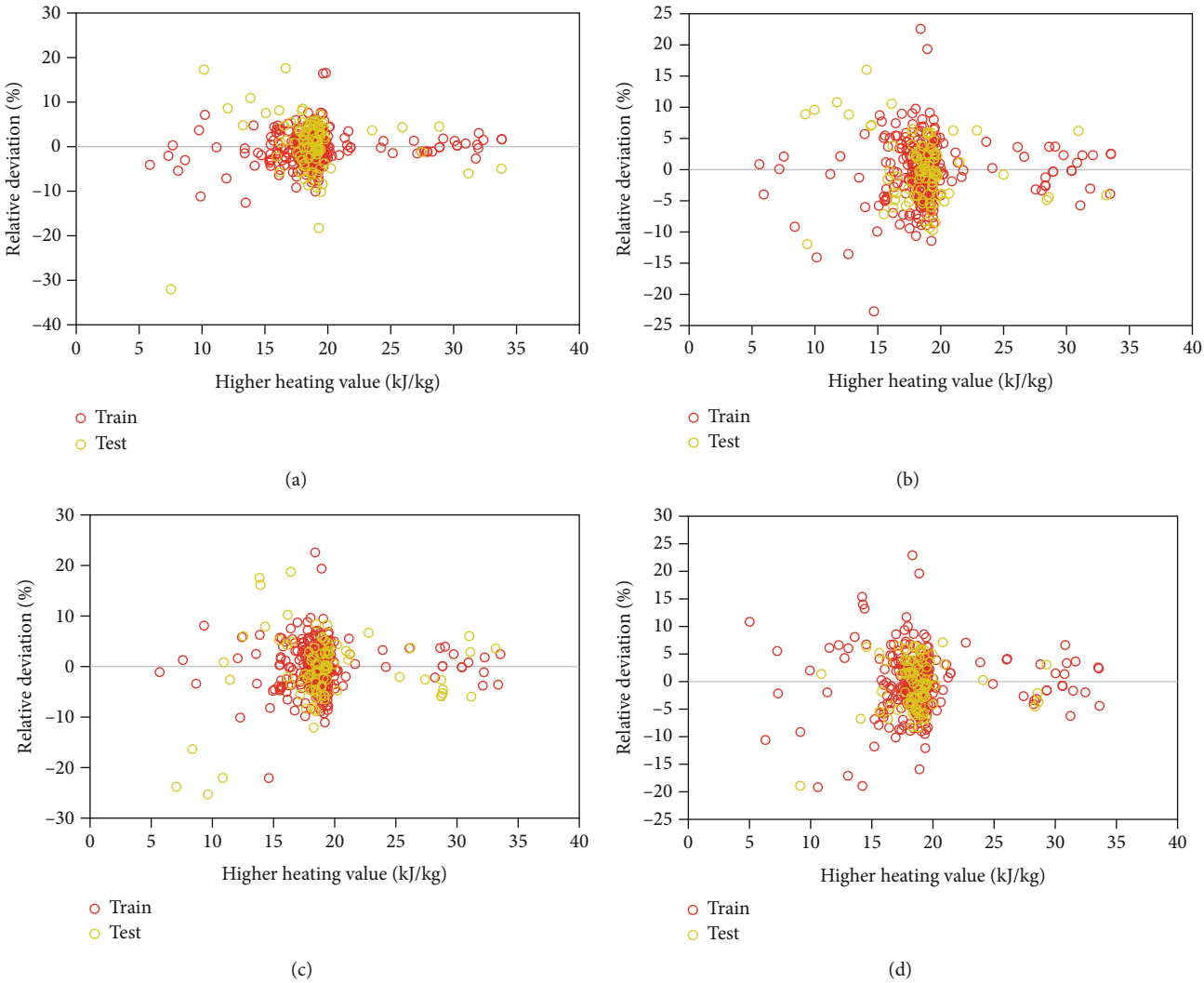


Figure 3: Relative deviations of the GPR models versus HHV measurements under the (a) exponential, (b) Matern, (c) squared exponential, and (d) rational quadratic kernel functions.

has the highest accuracy and lowest discrepancy. This is supportive of the statistical evaluation findings.

3.4. Cross Plot. Figure 2 illustrates the cross plots of experimental HHV quantities versus the corresponding estimates. It further supports the reliability of the proposed models. As can be seen, the linear trend with an R^2 range of 0.90-0.97 demonstrates that the predictions and measurements are consistent for both the training and testing datasets. As can be seen in Figure 2(a), the most accurate results were obtained by the exponential kernel function.

3.5. $y \sim y \sim$ Relative Deviation Distribution. Figure 3 depicts the relative deviation distributions of the HHV estimates of the developed GPR models. It should be noted that the relative deviation (RD) is calculated as:

$$\mathrm{RD}(\%) = 100 \times \left(\frac{y_{\exp.,i} - y_{\mathrm{pred.},i}}{y_{\exp.,i}} \right). \tag{15}$$

These graphs help determine the degree to which the calculations are realistic based on the experimental quantities. The reliability of the estimates is described by locations of the training and testing data points concerning the horizontal zero-line. According to Figure 3, most relative deviations were found to be from -20% to 20%, which is a favorable range. Furthermore, the points are mostly resting near the horizontal line (Figure 3(a)), in particular those of the exponential kernel function.

3.6. Sensitivity Analysis. The present study employed a sensitivity analysis to relate the exponential outputs to the independent input variables. Furthermore, this work employed the relevancy factor (RF) as Pearson's method as [29, 30]:

$$\mathrm{RF} = \frac{\sum_{i=1}^{n} (\overline{x_k} - x_k.i)^2 \sum_{i=1}^{n} (\bar{y} - y_i)}{\sqrt{\sum_{i=1}^{n} (\overline{x_k} - x_k.i)^2 \sum_{i=1}^{n} (\bar{y} - y_i)^2}}, \tag{16}$$

in which k denotes the input type, while n represents the

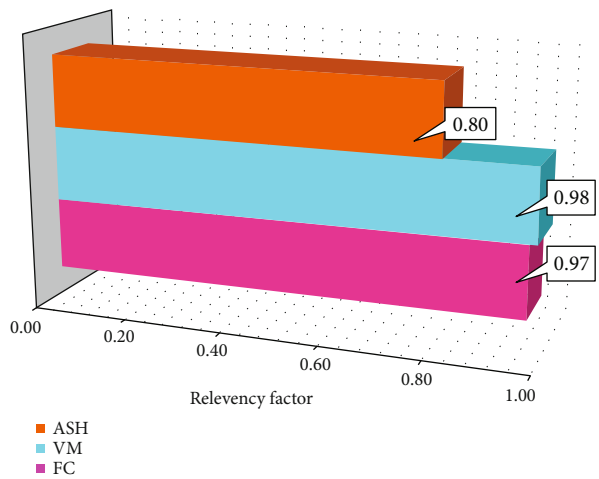


Figure 4: Sensitivity analysis results of the best GPR model.

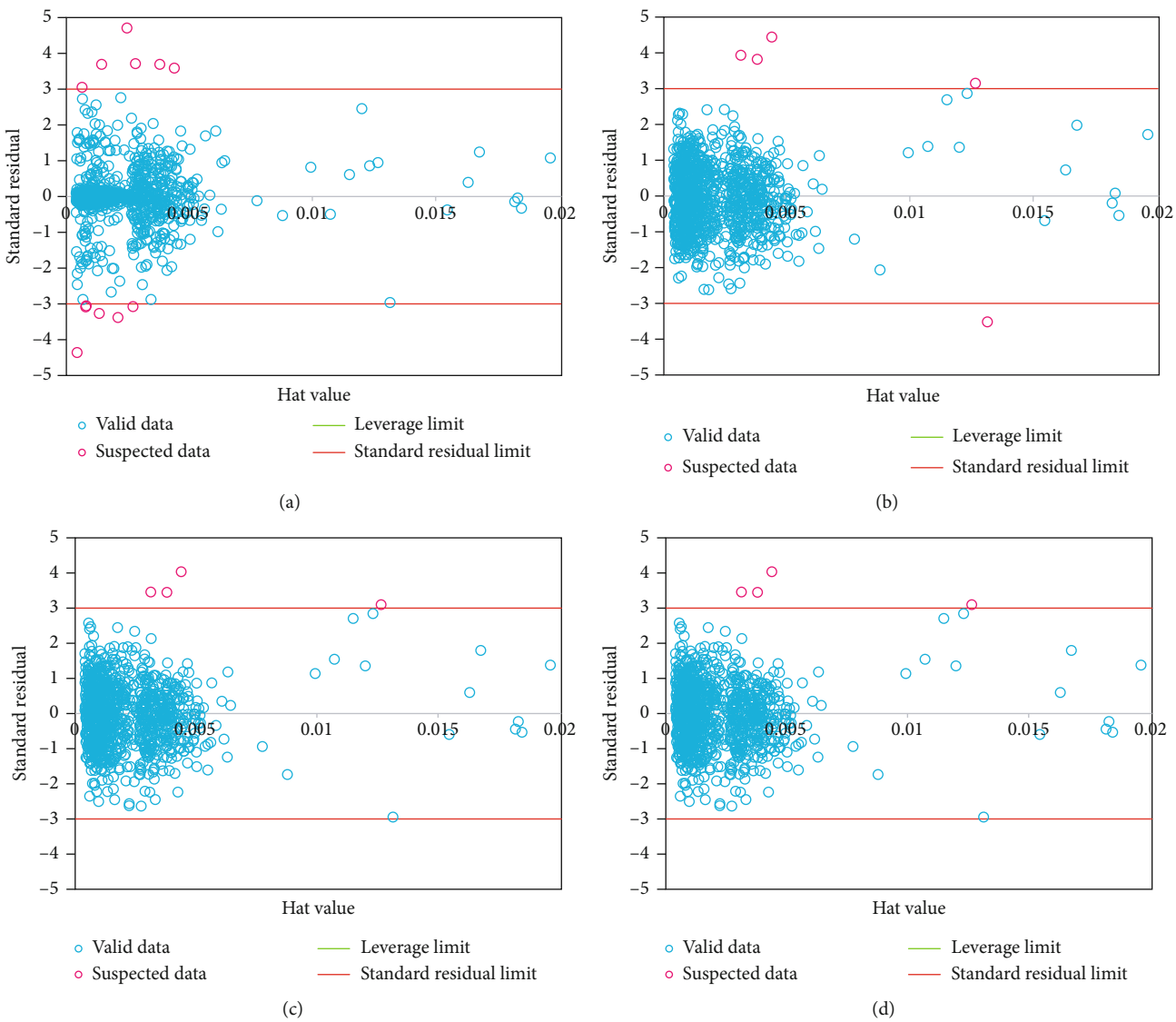


Figure 5: William plots of (a) exponential, (b) Matern, (c) squared exponential, and (d) rational quadratic kernel functions.

number of data points. Also, x is the input value, $\bar{x}_k$ is the average value of input k, y is the target, and $\bar{y}$ is the average value of the target [31, 32]. RF varies in the range of [-1, 1]; a negative RF represents an inverse relationship between the inputs and output, while a positive RF stands for a direct relationship. A smaller difference between RF and the above-mentioned limits would imply a stronger input-output relationship.

Figure 4 shows the relative deviation results of the proposed models. As can be seen, all input variables have a direct effect on the HHV. Hence, the proposed models can be said to be able to emulate the effects of several inputs on the target.

3.7. Outlier Detection. Laboratory data values are always accompanied by uncertainty. The present work employed the Williams plot of standardized residuals (R) versus leverage (H) to shed some light on uncertain points. The diagonal entries represent the leverage values in the projection matrix $H = X(X^T X)^{-1} X^T$, in which X represents the explanatory variable matrix, while T stands for the transpose matrix operator [33, 34]. A leverage value above the threshold implies uncertainty and a high-leverage point. The leverage threshold is obtained as [35, 36]:

$$H^* = \frac{3(\text{number of inputs} + 1)}{\text{number of data points}}. \tag{17}$$

Figure 5 illustrates the William plots of the proposed models. One can qualify the data points based on the corresponding locations in the plots. The model applicability domain is represented by the squared area of $-3 \leq R \leq 3$ and $H < H^*$. The area of $\leq R \leq 3$ and $h > H^*$ represents the good high leverage data. A question mark represents the model's ability to estimate data points resting in this area. The points that lie in the domains $R > 3$ or $R < -3$ are referred to as the bad high leverage data (i.e., outliers). According to Figure 5, a small number of points exist in the bad high leverage and good high leverage areas; the remaining points fall in the model applicability domain.

4. Conclusion

The present study adopted GPR and implemented a comprehensive modeling work on extensive data collected from the literature. HHV was modeled as a function of fixed carbon (FC), volatile matter (VM), and ash s by using four Kernel functions. The data were divided into training and testing datasets. This study utilized cross plots, relative deviation diagrams, sensitivity analyses, and Williams plots along with the parametric analysis of errors (including MRE, MSE, RMSE, and R^2). The developed GPR models were found to have high performance in the HHV estimation of biodiesels. The exponential function exhibited the highest accuracy, while the squared exponential function showed the lowest accuracy—the MRE and adjusted R^2 were calculated to be 3.11% and 0.961 for the exponential function, respectively, while they were obtained to be 3.99% and 0.94 for the squared exponential function, respectively. The cross plots and relative deviations demonstrated satisfactory consistency between the HHV measurements and estimates. Finally, the outlier analysis was performed to evaluate data validity and GPR model reliability.

Data Availability

The data used to support the findings of this study are provided within the paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Social Science Planning Project Fund of Sichuan Province (SC19C058) and Humanities Special Fund of Southwest Petroleum University (2018RW007).

References

[1] M. Simjoo, "Polyacrylamide gel polymer as water shut-off system: preparation and investigation of physical and chemical properties in one of the Iranian oil reservoirs conditions," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 26, no. 4, pp. 99–108, 2007.

[2] M. Tajmiri and M. R. Ehsani, "Wettability alteration of oil-wet and water-wet of iranian heavy oil reservoir by CuO nanoparticles," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 36, no. 4, pp. 171–182, 2017.

[3] M. Yazdizadeh, H. Nourbakhsh, and M. R. Jafari Nasr, "A solution model for predicting asphaltene precipitation," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 33, no. 1, pp. 93–102, 2014.

[4] X. Junming, J. Jianchun, and K. Sun, "An investigation of wood-plastic novolac modified by biomass pyrolysis oils," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 30, no. 1, pp. 83–87, 2011.

[5] M. Le Page, *The Great Carbon Scam*, Elsevier, 2016.

[6] D. R. Nhuchhen and P. A. Salam, "Estimation of higher heating value of biomass from proximate analysis: a new approach," *Fuel*, vol. 99, pp. 55–63, 2012.

[7] A. MAJUMDER, R. JAIN, P. BANERJEE, and J. BARNWAL, "Development of a new proximate analysis based correlation to predict calorific value of coal," *Fuel*, vol. 87, no. 13-14, pp. 3077–3081, 2008.

[8] A. Demirbaş, "Calculation of higher heating values of biomass fuels," *Fuel*, vol. 76, no. 5, pp. 431–434, 1997.

[9] P. Basu, *Biomass Gasification and Pyrolysis: Practical Design and Theory*, Academic press, 2010.

[10] J. Parikh, S. Channiwala, and G. Ghosal, "A correlation for calculating HHV from proximate analysis of solid fuels," *Fuel*, vol. 84, no. 5, pp. 487–494, 2005.

[11] T. Cordero, F. Marquez, J. Rodriguez-Mirasol, and J. J. Rodriguez, "Predicting heating values of lignocellulosics and carbonaceous materials from proximate analysis," *Fuel*, vol. 80, no. 11, pp. 1567–1571, 2001.

[12] R. Maachou, "Energy consumption modeling in activated sludge process using coupling PCA-ANFIS approach," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 38, no. 6, pp. 261–273, 2019.

[13] S. Mesroghli, E. Jorjani, and S. C. Chelgani, "Estimation of gross calorific value based on coal analysis using regression and artificial neural networks," *International Journal of Coal Geology*, vol. 79, no. 1-2, pp. 49–54, 2009.

[14] S. B. Ghugare, S. Tiwary, V. Elangovan, and S. S. Tambe, "Prediction of higher heating value of solid biomass fuels using artificial intelligence formalisms," *Bioenergy Research*, vol. 7, no. 2, pp. 681–692, 2014.

[15] S. Hosseinpour, M. Aghbashlo, M. Tabatabaei, and M. Mehrpooya, "Estimation of biomass higher heating value (HHV) based on the proximate analysis by using iterative neural network-adapted partial least squares (INNPLS)," *Energy*, vol. 138, pp. 473–479, 2017.

[16] E. Akkaya, "ANFIS based prediction model for biomass heating value using proximate analysis components," *Fuel*, vol. 180, pp. 687–693, 2016.

[17] H. Uzun, Z. Yıldız, J. L. Goldfarb, and S. Ceylan, "Improved prediction of higher heating value of biomass using an artificial neural network model based on proximate analysis," *Bioresource Technology*, vol. 234, pp. 122–130, 2017.

[18] I. Estiati, F. B. Freire, J. T. Freire, R. Aguado, and M. Olazar, "Fitting performance of artificial neural networks and empirical correlations to estimate higher heating values of biomass," *Fuel*, vol. 180, pp. 377–383, 2016.

[19] A. Dashti, A. S. Noushabadi, M. Raji, A. Razmi, S. Ceylan, and A. H. Mohammadi, "Estimation of biomass higher heating value (HHV) based on the proximate analysis: smart modeling and correlation," *Fuel*, vol. 257, p. 115931, 2019.

[20] A. Amani, P. York, H. Chrystyn, B. J. Clark, and D. Q. Do, "Determination of factors controlling the particle size in nanoemulsions using artificial neural networks," *European Journal of Pharmaceutical Sciences*, vol. 35, no. 1-2, pp. 42–51, 2008.

[21] H. Huang, A. Qin, H. Mao et al., "The prediction method on the early failure of hydropower units based on Gaussian process regression driven by monitoring data," *Applied Sciences*, vol. 11, no. 1, 2021.

[22] M. Mahdaviara, A. Rostami, F. Keivanimehr, and K. Shahbazi, "Accurate determination of permeability in carbonate reservoirs using Gaussian process regression," *Journal of Petroleum Science and Engineering*, vol. 196, p. 107807, 2021.

[23] C. Edward and K. Christopher, *Gaussian Processes for Machine Learning Adaptive Computation and Machine Learning*, MIT Press Cambridge, 2005.

[24] H. Yu, "The gaussian process regression for TOC estimation using wireline logs in shale gas reservoirs," in *In International Petroleum Technology Conference*, 2016.

[25] Q. Fu, W. Shen, X. Wei, P. Zheng, H. Xin, and C. Zhao, "Prediction of the diet nutrients digestibility of dairy cows using Gaussian process regression," *Information Processing in Agriculture*, vol. 6, no. 3, pp. 396–406, 2019.

[26] C. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*, vol. 2, MIT press Cambridge, MA, 2006.

[27] M. Zhu, S. Liu, and S. Gu, "Short-term tide level forecasting based on Gaussian process regression," in *in The Eleventh ISOPE Pacific/Asia Offshore Mechanics Symposium 2014 International Society of Offshore and Polar Engineers*, 2014.

[28] S. Asante-Okyere, C. Shen, Y. Yevenyo Ziggah, M. Moses Rulegeya, and X. Zhu, "Investigating the predictive performance of Gaussian process regression in evaluating reservoir porosity and permeability," *Energies*, vol. 11, no. 12, p. 3261, 2018.

[29] M. H. Ahmadi, A. Baghban, M. Ghazvini, M. Hadipoor, R. Ghasempour, and M. R. Nazemzadegan, "An insight into the prediction of TiO 2/water nanofluid viscosity through intelligence schemes," *Journal of Thermal Analysis and Calorimetry*, vol. 139, no. 3, pp. 2381–2394, 2020.

[30] A. Dargahi-Zarandi, A. Hemmati-Sarapardeh, S. Hajirezaie, B. Dabir, and S. Atashrouz, "Modeling gas/vapor viscosity of hydrocarbon fluids using a hybrid GMDH-type neural network system," *Journal of Molecular Liquids*, vol. 236, pp. 162–171, 2017.

[31] S. Hajirezaie, A. Hemmati-Sarapardeh, A. H. Mohammadi, M. Pournik, and A. Kamari, "A smooth model for the estimation of gas/vapor viscosity of hydrocarbon fluids," *Journal of Natural Gas Science and Engineering*, vol. 26, pp. 1452–1459, 2015.

[32] A. S. Noushabadi, A. Dashti, M. Raji, A. Zarei, and A. H. Mohammadi, "Estimation of cetane numbers of biodiesel and diesel oils using regression and PSO-ANFIS models," *Renewable Energy*, vol. 158, pp. 465–473, 2020.

[33] A. Kamari, A. H. Mohammadi, A. Bahadori, and S. Zendehboudi, "A reliable model for estimating the wax deposition rate during crude oil production and processing," *Petroleum Science and Technology*, vol. 32, no. 23, pp. 2837–2844, 2014.

[34] A. Kamari, A. Khaksar-Manshad, F. Gharagheizi, A. H. Mohammadi, and S. Ashoori, "Robust model for the determination of wax deposition in oil systems," *Industrial & Engineering Chemistry Research*, vol. 52, no. 44, pp. 15664–15672, 2013.

[35] D. C. Hoaglin and R. E. Welsch, "The hat matrix in regression and ANOVA," *The American Statistician*, vol. 32, no. 1, pp. 17–22, 1978.

[36] A. Kamari, A. Safiri, and A. H. Mohammadi, "Compositional model for estimating asphaltene precipitation conditions in live reservoir oil systems," *Journal of Dispersion Science and Technology*, vol. 36, no. 3, pp. 301–309, 2015.

Hindawi
BioMed Research International
Volume 2021, Article ID 5368987, 8 pages
https://doi.org/10.1155/2021/5368987

Research Article

Developing a Novel Method for Estimating the Speed of Sound in Biodiesel Known as Grey Wolf Optimizer Support Vector Machine Algorithm

Zhenzhen Lv,[1] **Ming Hu**,[1] **Yixin Yang**,[1] **and Jeren Makhdoumi**[2]

[1]*School of Electrical and Information Engineering, Anhui University of Technology, Maanshan, Anhui 243002, China*
[2]*Department of Educational Science, Payame Noor University, Damghan, Iran*

Correspondence should be addressed to Ming Hu; xiaoming-sudo@outlook.com
and Jeren Makhdoumi; jerenmkhd@student.pnu.ac.ir

Received 4 June 2021; Revised 13 June 2021; Accepted 25 June 2021; Published 2 July 2021

Academic Editor: Alireza Baghban

In the current study, our goal was to obtain a robust model to predict the speed of sound in biodiesel. For this purpose, an extensive databank has been extracted from previously published papers. Then, a Support Vector Machine (SVM) has been optimized by Grey Wolf Optimization (GWO) method to analyze these data and determine the correlation between speed of sound in biodiesel and its related properties including pressure, temperature, molecular weight, and normal melting point. The results were very satisfactory because the values of statistical parameters R^2 and RMSE were obtained 1 and 1.4024, respectively. Here, this is the first time that the sensitivity analysis is used to estimate this target value. This analysis shows that the pressure widely affects the output values with relevancy factor 87.92. Also, our proposed method is highly accurate than other machine learning methods used in papers employed for this objective.

1. Introduction

In the future, the use of petroleum and fossil fuels will be limited [1]. A large number of studies have recently investigated biodiesel utilization within engines [2]. The need for oil imports would be reduced by using animal or agricultural sources to produce methyl esters. This can improve energy security and the local economy and lead to a satisfactory carbon emission balance [3]. Moreover, biodiesel combustion within a diesel engine typically emits smaller quantities of carbon [3–5].

Biodiesel can be extracted from several chemical compositions of feedstock. Due to the dependence of fat/oil structures in fatty acids on the source of oil/fat, biodiesel highly varies in physicochemical properties, including the cold-flow properties and cetane number [3]. Hence, the biodiesel type is considerably important in combustion and emissions [6].

It is important to identify better fatty acid compositions in order to enhance engine performance and diminish emis2sions. This has been studied by numerous works [7–11]. It is rational to relate the properties of biodiesel to some important oil characteristics, e.g., fatty acid composition, chain length, number of double bonds, unsaturation degree, and molecular weight [12–16]. Earlier works related the fatty acid composition and cetane number through regression models [12]. The use of different methods of artificial intelligence has been widely used in various sciences [17–21], and the cetane number was studied using artificial neural networks (ANNs) and multiple linear regression models [13]. Furthermore, temperature and density were associated in previous studies [22]. Some researchers related the cetane number, viscosity, density, and increased heating value to the number of double bonds and molecular weight [23]. The cetane number, oxidative stability, cold filter plugging point, and iodine value were related to the long-chain saturated factor and methyl ester unsaturation degree [24]. Some studies related the number of double bonds and the number of C atoms in the fatty acid [25].

Biodiesel utilization in engines was broadly investigated [26] under transient conditions [27]. Furthermore, broad examinations were performed statistically to find and study the impacts of biodiesel feedstock on emissions of engines [28] and fuel properties [29].

As an important and practical property, the present study focuses on the prediction of speed of sound in biodiesel in order to develop an accurate predictive correlation formulation based on the fatty acid composition through a multiple linear regression known as SVM-GWO. This very important property of biodiesel has received less attention from researchers, so we were looking for an accurate model to be able to estimate this functional property with high accuracy. In this paper, an extensive database has been used and an attempt has been made to evaluate the accuracy of this model using various analyses.

2. Materials and Methods

2.1. Support Vector Machine (SVM). The SVM is one of the machine learning (ML) methods. Rather than other techniques, this technique works based on a minimum of structural risk expressed by statistical theory [30]. This technique, for the first time, is proposed by Vapnik in 1992 for classification problems [31]. Afterward, it was developed by Cortes and Vapnik, in 1995 and 1997, for regression problem adaptation [32, 33]. SVM can be employed for both linear and nonlinear problems, but for nonlinear problems, it must be improved by kernel functions. The SVM equations are given in the following [34]. In Equation (1), a sample dataset is used for training using the SVM regression model, where y_i, x_i, d, and R are output, input, input space dimension, and output space, respectively.

$$\left\{x_i, y_i | x \in R^d, \quad y \in R, i = 1, 2, \cdots, n\right\}. \tag{1}$$

In Equation (2), the input data is mapped from R^d space to a high-dimension one, R^k $(k > d)$.

$$\psi(x) = (\phi(x_1), \quad \phi(x_2), \cdots, \phi(x_n)). \tag{2}$$

Equation (3) introduces the prediction model for SVM as follows:

$$f(x) = \omega^T \phi(x) + b, \quad \omega \in R^k, \cdots, \phi(x_n)), \tag{3}$$

where b, ω, and $f(x)$ are bias constant, weight, and a nonlinear mapping function, respectively, and ω and b are defined by Equation (4) with minimal structural risk.

$$\min R = \frac{1}{2}\|\omega\|^2 + c \times R_{\text{emp}}, \tag{4}$$

where $\|\omega\|^2$ is used to handle the difficulty of the model, c is the regularization coefficient, and R_{emp} is a function for handling errors. Also, for optimizing the objective function, R_{emp} is defined as the linear term of the error of SVM. So, Equation (4) can be changed into Equation (5) using the relaxation factors, ξ_i and ξ_i^*, and insensitivity loss function, ε:

$$\min J = \frac{1}{2}\|\omega\|^2 + c\sum_{i=1}^{n} \xi_i + \xi_i^* \begin{cases} y_i - \omega^T \phi(x_i) - b \le \varepsilon + \xi_i, \\ \omega^T \phi(x_i) + b - y_i \le \varepsilon + \xi_i, \\ \xi_i, \xi_i^* \ge 0 (i = 1, 2, \cdots, n). \end{cases} \tag{5}$$

Also, the Lagrange function is given by Equation (6) to solve the SVM error.

$$\begin{aligned} L(\omega, \xi_i, \xi_i^*, \alpha, \alpha^*, c, \beta, \beta^*) &= \frac{1}{2}\|\omega\|^2 + c\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{n=1}^{n} \alpha_i [\omega^T \phi(x_i) + b - y_i + \varepsilon + \xi_i] \\ &- \sum_{n=1}^{n} \alpha^*{}_i [y_i - \omega^T \phi(x_i) - b + \varepsilon + \xi_i^*] \\ &- \sum_{n=1}^{n} (\beta_i \xi_i + \beta_i^* \xi_i^*), \alpha_i, \alpha_i^*, \beta_i, \beta_i^* > 0, \end{aligned} \tag{6}$$

where α_i, α_i^*, β_i, $\beta_i^* \infty$ are defined as Lagrange factors. According to Karush-Kuhn-Tucker optimization conditions (Equation (7)) and symmetric kernel function (Equation (8)), the optimization problem can be obtained as Equation (9).

$$\begin{cases} \dfrac{\partial L}{\partial \omega} = 0 \longrightarrow \omega = \displaystyle\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \phi(x_i), \\ \dfrac{\partial L}{\partial b} = 0 \longrightarrow \omega \displaystyle\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0, \\ \dfrac{\partial L}{\partial \xi_i} = 0 \longrightarrow c - \alpha_i - \beta_i = 0, \\ \dfrac{\partial L}{\partial \xi_i^*} = 0 \longrightarrow c - \alpha_i^* - \beta_i^* = 0, \end{cases} \tag{7}$$

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j), \tag{8}$$

$$\begin{aligned} \max W(\alpha_i, \alpha_i^*) = &-\frac{1}{2}\sum_{ij=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) \\ &+ \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) y_i - \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \varepsilon \end{aligned} \tag{9}$$

$$\text{s.t.} \begin{cases} \displaystyle\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le c. \end{cases} \tag{10}$$

So, the SVM regression function is given by

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x_j) + b. \tag{11}$$

Table 1: Details of the implemented GWO-SVM algorithm.

Parameter	Value/comment
Kernel function	RBF
No. of train data	786
No. of test data	262
Optimization technique	GWO
C	52263.664
ε	0.5033
Γ	0.07825

The SVM method utilizes different kernel functions. In the current work, we used the radial basis kernel function (Equation (12)), where σ is representative of the width parameter of this function.

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right). \quad (12)$$

2.2. Grey Wolf Optimization (GWO). The GWO algorithm was introduced by Mirjalili et al. in 2014 as a novel metaheuristic algorithm inspired by the social hunting of grey wolves [35]. Generally, this algorithm follows four classes including (1) decision-making is performed by alpha (α) wolves about everything, (2) the alpha wolves are supported/consulted by beta (β) wolves, (3) the delta (δ) wolves must surrender to α and β wolves, and finally in (4) the other wolves are defined by omega (ω), which have to follow α and β orders. The ω wolves must help others whenever required [36, 37]. So, the hierarchy of power reduces from α to ω. In four classes, a specific optimization issue is defined by solutions of the GWO algorithm. So, α, β, and ω are the best solutions in this algorithm, and others are considered as ω. With this definition, the algorithm is updated in every iteration. The process of the algorithm follows these rules for prey: searching, surrounding, chasing, and attacking. The surrounding is given as follows:

$$X(t+1) = X_p(t) - A \times D, \quad (13)$$

where A, D, $X_p(t)$, $X(t+1)$, and t are the matrix coefficient, the distance between the prey and grey wolf, the position vector of each wolf, the next position of a grey wolf, and the current iteration whose calculations are given by

$$\begin{aligned} D &= \left|C \times X_p(t) - X(t)\right|, \\ A &= 2ar_1 - a, \\ C &= 2ar_2, \end{aligned} \quad (14)$$

where r_1, r_2 are the random vectors from 0 to 1.

In the hypersphere form, the relocation around the prey is feasible with the help of these equations. So, the ω wolves can update their positions as follows:

$$X(t+1) = \frac{X_1 + X_2 + X_3}{3}, \quad (15)$$

where X_1, X_2, and X_3 are defined as the following:

$$\begin{aligned} X_1 &= X_\alpha(t) - A_1 \times D_\alpha, \\ X_2 &= X_\beta(t) - A_2 \times D_\beta, \\ X_3 &= X_\delta(t) - A_3 \times D_\delta, \\ D_\alpha &= |C_1 \times X_\alpha(t) - X(t)|, \\ D_\beta &= \left|C_2 \times X_\beta(t) - X(t)\right|, \\ D_\delta &= |C_3 \times X_\delta(t) - X(t)|. \end{aligned} \quad (16)$$

2.3. Designing the GWO-SVM Model. Concerning the previous discussion, C, ε, and γ are used to handle the SVM performance. So, the GWO can be optimized by these factors. Table 1 depicts the characteristics of the GWO-SVM algorithm.

2.4. Gathering Data and Selecting Features. In this study, the database containing 1048 data with various variables related to the test system of the speed of sound in biodiesel— i.e., temperature, melting point, pressure, and molecular weight, has been collected from previously published papers. The source and range of inputs and output data are given elsewhere [38]. Three-quarters of the data are selected as training phase data and one-quarter of them are randomly separated as testing phase data.

3. Results and Discussion

In this section, we evaluate the ability of the proposed model to predict the target parameter, which is followed by various analyses.

3.1. Sensitivity Analysis (SA). In terms of exploring the impact of input data on the output, SA is defined as a mathematical method and used to determine useful priorities after the recognition of methodological errors and vital regions [39]. There are two forms for SA including local and global. The assessment of an input effect on results, while others are constant, is performed by local SA whereas the global SA evaluates the effect which stemmed from inputs on the outcome whenever changed [40].

The impact of input parameters on the speed of sound has been shown in Figure 1 that the most effective one is related to pressure with the relevancy factor of 87.92%.

Also, the relative factors of temperature, melting point, and molecular weight, with scores of −29.56%, −25.63%, and 15.18%, are not so big.

3.2. Outlier Analysis. Another statistical method used in this study is outlier diagnosis. This method is considered a fundamental method applied to determine datasets with different behavior from all data [41, 42]. It uses leverage statistical technique to find the outliers having parameters such as

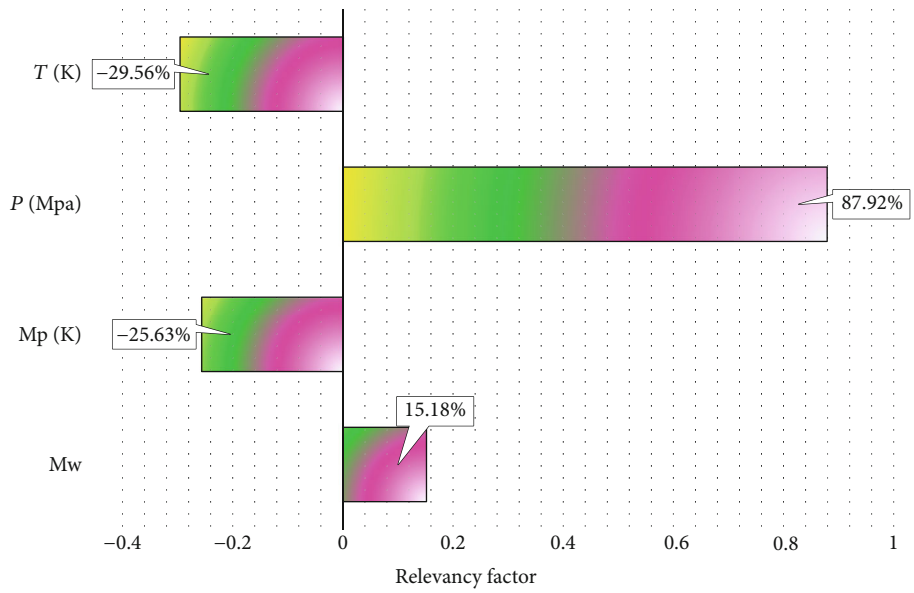


Figure 1: Sensitivity analysis on the input parameters.

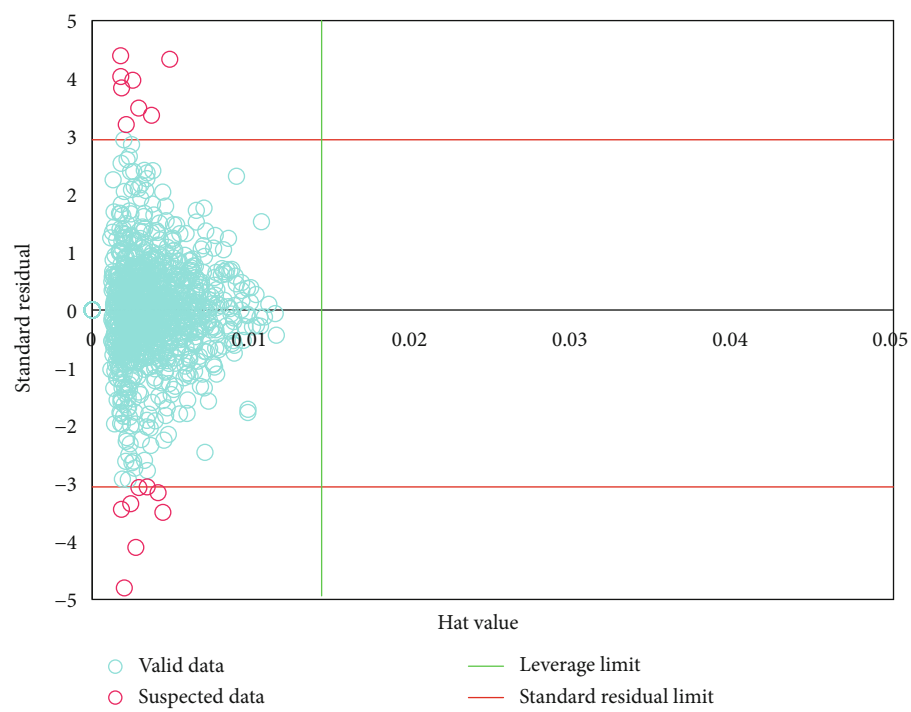


Figure 2: Outlier analysis to determine suspicious data points.

standardized residuals (R), critical leverage limit (H^*), and Hat indices (H) [43, 44]. H and H^* are defined as follows:

$$H = X\left(X'X\right)^{-1}X^t,$$
$$H^* = \frac{3n}{(p+1)}, \tag{17}$$

where X together with t are the two-dimensional $(n \times k)$ matrix and transpose matrix, respectively. Also, p and n are the numbers of input parameters and training points, respectively. Here, the likely Hat solutions include the main diagonal space of H. Also, Williams' plot, defined by R versus H, is used to determine the outlying candidates. Then, the feasible data region is introduced as a squared area to limit the warning leverage value on the horizontal and vertical axes and cut-off value, which is usually ±3, respectively. R and H are placed out of the valid area—i.e., $[-3, 3]$ and $[0, H^*]$—and classified as the outliers. Figure 2 depicts Williams' plot of GWO-SVM outputs. It is observed that most of the data values are placed in the valid area and the rest of them,

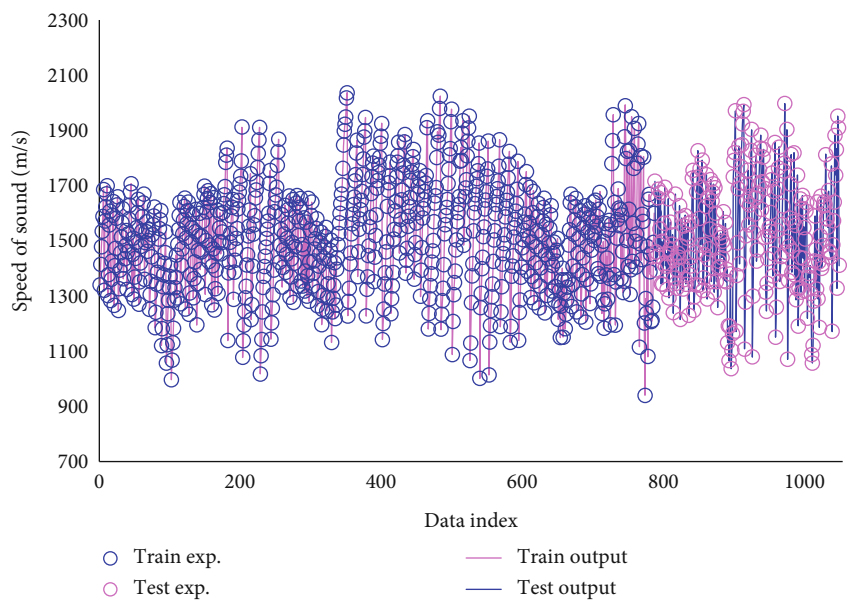


Figure 3: Observational comparison of real values and their corresponding modeled values for test and train data.

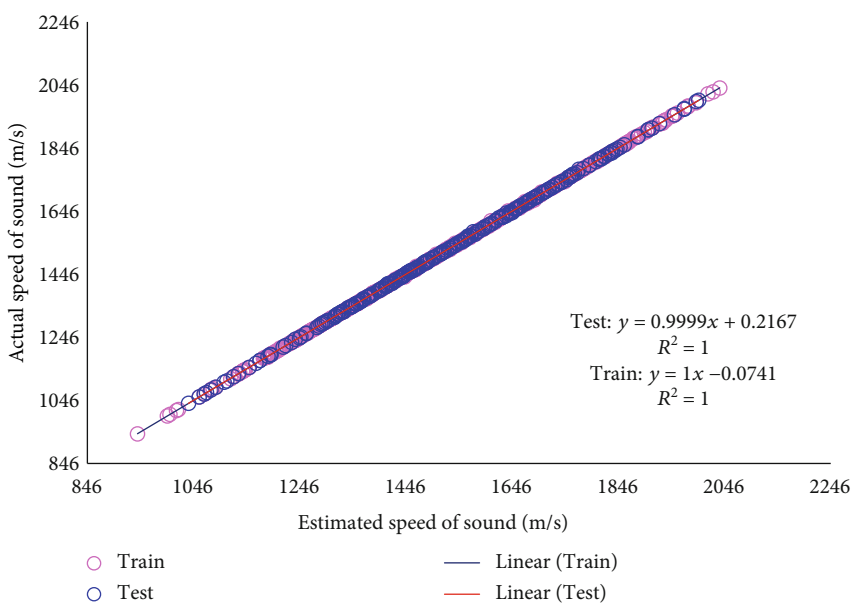


Figure 4: Regression plot to determine the accuracy of the proposed model in predicting actual values.

including 16 points, have a higher value than H^*. So it is demonstrated that the useful GWO-SVM algorithm can detect the inherent relationships between the speed of sound value and input parameters in addition to having a much more acceptable approach.

3.3. Model Assessment. The model assessment is performed by the speed of sound values that resulted in training and testing of the proposed model. Figure 3 shows these values versus the data index. It is proved that this model has the considerable capability to predict the speed of sound in biodiesel.

Also, to assess the accuracy of results with real values, the determination coefficient, R^2, is used and varies from 0 to 1. The R^2 values for testing and training the GWO-SVM dataset are 1 and 1, respectively. Thus, the accuracy of the predicted model is verified. The diagram of real values versus predicted values is shown in Figure 4.

The main part of the speed of sound in biodiesel values situates along the bisector line which shows how the GWO-SVM model is able to do prediction with high accuracy. Also, Figure 5 depicts the percentage of deviation for the GWO-SVM model which is not more than 0.6% that demonstrates the precision of the model.

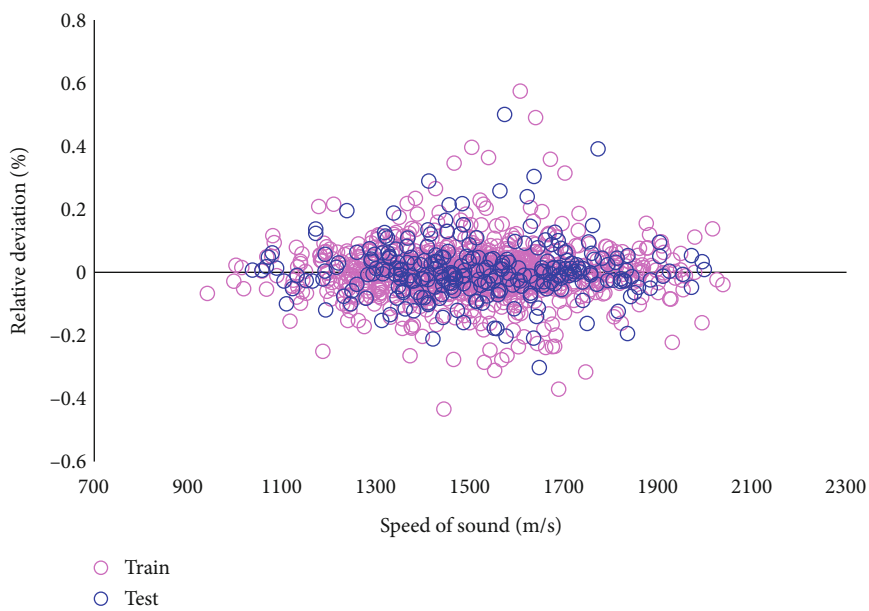


Figure 5: Relative deviation analysis to determine the accuracy of the SVM-GWO model.

Table 2: Statistical analyses based on the proposed SVM-GWO model.

Set	R^2	MRE (%)	MSE	RMSE	STD
Train	1.000	0.061	1.925445498	1.3876	1.0321
Test	1.000	0.061	1.966602271	1.4024	1.0542
Total	1.000	0.061	1.935734691	1.4024	1.0371

Table 3: Comparison between the accuracy of different models in predicting target outcomes.

Statistical parameter	SGB model	GP model	SVM-GWO model
R^2	0.99996	0.99803	1.000
RMSE	1.55	8.81907	1.4024

Also, Table 2 shows statistical analyses of the SVM-GWO model and verifies the accuracy of this model for predicting the speed of sound in biodiesel.

3.4. Comparison with Literature. Table 3 shows the comparison done between previously developed models (SGB and GP) by Abooali et al. [38] and our model for predicting the speed of sound in biodiesel. As it turns out, the model proposed in this study has a higher ability to predict output values because it has more R^2 and less RMSE compared to other models.

4. Conclusions

In this study, the SVM-GWO model has been proposed to investigate the effect of structural features on the performance of the speed of sound in biodiesel. The database containing large experimental data has been collected from previously published papers. Comparing all of the ML models, our model showed the best accuracy. So it has great capability to assist in the objective design of the speed of sound in biodiesel. Furthermore, it was shown that the pressure has the highest impact on the output values. In conclusion, according to the obtained maximum value of the coefficient of determination and minimum RMSE, our model is considered the most precise model to predict the speed of sound in biodiesel; therefore, it can be used to estimate this important property in related processes.

Data Availability

Data references are described in the text of the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. H. S. Dehaghani, M. S. Taleghani, M. H. Badizad, and R. Daneshfar, "Simulation study of the Gachsaran asphaltene behavior within the interface of oil/water emulsion: a case study," *Colloid and Interface Science Communications*, vol. 33, p. 100202, 2019.

[2] N. Nabipour, R. Daneshfar, O. Rezvanjou et al., "Estimating biofuel density via a soft computing approach based on intermolecular interactions," *Renewable Energy*, vol. 152, pp. 1086–1098, 2020.

[3] A. K. Agarwal, "Biofuels (alcohols and biodiesel) applications as fuels for internal combustion engines," *Progress in Energy and Combustion Science*, vol. 33, no. 3, pp. 233–271, 2007.

[4] A. Demirbas, "Biodiesel production from vegetable oils via catalytic and non-catalytic supercritical methanol transesterification methods," *Progress in Energy and Combustion Science*, vol. 31, no. 5-6, pp. 466–487, 2005.

[5] E. G. Giakoumis, C. D. Rakopoulos, A. M. Dimaratos, and D. C. Rakopoulos, "Exhaust emissions of diesel engines operating under transient conditions with biodiesel fuel blends," *Progress in Energy and Combustion Science*, vol. 38, no. 5, pp. 691–715, 2012.

[6] A. C. Hansen, D. C. Kyritsis, and C. F. F. Lee, "Characteristics of biofuels and renewable fuel standards," *Biomass to biofuels: strategies for global industries*, pp. 1–26, 2010.

[7] G. Knothe, ""Designer" biodiesel: optimizing fatty ester composition to improve fuel properties," *Energy & Fuels*, vol. 22, no. 2, pp. 1358–1364, 2008.

[8] S. Pinzi, D. Leiva, G. Arzamendi, L. M. Gandia, and M. P. Dorado, "Multiple response optimization of vegetable oils fatty acid composition to improve biodiesel physical properties," *Bioresource Technology*, vol. 102, no. 15, pp. 7280–7288, 2011.

[9] G. Knothe and K. R. Steidley, "Kinematic viscosity of biodiesel fuel components and related compounds. Influence of compound structure and comparison to petrodiesel fuel components," *Fuel*, vol. 84, no. 9, pp. 1059–1065, 2005.

[10] J. D. A. Rodrigues Jr., F. D. P. Cardoso, E. R. Lachter, L. R. Estevão, E. Lima, and R. S. Nascimento, "Correlating chemical structure and physical properties of vegetable oil esters," *Journal of the American Oil Chemists' Society*, vol. 83, no. 4, pp. 353–357, 2006.

[11] G. Knothe, "Improving biodiesel fuel properties by modifying fatty ester composition," *Energy & Environmental Science*, vol. 2, no. 7, pp. 759–766, 2009.

[12] A. Nomgboye and A. Hansen, "Prediction of cetane number of biodiesel fuel from the fatty acid methyl ester [FAME] composition," *International Agrophysics*, vol. 22, no. 1, pp. 21–29, 2008.

[13] R. Piloto-Rodríguez, Y. Sánchez-Borroto, M. Lapuerta, L. Goyos-Pérez, and S. Verhelst, "Prediction of the cetane number of biodiesel using artificial neural networks and multiple linear regression," *Energy Conversion and Management*, vol. 65, pp. 255–261, 2013.

[14] A. Ramadhas, S. Jayaraj, C. Muraleedharan, and K. Padmakumari, "Artificial neural networks used for the prediction of the cetane number of biodiesel," *Renewable Energy*, vol. 31, no. 15, pp. 2524–2533, 2006.

[15] A. Gopinath, S. Puhan, and G. Nagarajan, "Relating the cetane number of biodiesel fuels to their fatty acid composition: a critical study," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 223, no. 4, pp. 565–583, 2009.

[16] H. Yang, Z. Ring, Y. Briker, N. McLean, W. Friesen, and C. Fairbridge, "Neural network prediction of cetane number and density of diesel fuel from its chemical composition determined by LC and GC–MS," *Fuel*, vol. 81, no. 1, pp. 65–74, 2002.

[17] A. Erdal Tümer, S. Edebali, and Ş. Gülcü, "Modeling of removal of chromium (VI) from aqueous solutions using artificial neural network," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 39, no. 1, pp. 163–175, 2020.

[18] M. R. Ehsani, H. Bateni, and G. Razi Parchikolaei, "Modeling of oxidative coupling of methane over Mn/Na2WO4/SiO2 catalyst using artificial neural network," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 32, no. 3, pp. 107–114, 2013.

[19] R. Maachou, A. Lefkir, A. Bermad, and S. Abdelaziz, "Energy consumption modeling in activated sludge process using coupling PCA-ANFIS approach," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 38, no. 6, pp. 261–273, 2019.

[20] N. M. Ramli, M. A. Hussain, B. M. Jan, and B. Abdullah, "Online composition prediction of a debutanizer column using artificial neural network," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 36, no. 2, pp. 153–174, 2017.

[21] S. Ahadian, S. Moradian, F. Sharif, M. Amani Tehran, and M. Mohseni, "Prediction of time of capillary rise in porous media using artificial neural network (ANN)," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 26, no. 1, pp. 71–83, 2007.

[22] M. J. Pratas, S. V. D. Freitas, M. B. Oliveira, S. C. Monteiro, A. S. Lima, and J. A. P. Coutinho, "Biodiesel density: experimental measurements and prediction models," *Energy & Fuels*, vol. 25, no. 5, pp. 2333–2340, 2011.

[23] L. F. Ramírez-Verduzco, J. E. Rodríguez-Rodríguez, and A. del Rayo Jaramillo-Jacob, "Predicting cetane number, kinematic viscosity, density and higher heating value of biodiesel from its fatty acid methyl ester composition," *Fuel*, vol. 91, no. 1, pp. 102–111, 2012.

[24] M. J. Ramos, C. M. Fernández, A. Casas, L. Rodríguez, and Á. Pérez, "Influence of fatty acid composition of raw materials on biodiesel properties," *Bioresource Technology*, vol. 100, no. 1, pp. 261–268, 2009.

[25] M. Lapuerta, J. Rodríguez-Fernández, and E. F. De Mora, "Correlation for the estimation of the cetane number of biodiesel fuels and implications on the iodine number," *Energy Policy*, vol. 37, no. 11, pp. 4337–4344, 2009.

[26] C. Rakopoulos, K. A. Antonopoulos, D. C. Rakopoulos, D. T. Hountalas, and E. G. Giakoumis, "Comparative performance and emissions study of a direct injection diesel engine using blends of diesel fuel with vegetable oils or bio-diesels of various origins," *Energy Conversion and Management*, vol. 47, no. 18-19, pp. 3272–3287, 2006.

[27] C. D. Rakopoulos, A. M. Dimaratos, E. G. Giakoumis, and D. C. Rakopoulos, "Investigating the emissions during acceleration of a turbocharged diesel engine operating with bio-diesel or n-butanol diesel fuel blends," *Energy*, vol. 35, no. 12, pp. 5173–5184, 2010.

[28] E. G. Giakoumis, "A statistical investigation of biodiesel effects on regulated exhaust emissions during transient cycles," *Applied Energy*, vol. 98, pp. 273–291, 2012.

[29] E. G. Giakoumis, "A statistical investigation of biodiesel physical and chemical properties, and their correlation with the degree of unsaturation," *Renewable Energy*, vol. 50, pp. 858–878, 2013.

[30] İ. Güven and F. Şimşir, "Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods," *Computers & Industrial Engineering*, vol. 147, p. 106678, 2020.

[31] V. Vapnik, "Principles of risk minimization for learning theory," *Advances in neural information processing systems*, NIPS, 1992.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[33] V. Vapnik, S. E. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation,

and signal processing," *Advances in Neural Information Processing Systems*, pp. 281–287, 1997.

[34] S. Dai, D. Niu, and Y. Han, "Forecasting of power grid investment in China based on support vector machine optimized by differential evolution algorithm and grey wolf optimization algorithm," *Applied Sciences*, vol. 8, no. 4, p. 636, 2018.

[35] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.

[36] S. Mirjalili, S. Saremi, S. M. Mirjalili, and L. . S. Coelho, "Multi-objective grey wolf optimizer: a novel algorithm for multi-criterion optimization," *Expert Systems with Applications*, vol. 47, pp. 106–119, 2016.

[37] A. Naserbegi, M. Aghaie, and A. Zolfaghari, "Implementation of grey wolf optimization (GWO) algorithm to multi-objective loading pattern optimization of a PWR reactor," *Annals of Nuclear Energy*, vol. 148, p. 107703, 2020.

[38] D. Abooali, R. Soleimani, and S. Gholamreza-Ravi, "Characterization of physico-chemical properties of biodiesel components using smart data mining approaches," *Fuel*, vol. 266, p. 117075, 2020.

[39] A. Baghban, M. Kahani, M. A. Nazari, M. H. Ahmadi, and W. M. Yan, "Sensitivity analysis and application of machine learning methods to predict the heat transfer performance of CNT/water nanofluid flows through coils," *International Journal of Heat and Mass Transfer*, vol. 128, pp. 825–835, 2019.

[40] A. Baghban, A. H. Mohammadi, and M. S. Taleghani, "Rigorous modeling of CO_2 equilibrium absorption in ionic liquids," *International Journal of Greenhouse Gas Control*, vol. 58, pp. 19–41, 2017.

[41] R. Setiawan, R. Daneshfar, O. Rezvanjou, S. Ashoori, and M. Naseri, "Surface tension of binary mixtures containing environmentally friendly ionic liquids: insights from artificial intelligence," *Environment, Development and Sustainability*, pp. 1–22, 2021.

[42] R. Daneshfar, A. Bemani, M. Hadipoor et al., "Estimating the heat capacity of non-Newtonian ionanofluid systems using ANN, ANFIS, and SGB tree algorithms," *Applied Sciences*, vol. 10, no. 18, p. 6432, 2020.

[43] S. Alizadeh, I. Alruyemi, R. Daneshfar, M. Mohammadi-Khanaposhtani, and M. Naseri, "An insight into the estimation of drilling fluid density at HPHT condition using PSO-, ICA-, and GA-LSSVM strategies," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[44] F. Mousazadeh, M. H. T. Naeem, R. Daneshfar, B. S. Soulgani, and M. Naseri, "Predicting the Condensate Viscosity near the Wellbore by ELM and ANFIS-PSO Strategies," *Journal of Petroleum Science and Engineering*, vol. 204, p. 108708, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 5597222, 13 pages
https://doi.org/10.1155/2021/5597222

Research Article

An Extended Approach to Predict Retinopathy in Diabetic Patients Using the Genetic Algorithm and Fuzzy C-Means

Saeid Jafarzadeh Ghoushchi,[1] **Ramin Ranjbarzadeh**,[2] **Amir Hussein Dadkhah**,[1] **Yaghoub Pourasad**,[3] **and Malika Bendechache**[4]

[1]*Faculty of Industrial Engineering, Urmia University of Technology, Urmia, Iran*
[2]*Department of Telecommunications Engineering, Faculty of Engineering, University of Guilan, Rasht, Iran*
[3]*Department of Electrical Engineering, Urmia University of Technology, Urmia, Iran*
[4]*School of Computing, Faculty of Engineering and Computing, Dublin City University, Ireland*

Correspondence should be addressed to Saeid Jafarzadeh Ghoushchi; s.jafarzadeh@uut.ac.ir

Received 2 March 2021; Accepted 19 June 2021; Published 28 June 2021

Academic Editor: Alireza Baghban

The present study is developed a new approach using a computer diagnostic method to diagnosing diabetic diseases with the use of fluorescein images. In doing so, this study presented the growth region algorithm for the aim of diagnosing diabetes, considering the angiography images of the patients' eyes. In addition, this study integrated two methods, including fuzzy C-means (FCM) and genetic algorithm (GA) to predict the retinopathy in diabetic patients from angiography images. The developed algorithm was applied to a total of 224 images of patients' retinopathy eyes. As clearly confirmed by the obtained results, the GA-FCM method outperformed the hand method regarding the selection of initial points. The proposed method showed 0.78 sensitivity. The comparison of the fuzzy fitness function in GA with other techniques revealed that the approach introduced in this study is more applicable to the Jaccard index since it could offer the lowest Jaccard distance and, at the same time, the highest Jaccard values. The results of the analysis demonstrated that the proposed method was efficient and effective to predict the retinopathy in diabetic patients from angiography images.

1. Introduction

Patients suffering from diabetes, in some cases, lose their eyesight, which is often because of diabetic retinopathy (DR), though, there are some other causes of vision loss or poor eyesight, including other retinal and nonretina vision conditions such as macular degeneration and glaucoma (which are associated with age) and neuropathy vascular vision (which is Nonarteritic Anterior Ischemic Optic Neuropathy (NAION)) and cataract [1–4]. In addition, when an individual (a patient) with diabetes complains of visual disturbance, despite the 6.6's precision, the refractive errors, contrast sensitivity, direct light, and compliance range need to be taken into account. The doctors managing diabetic patients should consider these vision problems in order to make sure that timely referral and treatments are done to control vision impairment as much as possible. This control significantly affects peoples' daily life, particularly for activities like driving [5, 6].

Diabetic retinopathy can be considered a disease of the type of diabetes mellitus (DM), and it has been recognized as the primary basis for blindness in healthy adults in advanced countries. The retinal and choroidal circulation in the eye changes with the diabetic retinopathy. Multiple vascular dysfunctions may be the primary DR markers [7–9]. Also, Nonproliferative Diabetic Retinopathy (NPDR) can be defined by loss of capillary protein, thickening morphology of the glomerular basement membrane, and loss of smooth muscle cells, which leads to micro aneurysm, which is an essential indicator of the prediction of DR's progress [10, 11].

Additionally, in the foveal avascular zone (FAZ) and capillary outflow in florfenicol angiography (FA) studies,

patients with primary DR have been observed. Recently, OCT angiography has shown the presence of microenvironment and capillary excision in patients with diabetic retinopathy. Although DR is generally known as a retina, histology, electron microscopy, and green and endogenous angiography (ICGA), studies have shown that diabetes with degeneration of cholangiocarcinoma (CC) and also the formation of laminar deposition the term "diabetic cryotherapy " is used to refer to changes in the CC associated with diabetes [5, 12, 13]. Imaging of retinal and choroidal ducts in diabetic patients does not occur regularly, especially in patients referring to primary care centers, since color-based angiographic techniques, such as FA and ICGA, require an intravenous injection of external dyes [14].

In addition, considering its location and structure, CC visualization is challenging with both FA and ICGA. OCTA is a relatively new imaging technique that allows for indirect visualization of the retina and CC pathways. CATA is based on the contrast of the motor and, contrary to the FA and the ICGA, does not need to inject different colors [14, 15]. The OCTA B-scan is produced by acquiring OCT B-scans repeated in a fast series of a retina. If the texture is constant, repeat scans B are identical; however, if there is a flow of blood, moving erythrocytes create pixels in OCT B-scans that are converted to the amount of current flow rate. The volume of OCTA can be generated by acquiring redundant B-scans in several retina networks. Since the depth of the OCT depth has been resolved, the various vascular layers of the retina can be individually visualized, which is not possible with color-based angiography [2, 4, 16]. In addition, since it is undeniable, the OCTA can be repeated several times in subsequent reviews or even during the same visit. Finally, since OCTA is the standard OCT imaging format, structural information is obtained at the same time and is inherently recorded by angiography information. This feature provides a simultaneous representation of both structural and angiographic data [2, 4, 16].

Diabetes types I and II increase more not only in the elderly but also in the people suffering from obesity and also the young adolescents. Diabetes is a disease of multiple systems; thus, each part of the eye can be vulnerable. Although eye complications have been reported frequently by doctors for over two centuries, DR has played a significant role in its destructive impacts that result in the early loss of sight [15, 16]. DR is common today in the world, and prevention is difficult for people. The ophthalmologists typically detect the DR severity degree by various visual tests of the fundus through directly examining and evaluating the color photographs. A challenge in this sense is that this process is both time-consuming and costly. The diagnosis of DR and diagnosis of primary diseases remain somewhat personal, with statistics of agreement between trained specialists in studies previously recorded being different [2].

Additionally, 75% of patients with DR condition are currently living in underdeveloped or poor regions, where adequate doctors, medical health care facilities, and diagnostic infrastructures are not available. To control the rate of the increasing number of the DR patients which can lead to blindness all over the globe, many global screening solutions have been set up to deal with the spread of disease inside the eye and better maintain vision; however, such programs make extensive use of DR to diagnose and treat retinopathy and related problems on an individual basis in an efficient way [4, 15]. As a result, millions across the world are currently exposed to visual impairment without recognizing adequate eye care. For the aim of overcoming the current problematic conditions, automated solutions have been proposed for the diagnosis of retinal disease using a dyed fundus illustration. In fact, those methods that are based on only some obtained samples from different patients in different times in one medical clinic or screening center are not applicable to fundus images [2, 16]. This is due to the fact that various kinds of fundus cameras are used in different clinics for detecting eye dilatation. In addition, many of these algorithms are employed to manually extract the attributes related to the DR using some hand-crafted features, aimed at describing the prediction of anatomical structures in the fundus, such as optical discs, blood vessels, or macula. Though these hand-crafted representations might be run on the Foundation Individual Dataset, they are once again trying to accurately detect DR through fundus images that are tailor-made for different demographic purposes [15, 17]. The general-purpose features, comprising GLCM (Gray Level Cooccurrence Matrix) [18], GLRM (Gray Level Run Length Matrix) [19], and Histogram of Oriented Gradients (HOG) Histogram of Oriented Gradients, have been examined using nonspecific methods applicable to specifying DR properties; however, they have shown weaker and disproportionate properties, which cannot describe the nuances of retinopathy.

Regular screening of diabetic patients for DR is a highly technical and critical aspect of patient care. The care and timing of this care are very important for the cost of the treatment. If timely diagnosed, DR compensatory treatments are reachable, and this is a pivotal process for anyone who needs them. The DR classification involves weighting many features and the position of these features. This process is an arduous task for doctors and needs a lot of time. After training, systems can acquire much more rapid classifications, and this will enable them to help clinicians in a quick classification [17, 20, 21].

DR can be described by morphological lesions associated with abnormalities in the retinal blood flow. These lesions represent a regional distribution that encompasses risk factors in the initial phases of the disease and can predict disease progression. Moreover, this disease encounters two different threatening symptoms of Diabetic Maculopathy (DM) vision and Proliferative Diabetic Retinopathy (PDR) in the late stages of the disease. These symptoms can be identified through considering the difference in spreading and category of lesions [4, 20]. The progression of diabetic retinopathy is also associated with sectional changes in the retinal blood flow and the adjustment of the diameters of the retinal capillaries in the macular region and in the retinal environment. The literature consists of some research into the segmentation of images in medical research such as the fuzzy clustering method with residual driving and automatic fuzzy clustering methods [22, 23]. Algorithms that are without

supervision do not involve training networks; they are called clustering methods, e.g., Fast FCM, and Robust FCM [24–29].

On the other hand, methods with supervision (e.g., ANN and recurrent convolution neural network) train an optimum network for detection and segmentation of medical images [30–34]. Unsupervised methods usually are faster than training methods and are typically used as a preprocess for supervised methods. Methods such as FCM [21, 35], KNN [21, 36], and SVM [37] are commonly used as automated methods for generating ground truth data for automated methods. The advantage of automated methods like FCM is the detection of locations with distinguished color [29, 38]. In other words, tumors should be light or dark and different from other pixels. Gadekall et al. [39] suggested a deep neural network model based on a Grey Wolf Optimization (GWO) technique to classify the extracted features of diabetic retinopathy dataset. Employing the GWO can be beneficial to find optimal parameters for training the DNN structure. Behera and Chakravarty [40] utilized a scale-invariant feature transform (SIFT) approach to extract some key features. Also, to capture the exudate regions on each retinal images, Speeded Up Robust Features (SURF) has been employed.

In this study, the growth region technique based on the combination of the FCM and Genetic Optimizer Algorithm (GA) strategies for the aim of diagnosing diabetes is proposed, considering the angiography images of the patients' eyes. The main novelty of FCM and GA is the optimization of automated methods of clustering images. Several previous works are providing automated methods for the extraction of ground truth images. However, some are not flexible and need further improvement. The methods presented in this paper without using hand methods exploited a genetic algorithm to find the best initial seeds of FCM method in the growth region algorithm.

2. Extended Growth Region Method

2.1. Data Collection. The present paper was designed to assess the dimensions and shape of FAZ in patients with diabetic retinopathy in comparison with healthy controls, using OCT angiography. The database contains 224 images of a fluorescence-sized angiography, 250×250 depths of 8 bits per pixel, and each pixel size of 11×11 micrometers taken from the eye of 14 times over a different period. Among them, 12 are diabetic, and two are nondiabetic, with an average of 16 images per person (192 images from diabetic and 32 images from nondiabetic eyes). Indeed, 12 of the images are obtained from the right eye and 2 from the left eye. The Heidelberg Specialist apparatus performed imaging in fluorescing angiography from Urmia Hospital. These images are in the jpeg format. Figure 1 displays an instance of a database. The Heidelberg device is a prototype device with a bandwidth of 50 nm, a lateral resolution of 14 μm, and an axial resolution of 7 μm that obtain 85,000 A-scans per second. This device produced by Heidelberg Engineering (Heidelberg, Germany) and employs an amplitude decorrelation technique for applying to a volume scan on a $15 \times 10^\circ$ or $15 \times 5^\circ$ zone including an area with the size of 4.3×2.9 mm or 4.3×1.5 mm.

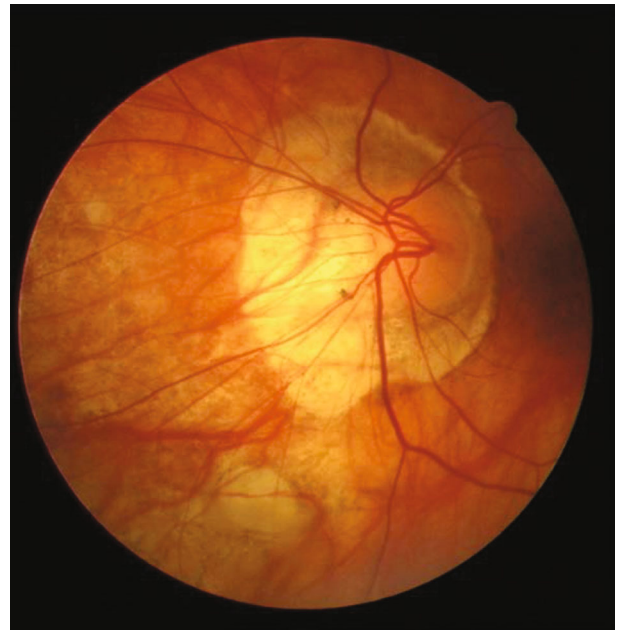


Figure 1: Example of angiography image.

2.2. Image Processing. In recent years, image processing has been extensively used, especially with the advent of advanced techniques such as discriminatory information processing, e.g., digital cameras and scanners. On the other hand, the images resulted from these techniques are generally associated with different degrees of noise, and even sometimes, these techniques fail to fade boundary inside an image [33, 41–43]. This problem finally decreases the resultant image resolution. In this context, image processing refers to all operations and techniques adopted by users for the purpose of decreasing the defects and increasing the image quality.

The region's growth method is an attempt for splitting the image into discrete regions on the basis of the degree of homogeneity or similarity between two or more parts of the image in neighboring pixels; thus, at another level, it depends on the criteria applied to homogeneity analysis [44, 45]. The pixels in each area are gathered together using some specific criteria: illumination, color, etc. This growing-base strategy is a simple method in the category of area-based methods and is based on the testing the intensity of an initial pixel with all touching pixels and adds them to the first pixel to search again for finding other pixels that belong to the segment. In the image segmentation context, the histogram-based methods are only focused upon the distribution of image pixels in the gray level, whereas the local growth strategies consider that the neighboring pixels also possess adjacent gray levels [44, 46].

In the following, the way the area-based methods work is explained step by step [47, 48]:

(1) The number of initial seeds is the beginning of the algorithm

(2) With the use of these seeds, the regions start their growth, and the pixels that resemble the original pixels will be added to that area

(3) Once the growth of area stops, the subsequent grain is taken into consideration, and the next area growth continues

(4) The above-mentioned steps will be continued until all of the pixels that exist in the image belong to one area

The area's growth method has the following steps (Figure 2).

2.3. Selecting Initial Seeds. For the algorithm to get started, the initial seeds need to be manually entered. In this state, the algorithm begins operating by choosing the user's initial points. Several methods in the field have the capacity of taking the start locations without having any prior knowledge about it. For instance, a random step strategy can be employed to explore the first points [47, 49].

For the aim of choosing the initial points, the present study proposes a new model using the combination of fuzzy clustering (FCM) and genetic optimization (GA) approaches. As in a FCM method every pixel inside the image can belong to more than a cluster, it helps to detect the border of any objects more accurately. Initially, a clustering process is carried out on the input image by applying the fuzzy clustering technique. This clustering method can be implemented based on the defining the membership grade F_M and cluster centers F_C. So, these parameters can be selected based on a trial and error method or by and optimization approach [29, 50]. To address the problem of experimenting all possible values for the parameters to obtain the maximum segmentation accuracy, the GA responsible for minimizing the cost function (sum of squares of the error) and determining the best value for these parameters. The cost function E is demonstrated in Equation (1) [51, 52]:

$$E(C) = \sum_{j=1}^{n}\sum_{i=1}^{K} m_{ij}\left\|x_j - c_i\right\|_A^2, \tag{1}$$

where the relation m can be expressed as follows:

$$\begin{aligned} 0 < \sum_{j=1}^{n} m_{ij} < n, \quad & j = 1, 2, \cdots, K, \\ \sum_{i=1}^{k} m_{ij} < 1, \quad & j = 1, 2, \cdots, n, \\ \sum_{i=1}^{k}\sum_{j=1}^{n} m_{ij} = n. & \end{aligned} \tag{2}$$

Moreover, m and C are computed employing Equation (3), as follows:

$$m_{ij} = \left[\sum_{k=1}^{K}\left(\frac{\left\|x_j - c_i\right\|_A}{\left\|x_k - c_i\right\|_A}\right)\right]^{-1}, \quad 1 \le i \le K, 1 \le j \le n, \tag{3}$$

$$c_i = \frac{\sum_{j=1}^{n} m_{ij} x_j}{\sum_{j=1}^{n} m_{ij}}, \quad 1 \le i \le K. \tag{4}$$

2.4. Determine the Similarity of Regions. After the determination of the initial points during the above step, the criterion of similarity is chosen for the regions. The objective of using this criterion is to examine all pixels around the new attached pixels to decide whether they can be added to the working area or not. This procedure specifies the attribution of the novel pixel to the corresponding area [53, 54].

A similarity criterion, which is widely used in this field, is the standard deviation that can be applied to each evaluating segment, meaning the new pixel $I_{n+1}(x, y)$ needs to be added in the segment area if it can pass the following condition:

$$\mu_n - X\sigma_n < I_{n+1}(x + y) < \mu_n + X\sigma_n, \tag{5}$$

where σ_n stands for the standard deviation, μ_n demonstrates the mean, and X illustrates a weighting parameter used for defining how many pixels are different within the region. In general, in case $X = 3$, more likely all pixels (about 99.70 percent) of the evaluating locations (pixels) are chosen in the same segmented area. It should be mentioned that whatever the value of X is smaller, the chance of finding more similar pixels in a vicinity is lower. In another word, by defining the low and high values for the parameter X, the number of the segmented area inside the image will be higher and lower, respectively. In addition, another criterion that needs to be considered is the image mean level (256 possible intensity levels). For doing this, the mean of the area which has been segmented by this technique is calculated, and then the intensity value of the new evaluating pixel $I_{n+1}(x, y)$ for assessing based on the Equation (6) is extracted. This means that Equation (6) must be true for adding the new pixel in the segmented area before.

$$\mu_n - X < I_{n+1}(x + y) < \mu_n + X. \tag{6}$$

2.5. Growth Region. When the initial seeds are determined for the algorithm commencement and also the similarity criterion is determined for pixels with areas, then, the area growth process gets started. The area growth, which starts from the initial seeds, is done through choosing the neighboring regions [53, 54].

2.5.1. Fuzzy C-Means (FCM). Grouping similar data or pixels in the same group (segmented data or pixels) using machine learning techniques based on some criteria can be defined as clustering [38, 55–59]. This diving data can be implemented by many approaches such as DBSCAN, fuzzy C-means, mean shift, and K-means algorithms. In the fuzzy C-means and K-means approaches, first, a number of groups/classes need to be defined, but it is not true when working with the mean shift and DBSCAN techniques (algorithms calculate the optimum number of the clusters) [43, 60].

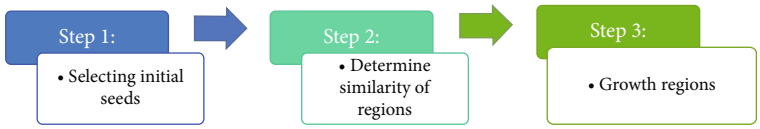


Figure 2: The region growth method steps.

The FCM method is based on finding the similarity between data points inside the image to define the groups using an unsupervised manner. Unlike some approaches such as K-means and C-means, in this method, each point is belonging to some clusters based on a weighting parameter. The objective function that needs to be minimized using FCM strategy is demonstrated in Equation (7) [38]:

$$E = \sum_{k=1}^{m} \sum_{i=1}^{N} \mathrm{mem}_{ki} \|\mathrm{center}_k - \mathrm{pixel}_i\|^2, \tag{7}$$

$$\mathrm{Mem}_{ki} = \frac{1}{\left(\sum_{j=1}^{m} (\mathrm{center}_k - \mathrm{pixel}_i) / (\mathrm{center}_j - \mathrm{pixel}_i)\right)^t}, \tag{8}$$

$$\sum_{k=1}^{m} \mathrm{mem}_{ki} = 1, \quad \mathrm{mem}_{ki} \epsilon [0, 1], \tag{9}$$

where pixel_i stands for the ith sample of I, center_k refers to the centre of the kth cluster, mm_{ki} stands for the membership score for the ith sample that relates to the kth divided group of pixels (clusters), m depicts the number of clusters, and N is a number for demonstrating how many pixels present in image I.

2.5.2. Genetic Algorithm (GA). Object detection or object recognition can be conducted by extracting key features of the image related to each possible object in the image. It means each pixel inside the image based on adjacent pixels can extract a lot of local or global features even by changing the image representation such as Local Directional Pattern (LDP) [61, 62] or Local Binary Pattern (LBP) [41, 63]. So, by investigating these features, we can find the structure and shape of each object in the image. This investigating process can be conducted by an optimization process to decrease the time of the consideration [64–67]. The genetic optimization algorithm was designed on the basis of the evolution theory and the survival of the fittest or natural selection proposed by Charles Darwin. This popular algorithm has been extensively applied to optimization problems [64]. In the genetic algorithm, at each implementation stage, random processing is performed on a group of search spots. It means a lot of random initial points are evaluated for finding the best possible points to apply in the next step. This way, each point is assigned with a sequence of traits, and then, the sequences are exposed to genetic operators. Next, based on minimizing the cost function (or increasing the fitness function), these obtained sequences are divided to separate parts and then each part adds to another part to create a new chain of spots within the searching space until stopping criteria are reached. The task of dividing and then adding sequences is done by considering their participation probability [68–70]. In the present research, the cost function is computed based on the difference between the input image and the image attained by the region growing technique. It gets started from the initial random point, as can be expressed in the following relation:

$$\cos t = (\mathrm{GM} - I_1)^2 + (\mathrm{WM} - I_2)^2 + (\mathrm{BM} - I_3)^2, \tag{10}$$

where WM, BM, and GM demonstrate the white matter, black matter, and gray matter, respectively, and I_1, I_2, and I_3 represent the images acquired from segmentation process.

2.5.3. Performance Analysis. The Specificity, Precision, Accuracy, False positive rate (FPR), Sensitivity, Relative volume difference (RVD), Volume overlap error (VOE), and Dice similarity (DICE) are the eight evaluation metrics that were employed to assess the result of the proposed model (can be calculated using Equations (11)–(18)). Sensitivity or true positive rate refers to the percentage of the important objects that identified correctly. Specificity refers to the percentage of the unimportant objects (healthy tissue) identified as unimportant correctly. Accuracy (measure of statistical bias) and precision (measure of statistical variability) represent the closeness of the measurements to a predefined value and to each other, respectively [38, 41, 71, 72].

(i) True Positive (TP): the important object (retinopathy) is recognized perfectly.

(ii) True Negative (TN): the healthy tissue is recognized perfectly.

(iii) False Positive (FP): the important object is recognized with mistakes.

(iv) False Negative (FN): the healthy tissue is recognized with mistakes.

$$\text{TPR or Sensitivity} = \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{TP}}, \tag{11}$$

$$\text{TNR or Specificity} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}, \tag{12}$$

$$\text{PPV or Precision} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP}}, \tag{13}$$

$$\text{FPR or False positive rate} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{FN}}, \tag{14}$$

$$\text{ACC or Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\text{All results}}, \tag{15}$$

$$\mathrm{Dice}(\mathrm{set}_1, \mathrm{set}_2) = 100 \times \left(\frac{2 \times |\mathrm{set}_1 \cap \mathrm{set}_2|}{|\mathrm{set}_1 + \mathrm{set}_2|}\right), \quad (16)$$

$$\mathrm{VOE}(\mathrm{set}_1, \mathrm{set}_2) = 100 \times \left(1 - \frac{\mathrm{set}_1 \cap \mathrm{set}_2}{\mathrm{set}_1 \cup \mathrm{set}_2}\right), \quad (17)$$

$$\mathrm{RVD}(\mathrm{set}_1, \mathrm{set}_2) = 100 \times \left(\frac{\mathrm{set}_1 - \mathrm{set}_2}{\mathrm{set}_2}\right), \quad (18)$$

where set_1 and set_2 represent the segmentation and ground-truth results, respectively.

In a given test, specificity and sensitivity are dependent upon the nature of the test and the type of test used. This is worth mentioning that a test result is not interpretable using Specificity and Sensitivity only. Receiver Operating Characteristic curve (ROC curve) is a probability curve that refers to a graphical plot (TPR or true positive rate against the FPR or False positive rate; FPR is on the x-axis and TPR is on y-axis) for checking any classification model's performance at various thresholds settings.

3. Experimental Results and Discussions

As the region growing algorithm is based on finding the homogeneity defined by calculating the pixel intensity statistics, the first step in the region growing is to identify a set of seed positions (initial set of small areas). The initial region starts as the exact location of these seeds that are based on some user criterion or can be selected based on the generating random number. The regions are then grown from these seed points to other points in the vicinity of pixels depending on a region membership criterion (pixel intensity). To determinate whether the new point is good enough to join the selected seed location (point) or not, the mean and standard deviation of the growth region need to be computed, as depicted in Equations (16) and (17).

Here, we use 8-connected neighborhood for our pixel's adjacent relationship to grow from the seed points. By examining the value of the pixels in the vicinity of the seed locations, all tested pixels are categorized into (1) seed points and (2) background. It is an iterated process until there are no changes in two successive iterative stages. The number and location of these initial location are determined by GA. By examining the random values for both the location and number of the seed points obtained by GA, the algorithm calculates cost function for each of them. Then, those locations that reach the best result can be selected for the final segmentation using region growing approach. This process is expressed in the following relations:

$$\mu_N = \frac{(N-1)(\mu_{N-1}) + I_N}{N}, \quad (19)$$

$$\sigma_N = \sqrt{\frac{(N-2)(\sigma_{N-1})^2 + (N/N-1)(I_N - \mu_N)^2}{N-1}}. \quad (20)$$

Afterwards, for the neighboring points that have been inserted to this class, they will also be found in the same neighborhood, and then the update of the parameters is done. The searching process is continued until the detection of the first class is completely done, and no other pixel is appended.

The numbers of clusters in FCM are 6-10, and for the genetic algorithm, the population is fixed at 20, and the mutation rate is set to 0.2. Figure 3 demonstrates the results of the clustering process performed with the use of the proposed FCM-GA method. As clearly shown, the yellow areas (some similar parts of the image) are decreasing inside the loop until reaching the best possible solution. Regarding Figure 3(a), the retina target region is almost above the right part of the image with high contrast. After classification using the FCM method, the results of the membership function are depicted as can be seen in Figure 3. The final image could detect the precise situation of the retina target region. The confusion matrix for testing the dataset is demonstrated in Figure 4.

The TPR (Sensitivity), ACC (accuracy), TNR (Specificity), PPV (Precision), and FPR (False positive rate) values using some approaches are described in Table 1. As it is clearly shown, our approach gains the best scores among all evaluated methods. The Deep membrane and Improved U-NET obtain the significant value in ACC criteria. The worst result (except FPR) was reached by the Artificial Neural Network (ACC = 83.4, TPR = 93.8, TNR = 44.7, PPV = 86.5). Besides, the PPV values of the Improved U-NET and Deep membrane are considerably higher as compared to the Ant colony, Pixel-based Segmentation, Morphological Watershed, Artificial Neural Network, and FCM results. For the TPR values, there is a small difference between Deep membrane and PCNN models. The FPR and TPR values of the Deep membrane model is partly similar to the proposed algorithm; however, Deep membrane approach has a meaning unlike TNR. Also, the Pixel-based Segmentation method represents the worst result in terms of the FPR. The Ant colony, Pixel-based Segmentation, Morphological Watershed, Artificial Neural Network, and FCM methods obtain the worst results in term of TNR, whilst their TPR values are acceptable.

Table 2 illustrates the evaluation of our segmentation technique and the results from recently published models. The attained values of RVD measurement imply the amount of oversegmentation or undersegmentation. Therefore, a zero score stands for the best possible segmentation result, whilst a negative value represents the segmented result image is not as large as that of the corresponding actual annotated image. DICE equals one for a precise segmentation. The Dice scores of the Deep membrane and PCNN models are the same and are partly similar to those of the proposed algorithm; however, these two techniques have a significantly different RVD. Also, the Deep membrane and Pixel-based Segmentation algorithms show the undersegmentation results with -2.46 and -5.64 values, respectively. Besides, the RVD values of the Artificial Neural Network and Morphological Watershed models were considerably higher as compared to our outcomes. Our strategy and Ant colony algorithm have significantly different RVDs [19]. Considering the VOE, the lowest and highest values belong to our model and Artificial Neural Network, respectively. Moreover,

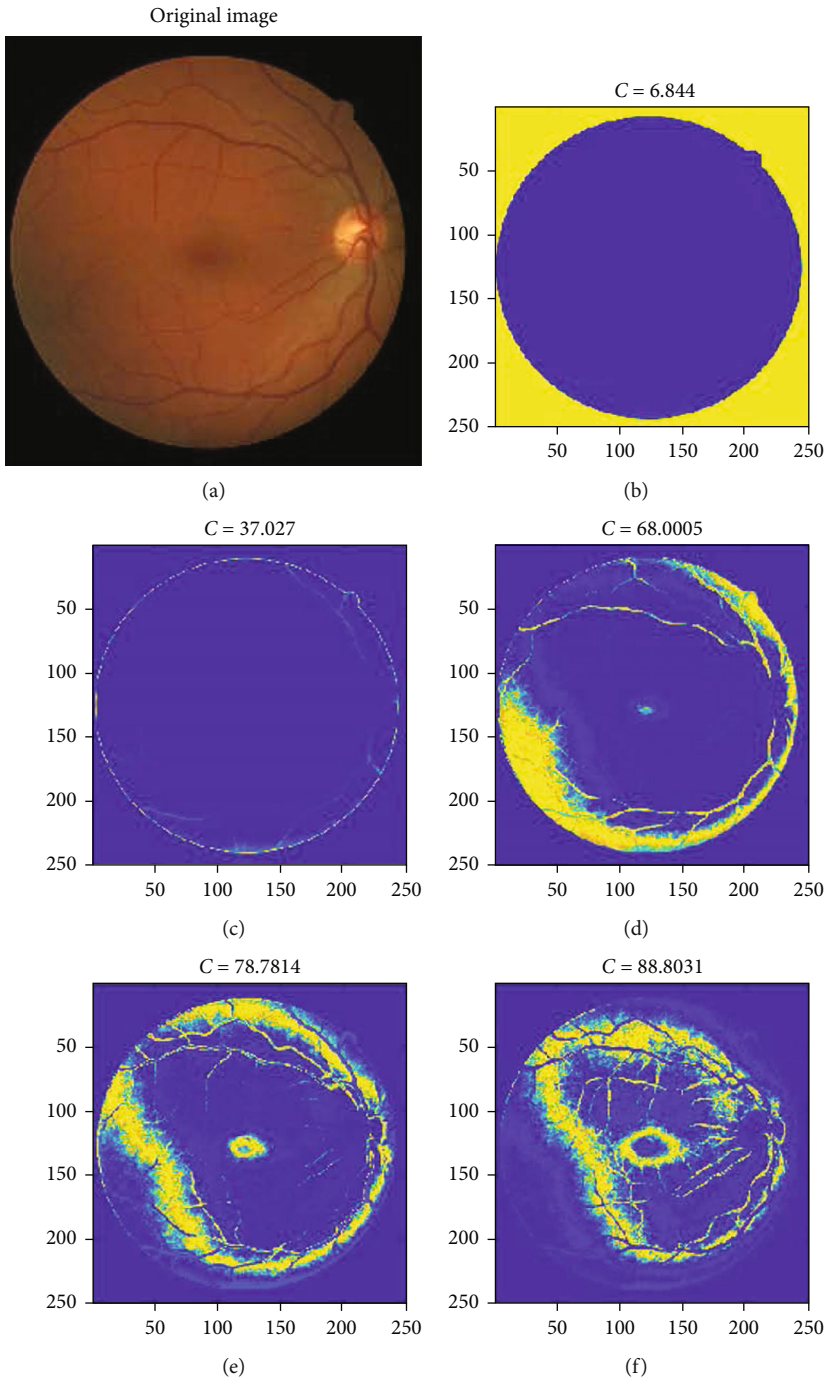


Figure 3: Continued.

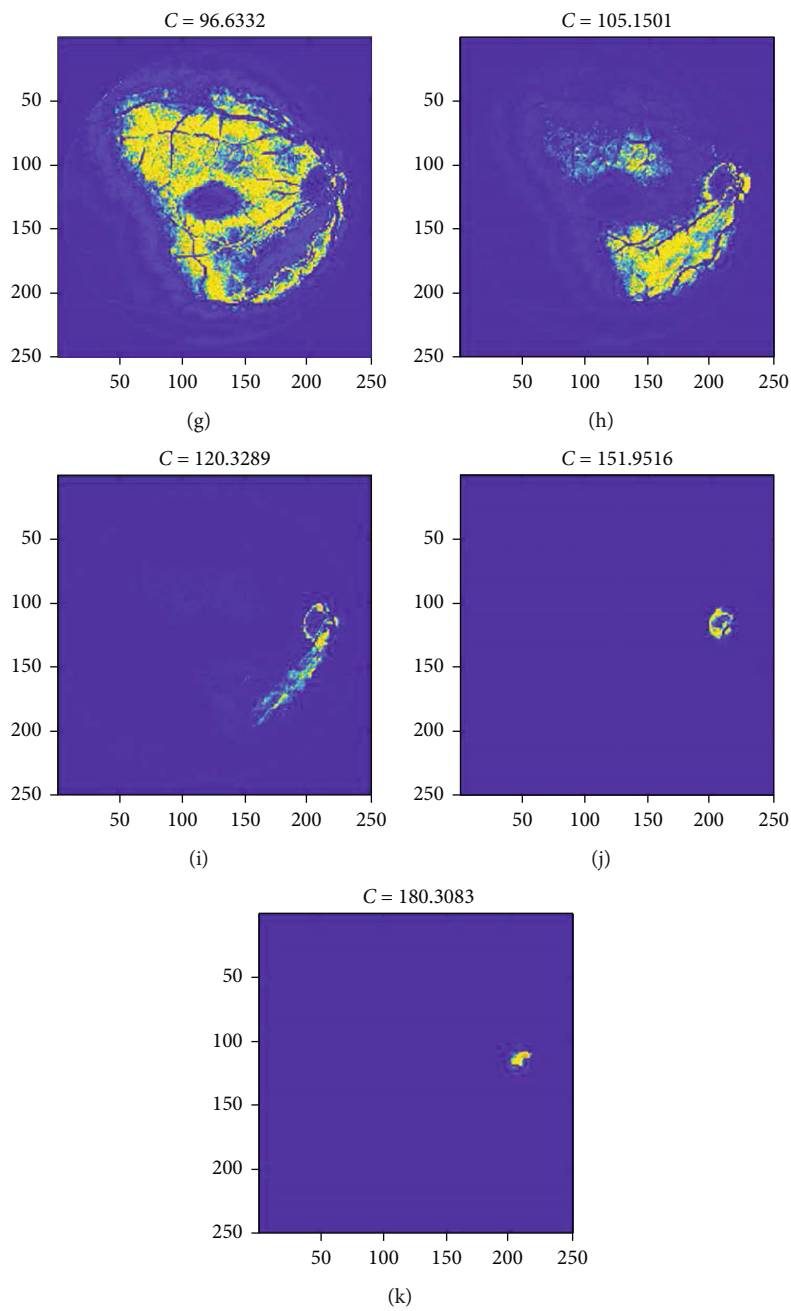


Figure 3: The results of the presented method for detection of retinopathy.

there are minimum differences between VOEs in Ant colony, Pixel-based Segmentation, and Artificial Neural Network techniques.

Therefore, the division of an ideal abnormal area into angiography images plays a vital role in medical imaging programs. Segmentation algorithms are applied to promoting the precision and function in case of images in medical contexts. The method proposed in this paper was applied to a number of images, and also, the actual image analyses were compared to each other in order to examine the results of the reading. This study conducted comparative research on various performance criteria. Generally, in the GA classification, FCM returned more accurate results. Extending this work to improve GA segmentation with some other algorithms, without changing the primary nature of medical images, is required.

As the obtained results indicate obviously, the proposed method performed successfully in reducing the VOE score. It indicates the fact that the image has been split with higher precision. In the present research, the initial points were segregated automatically with the use of the region growth method. In addition, a genetic algorithm was employed for the purpose of selecting the most proper initial points in an automatic way. The GA approach was utilized, and the best possible fitness function was defined for the aim of image segmentation. This way, the proper initial

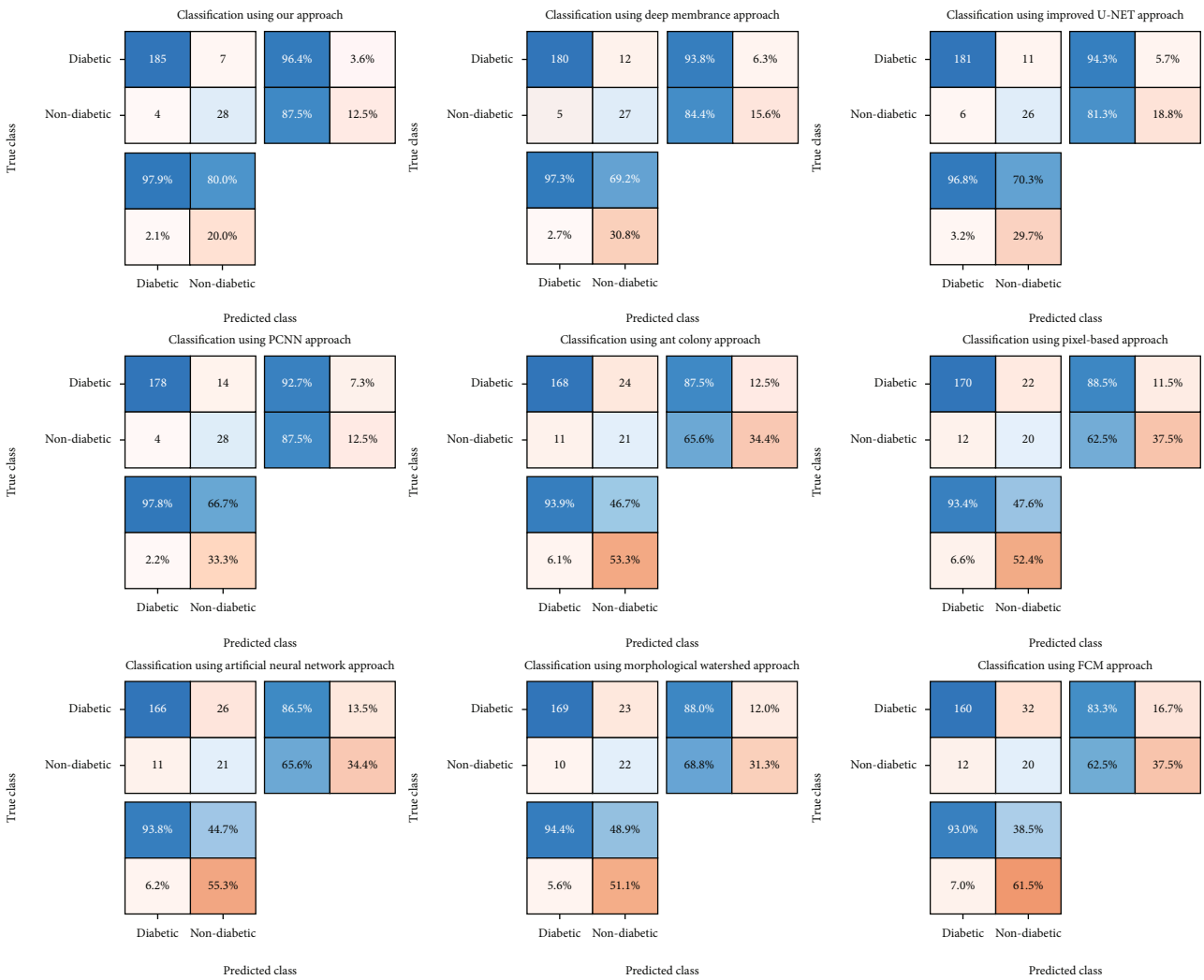


Figure 4: Demonstration of confusion matrix for our approach and 8 other methods.

Table 1: Comparative outcomes of the proposed strategy and recently published studies.

Method	ACC	TPR	TNR	PPV	FPR
Deep membrane [4]	*92.4%*	97.3%	69.2%	93.8%	84.4%
Improved U-NET [6]	*92.4%*	96.8%	70.3%	94.3%	81.3%
PCNN model [2]	*91%*	97.8%	66.7%	92.7%	87.5%
Ant colony algorithm [73]	*84.3%*	93.9%	46.7%	87.5%	65.6%
Pixel-based Segmentation [1]	*84.8%*	93.4%	47.6%	88.5%	62.5%
Artificial Neural Network [74]	*83.4%*	93.8%	44.7%	86.5%	65.6%
Morphological Watershed [75]	*85.2%*	94.4%	48.9%	88%	68.8%
Presented FCM	*80.3%*	93%	38.5%	83.3%	62.5%
Presented FCM+GA	*95%*	97.9%	80%	96.4%	85.4%

locations were determined to get the algorithm started. At the final step, the images were exposed to the proposed algorithm, and the obtained outcomes were compared to those of the growth method of the area wherein the manually chosen points were chosen. Findings confirmed the capability of our technique in decreasing the fragmentation errors.

In the present study, a total of 224 images of the retina were employed, which included 192 images from diabetic and 32 images from nondiabetic eyes. We used 70%, 20%, and 10% of total data for training, validation, and test, respectively. The proposed method showed the highest level of sensitivity among all. That indicates that GA can be used more effectively in selecting initial points.

Table 2: Quantitative comparative outcomes for segmentation of the retinopathy. This evaluation is conducted between our model and baseline studies. The assessments are based on Relative volume difference (RVD), Volume overlap error (VOE), and Dice similarity (DICE).

Method	Dice	RVD (%)	VOE (%)
Deep membrane [4]	92%	-2.46	6.16
Improved U-NET [6]	90%	3.74	5.78
PCNN model [2]	92%	3.55	6.41
Ant colony algorithm [73]	88%	-5.64	7.94
Pixel-based Segmentation [1]	88%	-4.91	7.42
Artificial Neural Network [74]	89%	5.37	8.12
Morphological Watershed [75]	87%	4.96	7.76
Presented FCM	86%	4.07	7.39
Presented FCM+GA	94%	2.32	4.28

4. Conclusions

In the present paper, a segmentation approach using the growth region technique is represented. This model is based on the combination of the FCM and GA methods for the purpose of diagnosing diabetes from the angiography images of the patients' eyes. The algorithm started with the early locations inside the image, and the mean of these locations were assumed as the mean of the objective area, and the early value for standard deviation was assumed as zero. For detecting the extract target area, the retinopathy images were recommended with the employed dynamic image analysis on the basis of the genetic algorithm. The eight evaluation metrics including Specificity, Precision, Accuracy, False positive rate (FPR), Sensitivity, Relative volume difference (RVD), Volume overlap error (VOE), and Dice similarity (DICE) were employed to assess the result of the proposed structure. The fuzzy fit function in GA was compared to another technique, and the outcomes indicated that the proposed method was more suitable to the sensitivity and Dice measures, with the maximum sensitivity and dice index. Considering the VOE index, the lowest and highest values are related to the suggested model and Artificial Neural Network, respectively. Additionally, there are minimum differences between VOEs in the Artificial Neural Network, Pixel-based Segmentation, and Ant colony techniques.

The fuzzy fitness function c-means evaluate the segmentation of threshold-based physical fuzzy tools, the fuzzy C-means tool in GA. The method proposed in the GA classification returned more proper results in its general performance, extending this to improve GA segmentation with some other algorithms, without changing the nature of the main medical angiography images is required. As the results demonstrated, the method introduced in this study was successful in decreasing the value of the VOE and RVD indexes. It indicates that the image has been split with higher precision.

Furthermore, a genetic optimizer was employed in this study in order to apply the proper early locations for the segmentation task automatically. The suitable early points were identified to start the algorithm with the use of GA and through giving a definition for the proper fitness function for the image segmentation purposes. At the final step, the prepared images were exposed to the proposed algorithm, and the obtained outcomes were compared with the growth manner of the region where the points are chosen manually. The obtained results confirmed the high capacity of our proposed algorithm in terms of reducing fragmentation errors.

Data Availability

The dataset is available online: http://www.med.harvard.edu/AANLIB/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. Al-Antary, M. Hassouna, Y. Arafa, and R. Khalifah, "Automated identification of diabetic retinopathy using pixel-based segmentation approach," in *ACM International Conference Proceeding Series, hlm*, pp. 16–20, Association for Computing Machinery, New York, NY, USA, 2019.

[2] T. J. Jebaseeli, C. A. Deva Durai, and J. D. Peter, "Retinal blood vessel segmentation from diabetic retinopathy images using tandem PCNN model and deep learning based SVM," *Optik*, vol. 199, p. 163328, 2019.

[3] Y. Moharamzad, A. H. Davarpanah, A. Yaghobi Joybari et al., "Diagnostic performance of apparent diffusion coefficient (ADC) for differentiating endometrial carcinoma from benign lesions: a systematic review and meta-analysis," *Abdominal Radiology*, vol. 46, no. 3, pp. 1115–1128, 2021.

[4] J. Xue, S. Yan, J. Qu et al., "Deep membrane systems for multitask segmentation in diabetic retinopathy," *Knowledge-Based Systems*, vol. 183, p. 104887, 2019.

[5] T. J. Jebaseeli, C. A. D. Durai, and J. D. Peter, "Segmentation of retinal blood vessels from ophthalmologic diabetic retinopathy images," *Computers and Electrical Engineering*, vol. 73, pp. 245–258, 2019.

[6] Q. Li, S. Fan, and C. Chen, "An intelligent segmentation and diagnosis method for diabetic retinopathy based on improved U-NET network," *Journal of Medical Systems*, vol. 43, no. 9, pp. 1–9, 2019.

[7] M. He, W. Wang, H. Yu et al., "Comparison of expression profiling of circular RNAs in vitreous humour between diabetic retinopathy and non-diabetes mellitus patients," *Acta Diabetologica*, vol. 57, no. 4, pp. 479–489, 2020.

[8] A. Khan, I. N. Petropoulos, G. Ponirakis, and R. A. Malik, "Visual complications in diabetes mellitus: beyond retinopathy," *Diabetic Medicine*, vol. 34, no. 4, pp. 478–484, 2017.

[9] K. A. Ponto, J. Koenig, T. Peto et al., "Prevalence of diabetic retinopathy in screening-detected diabetes mellitus: results from the Gutenberg Health Study (GHS)," *Diabetologia*, vol. 59, no. 9, pp. 1913–1919, 2016.

[10] S. P. Mortensen, K. M. Winding, U. W. Iepsen et al., "The effect of two exercise modalities on skeletal muscle capillary ultrastructure in individuals with type 2 diabetes," *Scandinavian Journal of Medicine & Science in Sports*, vol. 29, no. 3, pp. 360–368, 2019.

[11] L. Perrone, T. S. Devi, K. I. Hosoya, T. Terasaki, and L. P. Singh, "Thioredoxin interacting protein (TXNIP) induces inflammation through chromatin modification in retinal capillary endothelial cells under diabetic conditions," *Journal of Cellular Physiology*, vol. 221, no. 1, pp. 262–272, 2009.

[12] A. Mirshahi, F. Ghassemi, K. Fadakar, R. Mirshahi, F. Bazvand, and H. Riazi-Esfahani, "Effects of panretinal photocoagulation on retinal vasculature and foveal avascular zone in diabetic retinopathy using optical coherence tomography angiography: a pilot study," *Journal of Current Ophthalmology*, vol. 31, no. 3, pp. 287–291, 2019.

[13] M. Niestrata-Ortiz, P. Fichna, W. Stankiewicz, and M. Stopa, "Enlargement of the foveal avascular zone detected by optical coherence tomography angiography in diabetic children without diabetic retinopathy," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 257, no. 4, pp. 689–697, 2019.

[14] C. Yu, S. Xie, S. Niu et al., "Hyper-reflective foci segmentation in SD-OCT retinal images with diabetic retinopathy using deep convolutional neural networks," *Medical Physics*, vol. 46, no. 10, pp. 4502–4519, 2019.

[15] X. Li, L. Shen, M. Shen, F. Tan, and C. S. Qiu, "Deep learning based early stage diabetic retinopathy detection using optical coherence tomography," *Neurocomputing*, vol. 369, pp. 134–144, 2019.

[16] P. Brata Chanda and S. K. Sarkar, "Automatic identification of blood vessels, exaudates and abnormalities in retinal images for diabetic retinopathy analysis," *SSRN Electronic Journal*, vol. 1, pp. 1–9, 2019.

[17] L. Ye, W. Zhu, S. Feng, and X. Chen, "GANet: group attention network for diabetic retinopathy image segmentation," in *Medical Imaging 2020: Image Processing*, pp. 1–13, Texas, USA, 2020.

[18] J. Rahebi and F. Hardalaç, "Retinal blood vessel segmentation with neural network by using gray-level co-occurrence matrix-based features patient facing systems," *Journal of Medical Systems*, vol. 38, no. 8, pp. 1–12, 2014.

[19] B. K. Wardani, N. E. Belinda, R. Rulaningtyas, and E. Purwanti, "Application of gray level run length matrices features extraction for diabetic retinopathy detection based on artificial neural network," in *The 2nd International Conference On Physical Instrumentation And Advanced Materials 2019*, Indonesia, 2020.

[20] M. M. Khansari, J. Zhang, Y. Qiao et al., "Automated deformation-based analysis of 3D optical coherence tomography in diabetic retinopathy," *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 236–245, 2020.

[21] D. Palani, K. Venkatalakshmi, A. Reshma Jabeen, and V. M. Arun Bharath Ram, "Effective detection of diabetic retinopathy from human retinal fundus images using modified FCM and IWPSO," in *2019 IEEE International Conference on System, Computation, Automation and Networking, ICSCAN*, Institute of Electrical and Electronics Engineers Inc, India, 2019.

[22] S. Dorosti, M. Fathi, S. J. Ghoushchi, M. Khakifirooz, and M. Khazaeili, "Patient waiting time management through fuzzy based failure mode and effect analysis," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 2, pp. 2069–2080, 2020.

[23] H. Torabi Dashti, A. Masoudi-Nejad, and F. Zare, "Finding exact and solo LTR-retrotransposons in biological sequences using SVM," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 31, no. 2, pp. 111–116, 2012.

[24] Y. Ganjkhanlou, A. Bayandori Moghaddam, S. Hosseini, T. Nazari, A. Gazmeh, and J. Badraghi, "Application of image analysis in the characterization of electrospun nanofibers," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 33, no. 2, pp. 37–45, 2014.

[25] S. Jafarzadeh Ghoushchi, M. Khazaeili, A. Amini, and E. Osgooei, "Multi-criteria sustainable supplier selection using piecewise linear value function and fuzzy best-worst method," *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 2, pp. 2309–2325, 2019.

[26] E. Kamari, A. A. Hajizadeh, and M. R. Kamali, "Experimental investigation and estimation of light hydrocarbons gas-liquid equilibrium ratio in gas condensate reservoirs through artificial neural networks," *Iranian journal of chemistry and chemical engineering*, vol. 39, no. 6, pp. 163–172, 2020.

[27] B. Kavitha and D. Sarala Thambavani, "Artificial neural network optimization of adsorption parameters for Cr(VI), Ni(II) and Cu(II) ions removal from aqueous solutions by riverbed sand," *Iranian journal of chemistry and chemical engineering*, vol. 39, no. 5, pp. 203–223, 2020.

[28] H. A. Purwanithami, C. Atika Sari, E. H. Rachmawanto, and D. R. I. M. Setiadi, "Hemorrhage diabetic retinopathy detection based on fundus image using neural network and FCM segmentation," in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 45–49, Institute of Electrical and Electronics Engineers Inc. Indonesia, 2020.

[29] W. Wiharto and E. Suryani, "The analysis effect of cluster numbers on fuzzy c-means algorithm for blood vessel segmentation of retinal fundus image," in *2019 International Conference on Information and Communications Technology, ICOIACT*, pp. 106–110, Institute of Electrical and Electronics Engineers Inc, Indonesia, 2019.

[30] A. Azari, M. Shariaty-Niassar, and M. Alborzi, "Short-term and medium-term gas demand load forecasting by neural networks," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 31, no. 4, pp. 77–84, 2012.

[31] Z. Hosseini-Dastgerdi and S. Jafarzadeh-Ghoushchi, "Investigation of asphaltene precipitation using response surface methodology combined with artificial neural network," *Journal of Chemical and Petroleum Engineering*, vol. 2019, no. 2, pp. 153–167, 2019.

[32] N. M. Ramli, M. A. Hussain, B. M. Jan, and B. Abdullah, "Online composition prediction of a debutanizer column using artificial neural network," *Iranian Journal of Chemistry and Chemical Engineering (IJCCE)*, vol. 36, no. 2, pp. 153–174, 2017.

[33] R. Ranjbarzadeh, A. Bagherian Kasgari, S. Jafarzadeh Ghoushchi, S. Anari, M. Naseri, and M. Bendechache, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," *Scientific Reports*, vol. 11, no. 1, p. 10930, 2021.

[34] S. Salem and S. Jafarzadeh-Ghoushchi, "Estimation of optimal physico-chemical characteristics of nano-sized inorganic blue pigment by combined artificial neural network and response surface methodology," *Chemometrics and Intelligent Laboratory Systems*, vol. 159, pp. 80–88, 2016.

[35] S. Anbazhagan, V. Thiruvengadam, and K. Kulanthai, "Adaptive neuro-fuzzy inference system and artificial neural network modeling for the adsorption of methylene blue by novel adsorbent in a fixed-bed column method," *Iranian journal of chemistry and chemical engineering*, vol. 39, no. 6, pp. 75–93, 2020.

[36] S. Bandyopadhyay, S. Choudhury, S. K. Latib, D. K. Kole, and C. Giri, "Gradation of diabetic retinopathy using KNN classifier by morphological segmentation of retinal vessels," in *Advances in Intelligent Systems and Computing, hlm*, pp. 189–198, Springer, Verlag, 2018.

[37] T. Vandarkuzhali and C. S. Ravichandran, "Detection of fovea region in retinal images using optimisation-based modified FCM and ARMD disease classification with SVM," *International Journal of Biomedical Engineering and Technology*, vol. 32, no. 1, pp. 83–107, 2020.

[38] R. Ranjbarzadeh and S. B. Saadi, "Automated liver and tumor segmentation based on concave and convex points using fuzzy c-means and mean shift clustering," *Measurement*, vol. 150, p. 107086, 2020.

[39] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, and G. Srivastava, "Deep neural networks to predict diabetic retinopathy," *Journal of Ambient Intelligence and Humanized Computing*, vol. 1, pp. 1–14, 2020.

[40] M. K. Behera and S. Chakravarty, "Diabetic Retinopathy Image Classification Using Support Vector Machine," in *2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020, hlm*, Institute of Electrical and Electronics Engineers Inc, India, 2020.

[41] N. Karimi, R. Ranjbarzadeh Kondrood, and T. Alizadeh, "An intelligent system for quality measurement of golden bleached raisins using two comparative machine learning algorithms," *Measurement: Journal of the International Measurement Confederation*, vol. 107, pp. 68–76, 2017.

[42] Y. Pourasad, R. Ranjbarzadeh, and A. Mardani, "A new algorithm for digital image encryption based on chaos theory," *Entropy*, vol. 23, no. 3, p. 341, 2021.

[43] R. Ranjbarzadeh, S. Baseri Saadi, and A. Amirabadi, "LNPSS: SAR image despeckling based on local and non-local features using patch shape selection and edges linking," *Measurement*, vol. 164, p. 107989, 2020.

[44] A. F. A. Fadzil, S. Ibrahim, and N. E. A. Khalid, "Blood vessels segmentation of retinal fundus image via wStack-based object-oriented region growing," *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*, , pp. 31–37, Springer, Singapore, 2019.

[45] A. Hamzenejad, S. J. Ghoushchi, V. Baradaran, and A. Mardani, "A robust algorithm for classification and diagnosis of brain disease using local linear approximation and generalized autoregressive conditional heteroscedasticity model," *Mathematics*, vol. 8, no. 8, p. 1268, 2020.

[46] M. V. Maheswari and G. Murugeswari, "A survey on computer algorithms for retinal image preprocessing and vessel segmentation," in *2020 Proceedings of the 5th International Conference on Inventive Computation Technologies, ICICT, hlm*, pp. 403–408, Institute of Electrical and Electronics Engineers Inc, India, 2020.

[47] R. Panda, N. B. Puhan, and G. Panda, "New binary Hausdorff symmetry measure based seeded region growing for retinal vessel segmentation," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 119–129, 2016.

[48] E. Rodrigues, A. Conci, and P. Liatsis, "ELEMENT: multimodal retinal vessel segmentation based on a coupled region growing and machine learning approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3507–3519, 2020.

[49] N. D. Salih, M. D. Saleh, C. Eswaran, and J. Abdullah, "Fast optic disc segmentation using FFT-based template-matching and region-growing techniques," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, vol. 6, no. 1, pp. 101–112, 2018.

[50] R. Ranjbarzadeh and S. Baseri Saadi, "Corrigendum to "Automated liver and tumor segmentation based on concave and convex points using fuzzy c-means and mean shift clustering" [Measurement 150 (2020) 107086]," *Measurement*, vol. 151, p. 107230, 2020.

[51] H. Huang, F. Meng, S. Zhou, F. Jiang, and G. Manogaran, "Brain image segmentation based on FCM clustering algorithm and rough set," *IEEE Access*, vol. 7, pp. 12386–12396, 2019.

[52] K. Soppari and N. S. Chandra, "Development of improved whale optimization-based FCM clustering for image watermarking," *Computer Science Review*, vol. 37, p. 100287, 2020.

[53] N. Jothiaruna, K. Joseph Abraham Sundar, and B. Karthikeyan, "A segmentation method for disease spot images incorporating chrominance in comprehensive color feature and region growing," *Computers and Electronics in Agriculture*, vol. 165, p. 104934, 2019.

[54] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, "Machine learning and region growing for breast cancer segmentation," in *2019 International Conference on Advanced Science and Engineering, ICOASE*, pp. 88–93, Institute of Electrical and Electronics Engineers Inc. Iraq, 2019.

[55] M. Bendechache and M. T. Kechadi, "Distributed clustering algorithm for spatial data mining," in *ICSDM 2015 - Proceedings 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services, hlm*, pp. 60–65, Institute of Electrical and Electronics Engineers Inc, China, 2015.

[56] M. Bendechache, M. T. Kechadi, and N. A. Le-Khac, "Efficient large scale clustering based on data partitioning," in *Proceedings -3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, hlm*, pp. 612–621, Institute of Electrical and Electronics Engineers Inc, Canada, 2016.

[57] M. Bendechache, A. K. Tari, and M. T. Kechadi, "Parallel and distributed clustering framework for big spatial data mining," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 34, no. 6, pp. 671–689, 2019.

[58] R. Geetha Ramani and L. Balasubramanian, "Macula segmentation and fovea localization employing image processing and heuristic based clustering for automated retinal screening," *Computer Methods and Programs in Biomedicine*, vol. 160, pp. 153–163, 2018.

[59] H. Xia, S. Deng, M. Li, and F. Jiang, "Robust retinal vessel segmentation via clustering-based patch mapping functions," in *2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM*, pp. 520–523, Institute of Electrical and Electronics Engineers Inc. China, 2017.

[60] M. Bendechache, *Study of Distributed Dynamic Clustering Framework for Spatial Data Mining*, Theses and Dissertations of University College Dublin, Irland, 2019, http://oatd.org/oatd/record?record=handle%5C%3A10197%5C%2F10614.

[61] R. Srinivasa Perumal and P. V. S. S. R. Chandra Mouli, "Dimensionality reduced local directional pattern (DR-LDP) for face recognition," *Expert Systems with Applications*, vol. 63, pp. 66–73, 2016.

[62] P. P. Sarangi, B. S. P. Mishra, and S. Dehuri, "Fusion of PHOG and LDP local descriptors for kernel-based ear biometric recognition," *Multimedia Tools and Applications*, vol. 78, no. 8, pp. 9595–9623, 2019.

[63] T. Tuncer, S. Dogan, and F. Ozyurt, "An automated residual exemplar local binary pattern and iterative relief based COVID-19 detection method using chest X-ray image," *Chemometrics and Intelligent Laboratory Systems*, vol. 203, p. 104054, 2020.

[64] H. Liu, R. Zhai, J. Fu, Y. Wang, and Y. Yang, "Optimization study of thermal-storage PV-CSP integrated system based on GA-PSO algorithm," *Solar Energy*, vol. 184, pp. 391–409, 2019.

[65] H. Moayedi, M. Raftari, A. Sharifi, W. A. W. Jusoh, and A. S. A. Rashid, "Optimization of ANFIS with GA and PSO estimating α ratio in driven piles," *Engineering with Computers*, vol. 36, no. 1, pp. 227–238, 2020.

[66] M. Shirmohammadi, S. J. Goushchi, and P. M. Keshtiban, "Optimization of 3D printing process parameters to minimize surface roughness with hybrid artificial neural network model and particle swarm algorithm," *Progress in Additive Manufacturing*, vol. 6, no. 2, pp. 199–215, 2021.

[67] S. Zandevakili, M. R. Akhondi, R. Hosseini, and S. Mohammad, "Leaching optimization of Sarcheshmeh copper concentrate by application of Taguchi experimental design method," *Iranian journal of chemistry and chemical engineering*, vol. 39, no. 6, pp. 229–236, 2020.

[68] A. Khan, Z. u. Rehman, M. A. Jaffar et al., "Color image segmentation using genetic algorithm with aggregation-based clustering validity index (CVI)," *Signal, Image and Video Processing*, vol. 13, no. 5, pp. 833–841, 2019.

[69] L. Khrissi, N. El Akkad, H. Satori, and K. Satori, "Image segmentation based on K-means and genetic algorithms," in *Advances in Intelligent Systems and Computing, hlm*, pp. 489–497, Springer, Singapore, 2020.

[70] F. Torres, B. Escalante-Ramirez, J. Olveres, and P. L. Yen, "Lesion detection in breast ultrasound images using a machine learning approach and genetic optimization," in *Pattern Recognition and Image Analysis*, pp. 289–301, Springer, Cham, 2019.

[71] R. Gholipour Peyvandi and S. Z. Islami Rad, "Precise prediction of interface distribution of materials in multiphase separation facilities using a low-cost and simple technique: ANN," *Iranian journal of chemistry and chemical engineering*, vol. 39, no. 5, pp. 285–291, 2020.

[72] H. Mirshahvalad, R. Ghasemiasl, N. Raufi, and M. Malekzadeh Dirin, "A neural networks model for accurate prediction of the flash point of chemical compounds," *Iranian journal of chemistry and chemical engineering*, vol. 39, no. 4, pp. 297–304, 2020.

[73] T. SELÇUK and A. ALKAN, "Detection of microaneurysms using ant colony algorithm in the early diagnosis of diabetic retinopathy," *Medical Hypotheses*, vol. 129, p. 109242, 2019.

[74] A. T. Nair and K. Muthuvel, "Blood vessel segmentation and diabetic retinopathy recognition: an intelligent approach," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, vol. 8, no. 2, pp. 169–181, 2020.

[75] P. Siva Kalyani and G. Sasikala, "Morphological watershed approach for the analysis of diabetic nephropathy," in *Lecture Notes in Electrical Engineering*, pp. 547–554, Springer, Singapore, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 6069010, 13 pages
https://doi.org/10.1155/2021/6069010

Research Article

Comprehensive Modeling in Predicting Biodiesel Density Using Gaussian Process Regression Approach

Bingxian Wang [1] **and Issam Alruyemi** [2]

[1] *School of Mathematics and Statistics, Huaiyin Normal University, Huaian, Jiangsu 223300, China*
[2] *Fouman Faculty of Engineering, College of Engineering, University of Tehran, Fouman, Iran*

Correspondence should be addressed to Issam Alruyemi; essamkhudur@ut.ac.ir

Received 17 May 2021; Accepted 8 June 2021; Published 17 June 2021

Academic Editor: Alireza Baghban

In this study, four Gaussian process regression (GPR) approaches by various kernel functions have been proposed for the estimation of biodiesel density as the functions of pressure, temperature, molecular weight, and the normal melting point of fatty acid esters. Comparing the actual values with GPR outputs shows that these approaches have good accuracy, but the performance of the rational quadratic GPR model is better than others. In this GPR model, RMSE = 0.47, MSE = 0.22, MRE = 0.04, $R^2 = 1$, and STD is equal to 0.3. In addition, for the first time, this study shows that the effective parameters affect the biodiesel density. According to this analysis, it was shown that among the input parameters, pressure has the greatest effect on the target values with a relevancy factor of 0.59. This study can be used as a suitable and valuable work/tool for chemical and petroleum engineers who attempt environment protection and recovery improvement.

1. Introduction

Recently, in various countries, the issue of energy is deep and complex [1–3]. The greenhouse gas concentration in the atmosphere has been increased because the consumption of fossil fuels increased, and in that case, the earth's temperature increased too [4, 5]. For the solution to these problems, several agreements have been signed to reduce greenhouse gas emissions [6]. There are so many reasons but the main reasons for blooming the renewable energy resources are the aim of controlling the emission of pollutants and the implementation and credibility of these agreements [7, 8]. The two things that mainly can form biodiesel are the oils that come from vegetables and the fat of an animal's body [9]. Biodiesel is a combination of fatty acid alkyl esters [10]. Ethanol and methanol are some of the alcohols that the oils that come from vegetables and the fat of animal's body are transesterification catalytically by them [10]. Biodiesel is a clean fuel to burn, and it is not a toxic fuel because of low concentrations of sulfur. In that case, this fuel has the minimum bad effect on the emission of greenhouse gas changes, and also unlike fossil fuels, the biodiesel fuel has a less negative effect on our environment too [11, 12]. Additionally, in diesel engines, we can use biodiesel fuels lonely or use this fuel with fossil diesel because it improves engine life [13], although this fuel costs more than petroleum-based diesel and has higher viscosity too. Some other disadvantages of this fuel are lower oxidation stability, higher cloud point, and lower energy content in comparison with petroleum-based diesel [14]. The observational and modeling studies for the establishment of these properties become important because the concentration on usages of the biofuels and their properties have been grown [15, 16]. For example, the density of biodiesel is the important property that has a significant matter in the thermophysical process. One of the important topics in diesel fuels is the investigation of density because diesel fuels have so many technical and economic parts for the usage of the fuel and also the environmental effects [17–19].

For estimating the properties of fossil fuels, lots of investigations have been carried out in the literature and also there are different advancements in this study, such as a new way for guessing many properties of fuels consisting of surface

Table 1: Obtained statistical parameters to evaluate the performance of the models.

Model	Phase	R^2	MRE (%)	MSE	RMSE	STD
GPR (exponential)	Train	1.000	0.01	0.01	0.11	0.10
	Test	1.000	0.05	0.38	0.61	0.45
	Total	1.000	0.02	0.10	0.61	0.29
GPR (Matern)	Train	1.000	0.04	0.18	0.42	0.27
	Test	1.000	0.04	0.25	0.50	0.33
	Total	1.000	0.04	0.19	0.50	0.28
GPR (squared exponential)	Train	1.000	0.04	0.21	0.46	0.30
	Test	1.000	0.04	0.23	0.48	0.32
	Total	1.000	0.04	0.22	0.48	0.30
GPR (rational quadratic)	Train	1.000	0.04	0.22	0.47	0.30
	Test	1.000	0.04	0.22	0.47	0.30
	Total	1.000	0.04	0.22	0.47	0.30

tension and viscosity has been proposed by Queimada and his colleagues [20]. Barati-Harooni and his colleagues estimate interfacial tension between oil and brine with developed least-squares support vector machine in terms of pressure, salinity, and temperature [21]. Rostami and his colleagues correlate interfacial tension of hydrocarbon and water using genetic programming (GP) with an R-squared of 0.910 [22]. The Kay's model is a model that can determine the density of biodiesels that was estimated by Pratas and his colleagues [17]. The alkane density with an average absolute relative deviation of 60% was predicted by Gahk and his colleagues [23]. Miraboutalebi with the help of his coworkers estimate cetane numbers with a root mean squared error (RMSE) of 2.530 and R-squared of 0.950 by implementing an artificial neural network (ANN) [24]. On the other hand, the cetane number in terms of fatty acid methyl ester (FAME) was predicted by Mostafaei with developing the adaptive neuro-fuzzy inference system [25].

For biodiesel properties, there are some investigational searches in the literature. For example, the density of biodiesels was measured by Paratas et al. at atmospheric pressure with ten individual samples in temperatures between 278.150 and 373.170 K [17]. The viscosities and densities of three individual mixtures of methylcyclohexane and fatty acid methyl esters were obtained by Li and his colleagues in the atmosphere in temperatures between 293.150 and 324.150 K [26]. The density of soybean oil biodiesel was experimentally determined by Aitbelale and his colleagues at temperatures between 298.150 and 393.150 K and pressures up to 140.0 MPa [27]. The surface tension values were measured by Aitbelale et al. for three different biodiesels in temperature and pressure of 473.0 K and 7.0 MPa [27].

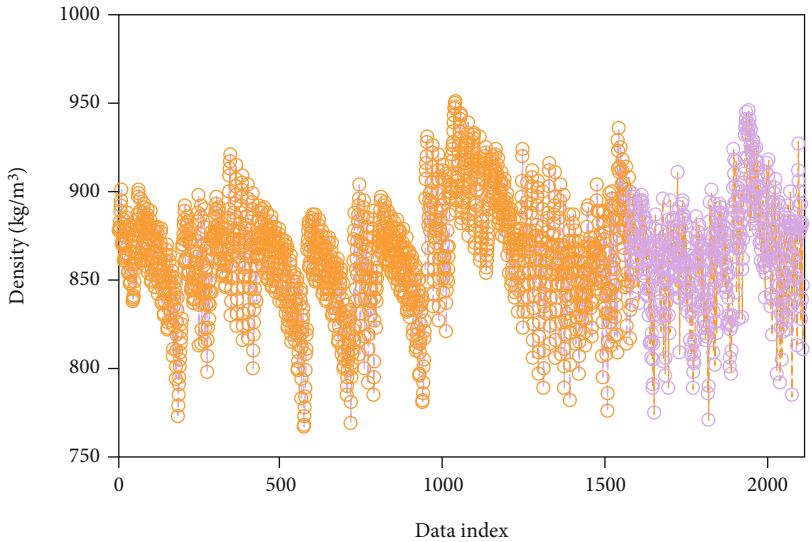
The lack of sufficient accuracy and difficulties of computations cause more attention in the aforementioned literature. On the other hand, much attention has been paid to artificial intelligence methods to a precise solution in order to model different processes [28–33]. One of the things that are crucial for the process design and the same operation is the accuracy of thermophysical properties. The development of a precise and low-cost advancement towards the estimation of biodiesel properties is worthy because of the necessity of these properties most importantly in the clean energy resource topics. For this job, the density of biodiesel fuels has been researched thoroughly. The development of the GPR algorithm model is the main purpose of this research for the estimation of biodiesel density. This algorithm is better for the estimation of individual properties compared to other models because of the independence of this algorithm from the outliers. In the process of development of this model, the reliability and accuracy of the collected dataset are important so for the first time for the identification of suspected data points of biodiesel density, a throughout analysis has been carried out. On the other hand, the effects of input variables on the output have been researched statistically as an important part of this work.

2. Material and Methods

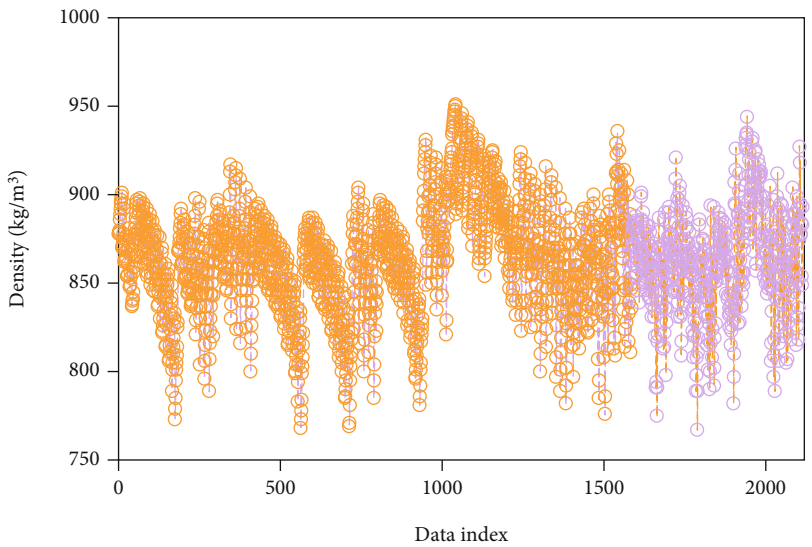
2.1. Gaussian Process Regression. In recent years, the trend of neural networks, which is a branch of artificial intelligence, has made significant progress in solving problems related to the engineering field. The main disadvantage of this method is overfitting, which of course can be improved by adjusting the weight [34, 35].

Of course, setting these parameters is also complex and difficult, and to solve this problem, a conventional mathematical method called Bayesian network is used. It should be noted that this method is probabilistic and uses Bayesian interference to calculate the probability [36]. This network plots each variable graphically, and these variables are connected by an arc, and each variable shows its knowledge content as a distribution of probabilities. It should be noted that the potential specificity of BNs is of great importance for assessing uncertainty. High distribution leads to more uncertainty. One of the factors influencing the increase of complex prior distribution on functions in the Bayesian method to neural networks is the prior overweight [37].

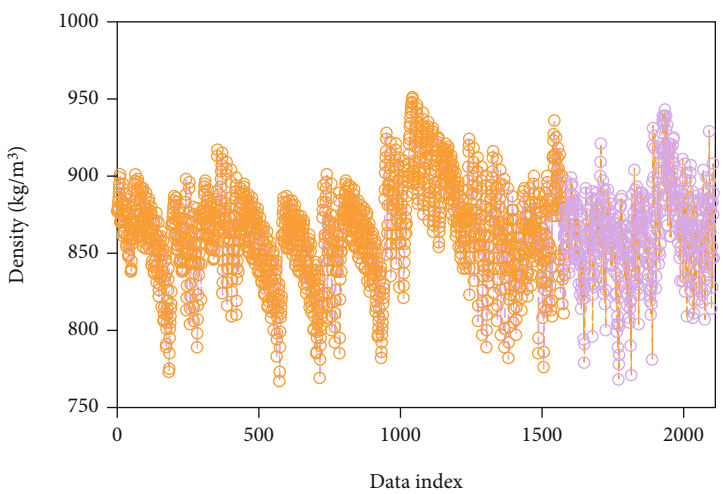
Gaussian process regression is a developed method for the abovementioned problem. It should be noted that this method is nonparametric. The advantages of the GPR



(a)

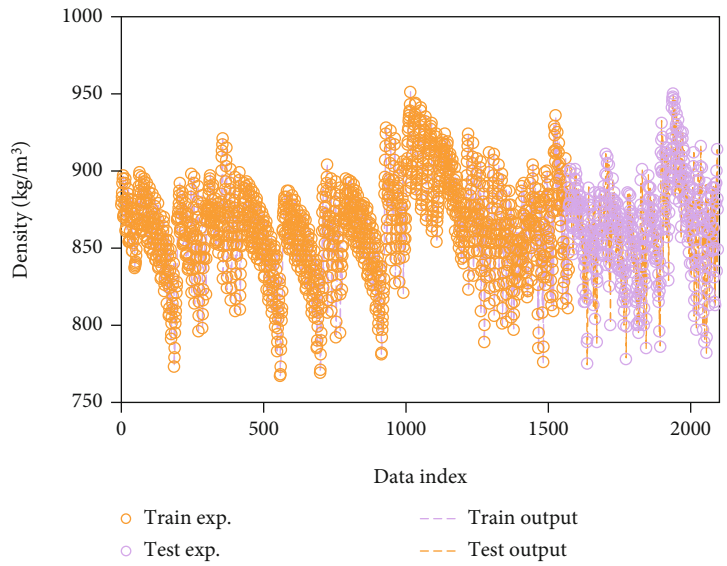


(b)



(c)

Figure 1: Continued.



(d)

Figure 1: The performance of the values predicted against their corresponding actual values using different kernel functions including (a) exponential, (b) Matern, (c) squared exponential, and (d) rational quadratic.

algorithm include measuring uncertainty for predictions and working well on small data. It is worth noting that the GPR method has important advantages over Bayesian, including simplicity, nonlinearity, easy generalization, and has several dimensions. The model parameters are determined using the sample training information in the GPR method.

The GP model is obtained by linking previous knowledge to the current process of modeling and integrating real laboratory data.

One of the salient and important differences between the old methods of machine learning and GPR is not finding the best approximation with experimental points and is a complete BN core. GPR works periodically by obtaining posterior distributions on models. In the following, we will explain how to create GP regression [38].

Randomly selected points $T = \{x_{L,i}, y_{L,i}\}$ and $L = \{x_{T,i}, y_{T,i}\}, i = 1, 2, 3, \cdots, n$, which are test and learning data from a particular distribution, are assumed as follows [39]:

$$T = \{x_{L,i}, y_{L,i}\}, L = \{x_{T,i}, y_{T,i}\}, i = 1, 2, \cdots, n. \quad (1)$$

As an important point, we remind that the model parameters are adjusted based on the learning data [40].

As the input and goal data, respectively, x and y have been assumed that noise has affected them.

The general formula of the GPR described as follows [41]:

$$y_{L,i} = f(x_{L,i}) + e_{L,i}, n = 1, 2, 3, \cdots, n. \quad (2)$$

In the abovementioned equation, X_L and Y_L denote independent and objective variables of the training data, and subsequently, the $\varepsilon \tilde{} N(0, \sigma^2_{\text{noise}} l_n)$ denotes for the observation noise with the independent Gaussian distribution that I_{noise} symbolizes and σ^2_{noise} the variance of noise and unit array [42].

Note: GP assumes the output $f(x)$ is random. So the following equation is obtained:

$$y_{T,i} = f(x_{T,i}) + \varepsilon_{T,i}, i = 1, 2, 3, \cdots, n. \quad (3)$$

In the above equation, y_T and x_T represent the goals and independent variables and the $f(x)$ represents a Gaussian process with covariance function $k(x, x')$ and mean function $m(x)$.

$$f(x_{L,i}) \sim \text{GP}\left(m(x), k\left(x.x'\right)\right). \quad (4)$$

But in practice, the exact determination of $m(x)$ can be complicated, so the value of $m(x)$ is taken to be zero to make the calculations easier, so we have the following [43]:

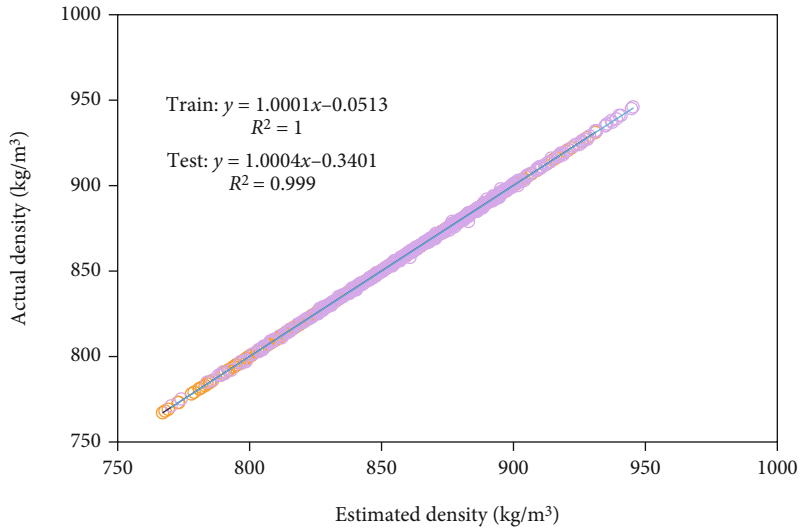
$$f(x_{L,i}) \sim \text{GP}\left(0, k\left(x.x'\right)\right). \quad (5)$$

From Eqs. (2) and (5), we can conclude the following equation:

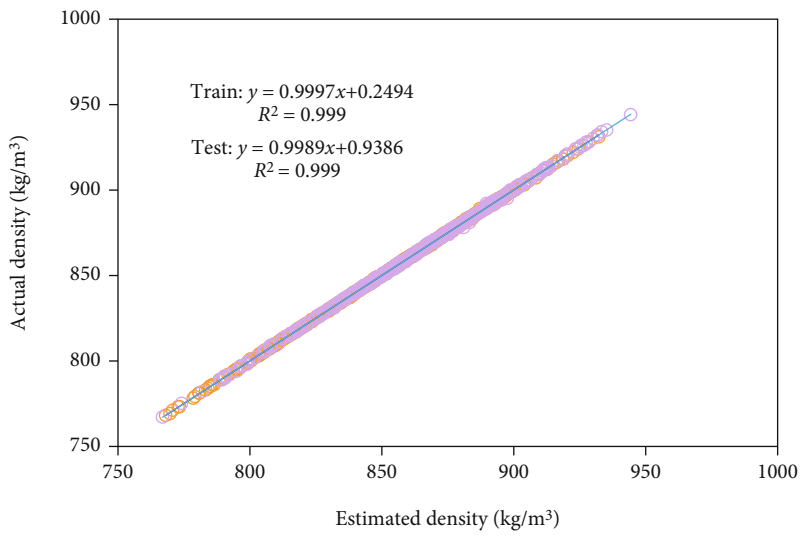
$$y \sim N\left(0, k\left(x.x'\right) + \sigma^2_{\text{noise}} I_n\right). \quad (6)$$

A better representation of the variables mentioned in the text above can be provided as follows [44]:

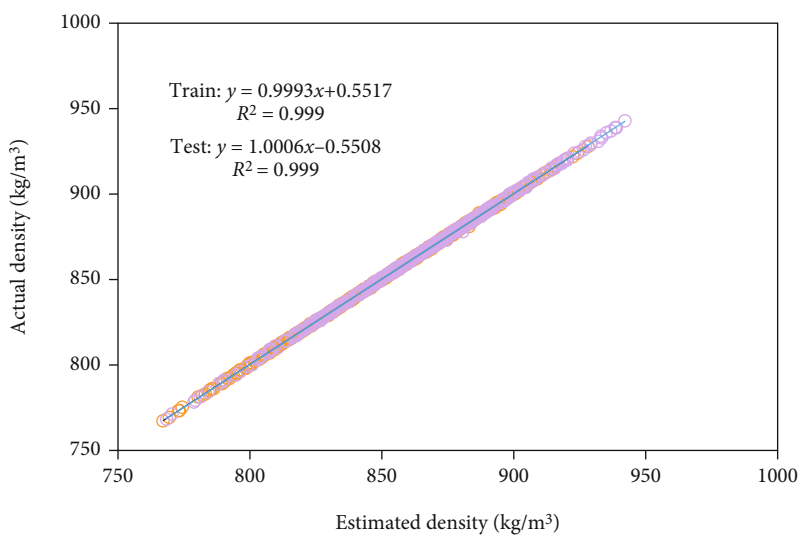
$$\begin{bmatrix} \overrightarrow{f_L} \\ \overrightarrow{f_T} \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x_L.x_L) & k(x_L.x_T) \\ k(x_T.x_L) & k(x_T.x_T) \end{bmatrix}\right), \quad (7)$$



(a)

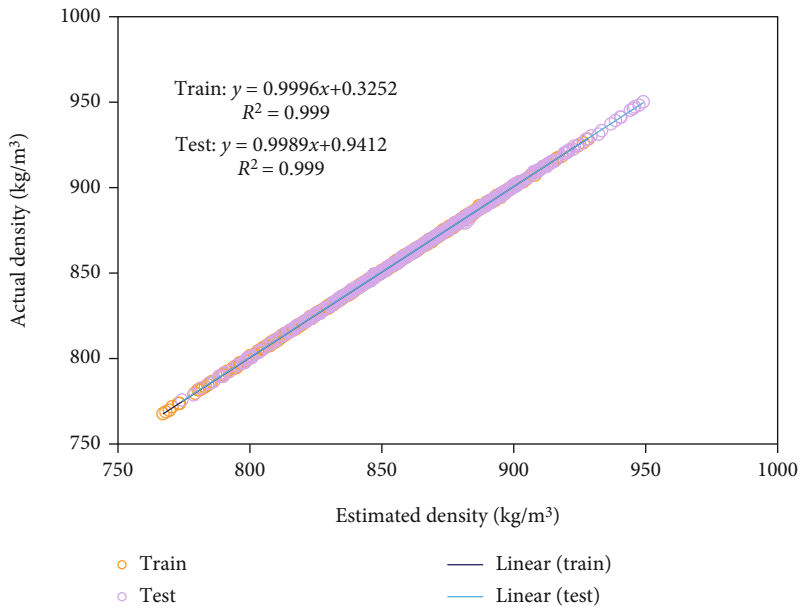


(b)



(c)

Figure 2: Continued.



(d)

Figure 2: Cross plots of predicted output values for different kernel functions including (a) exponential, (b) Matern, (c) squared exponential, and (d) rational quadratic.

$$\begin{bmatrix} \overrightarrow{\varepsilon_L} \\ \overrightarrow{\varepsilon_T} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma^2_{\text{noise}} I_n & 0 \\ 0 & \sigma^2_{\text{noise}} I_n \end{bmatrix}\right). \tag{8}$$

The following Gaussian function is obtained by adding the sum of Eqs. (7) and (8):

$$\begin{bmatrix} \overrightarrow{y_L} \\ \overrightarrow{y_T} \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x_L.x_L) + \sigma^2_{\text{noise}} I_n & k(x_L.x_T) \\ k(x_T.x_L) & k(x_T.x_T) + \sigma^2_{\text{noise}} I_n \end{bmatrix}\right). \tag{9}$$

Therefore, the previous distribution of Y_T is obtained from Gaussian conditions as below:

$$\left(\overrightarrow{y_T} \mid \overrightarrow{y_L}\right) \sim N\left(\mu_T, \Sigma_T + \sigma^2_{\text{noise}} I_n\right). \tag{10}$$

And values Σ_T and μ_T are assumed as follows:

$$\begin{aligned} \Sigma_T = k(x_T.x_T) &= k(x_T.x_T) + \sigma^2_{\text{noise}} I_n \\ &- k(x_T.x_L).\left(k\left(x_L.x_L + \sigma^2_{\text{noise}} I_n\right)^{-1} k(x_L.x_T),\right. \end{aligned}$$

$$\mu_T = m\left(\overrightarrow{y_T}\right) = k(x_T.x_L).\left(k\left(x_L.x_L + \sigma^2_{\text{noise}} I_n\right)^{-1} \overrightarrow{y_T}.\right. \tag{11}$$

In GPR modeling, the following theoretical concept is obtained by predicting the output of experimental data through independent variables and training data. From the above equations, it can be concluded that the covariance and the mean function of both together with the Gaussian distribution represent a GP.

To better predict the goals of the developed GPR model, the selection of the core function in the training phase is of great importance. Therefore, in this research, to find the best kernel function, we use four different and conventional kernel functions, of course, with changes.

These functions are described as follows [45]:

(i) Rational quadratic covariance function

$$k_{\text{RQ}}\left(x.x'\right) = \sigma^2\left(1 + \frac{x - x'2}{2al^2}\right)^{-a}. \tag{12}$$

In rational quadratic covariance function equation $\sigma^2, 1$, $a > 0$ represents the variance, length, and weight scale changes.

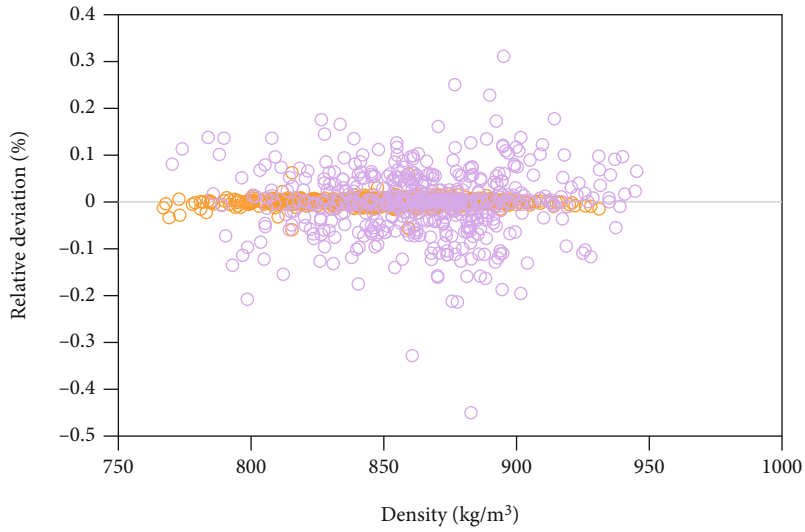
(ii) Squared exponential covariance function

$$k_{\text{SE}}\left(x.x'\right) = \sigma^2\left(-\frac{x - x'2}{l^2}\right). \tag{13}$$

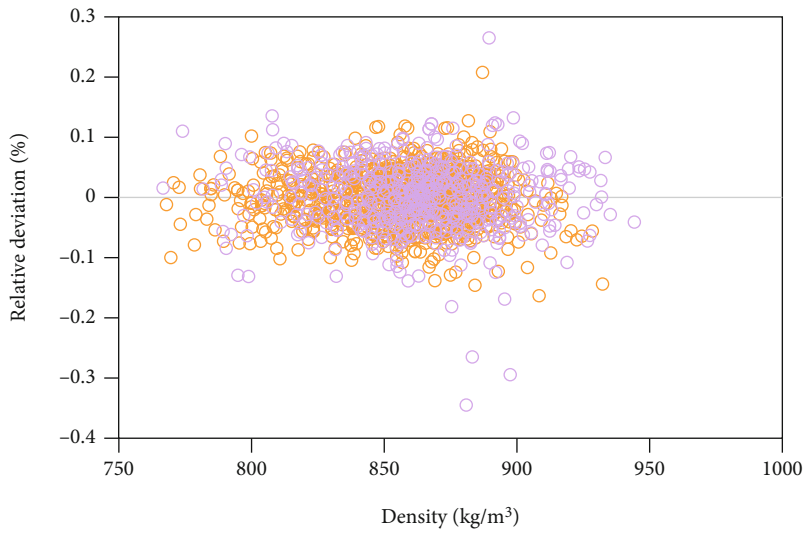
(iii) Exponential covariance function

$$k_E\left(x.x'\right) = \sigma^2 \exp\left(-\frac{x - x'}{l}\right). \tag{14}$$

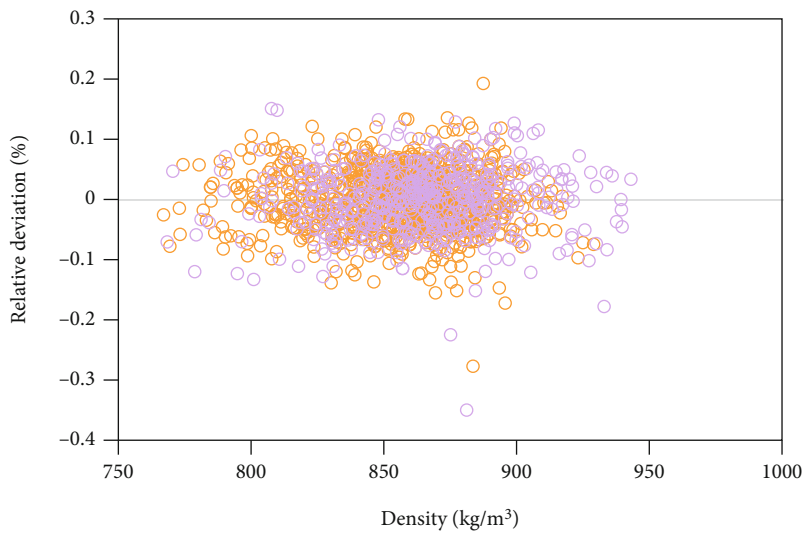
(iv) Matern covariance function



(a)

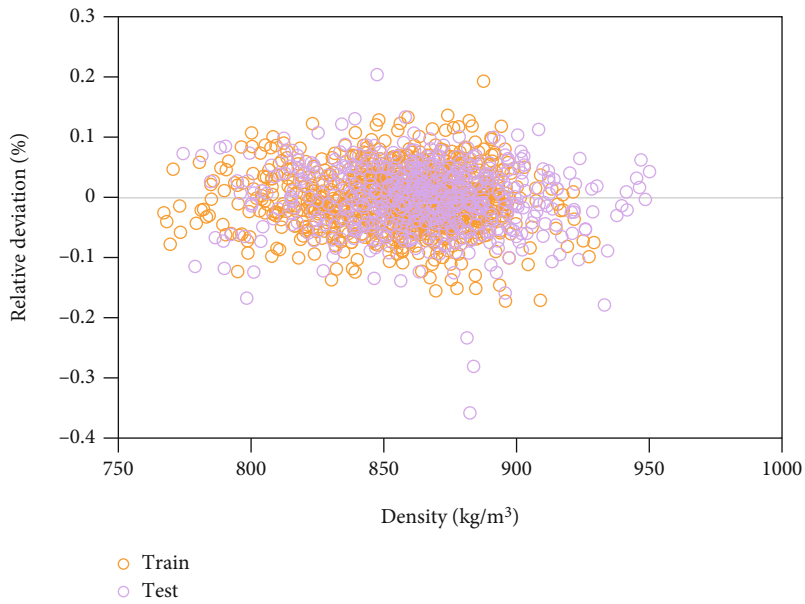


(b)



(c)

Figure 3: Continued.



(d)

Figure 3: The relative deviation between the real density and GPR outputs using different kernel functions including (a) exponential, (b) Matern, (c) squared exponential, and (d) rational quadratic.

$$k_M\left(x.x'\right) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{x-x'}{l}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{x-x'}{l}\right). \tag{15}$$

The symbol represents the gamma function and x, y are positive parameters, and the adjusted K_ν is the Bessel function [46].

In the Matern equation, exponential covariance and quadratic are important functions. The Matern function is an exponential when the value of $V = 0.5$ and also when the exponential is a square when V is inclined to infinity. Since the Matern equation has a greater degree of freedom, it also performs better than the other two [47].

Since the GPR method is nonparametric, the learning stage tries a lot to modify the parameters of the above equations.

2.2. Data Collection. The dataset containing 2117 real density points has been collected from different resources. References to this data have been reported elsewhere [48]. These points are in the pressure range of 0.1-129.78 MPa, melting point of 238.15-304.15 K, molecular weight of 186.291-310.514 g/mol, and temperature range of 278.36-413.15 K. The density values are different between 769.4 and 951.3 kg/m^3 in terms of these conditions.

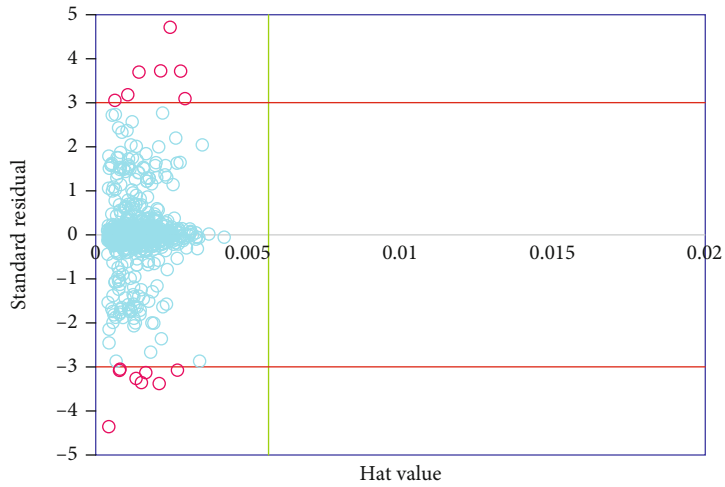
3. Results and Discussion

As mentioned before, here, four various GPR algorithms including kernel functions in Matern, rational quadratic, exponential, and square exponential forms are used for estimating the density of biodiesels. To evaluate the precision of these algorithms, statistical analysis of parameters is determined as follows:

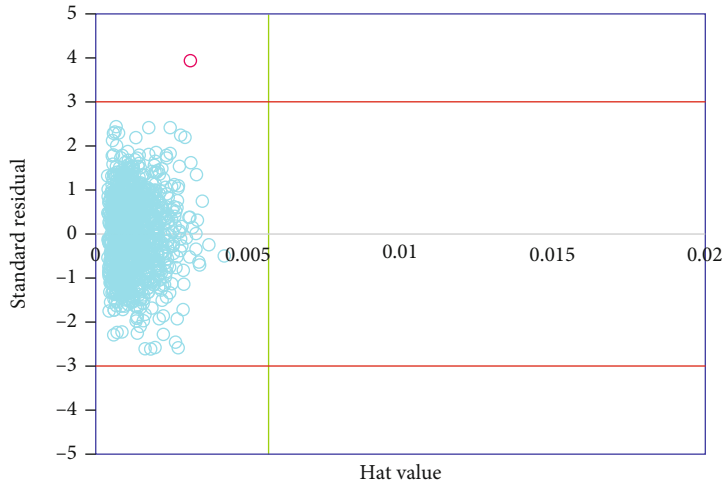
$$\begin{aligned}
\mathrm{MRE} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\left|y_{\mathrm{exp}.,i} - y_{\mathrm{pred}.,i}\right|}{y_{\mathrm{pred}.,i}},\\
\mathrm{MSE} &= \frac{1}{n}\sum_{i=1}^{n} \left(y_{\mathrm{exp}.i} - y_{\mathrm{pred}.i}\right)^2,\\
\mathrm{RMSE} &= \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(y_{\mathrm{exp}.i} - y_{\mathrm{pred}.i}\right)^2},\\
\mathrm{STD} &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \left(\frac{y_{\mathrm{exp}.i} - y_{\mathrm{pred}.i}}{y_{\mathrm{exp}.i}}\right)^2},\\
R^2 &= 1 - \frac{\sum_{i=1}^{n}\left(y_{\mathrm{pred}.i} - y_{\mathrm{exp}.i}\right)^2}{\sum_{i=1}^{n}\left(y_{\mathrm{pred}.i} - \overline{y_{\mathrm{exp}.i}}\right)^2}.
\end{aligned} \tag{16}$$

As you see in Table 1, R^2 values of rational quadratic, Matern, exponential, and square exponential forms are equal to 1. According to other statistical parameters, the rational quadratic form depicts a better performance than the other kernel functions. In this kernel function form, RMSE, MSE, MRE, and STD are obtained 0.47, 0.22, 0.04, and 0.30, respectively. These values exhibit the rational quadratic formability in the forecast of density values.

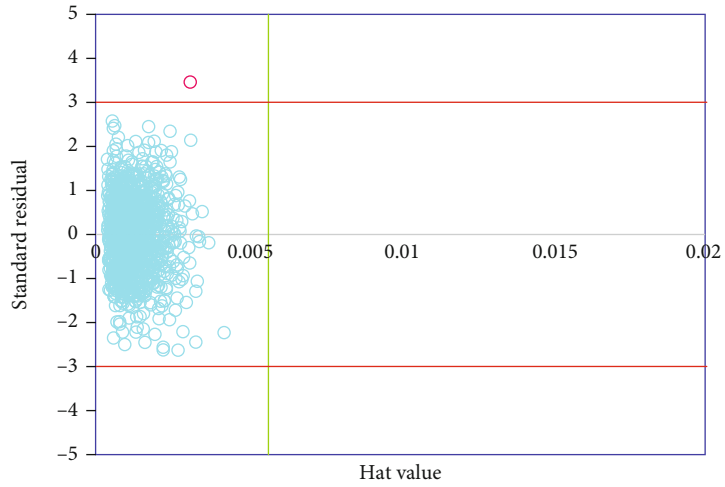
Comparing the results of R^2 from this table with the results published by Abooali et al. who used SGB and GP models to predict biodiesel density, it was concluded that the models presented by us have a higher ability to



(a)

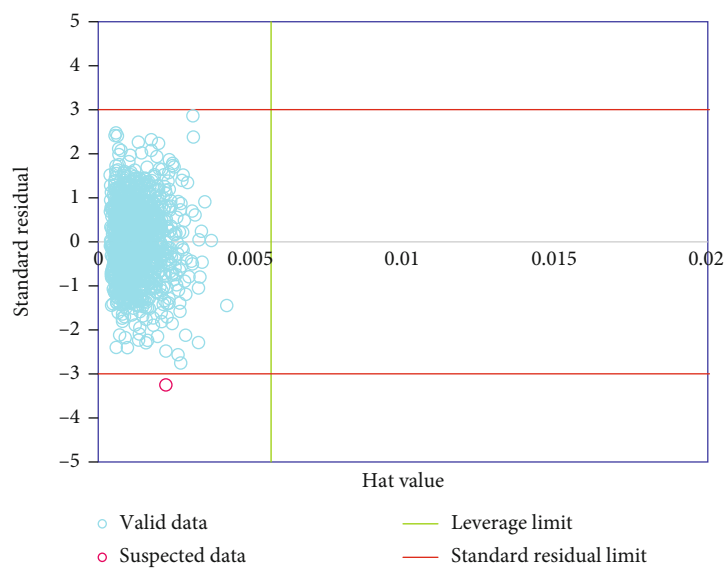


(b)



(c)

Figure 4: Continued.



(d)

Figure 4: Determination of suspected data point for the various kernel functions including (a) exponential, (b) Matern, (c) squared exponential, and (d) rational quadratic.

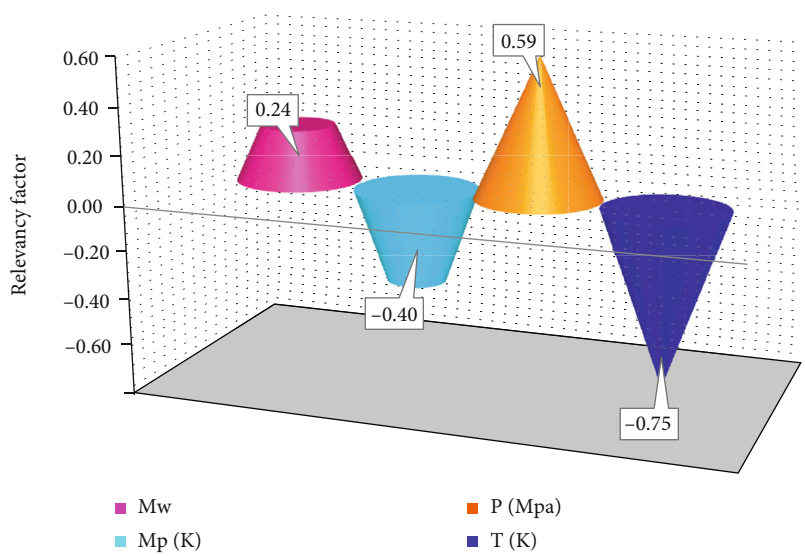


Figure 5: Sensitivity analysis on various input parameters.

predict the target values because the R^2 values for these two models were obtained 0.99988 and 0.99635, respectively [48].

To make a better decision about all these models, Figure 1 shows the experimental and estimated density values, simultaneously. In this figure, there is a good agreement between the real density values and GPR outputs.

Also, the cross/regression plot of predicted and real density has been shown in Figure 2.

To express the quality of GPR outputs, we refer to the density data located on bisector lines in the analysis. Moreover, Figure 3 shows the relative deviation between the real density and GPR outputs.

The accuracy of these density point data affects the validity of models. In this examination, too many data points have been used. It is important to know that these data may have errors due to measurements done in laboratories. So, these types of data are separated from the other data points. In this regard, some strongly developed strategies are required to remove these data and enhance the model accuracy. Here, the separation of these suspected data is accomplished by the Leverage method. In this method, after the determination of residual values, a Hat matrix is created as follows [49]:

$$H = A\left(A^T A\right)^{-1} A^T, \tag{17}$$

where A is an $i \times j$ dimensional matrix and i and j are defined for the model parameters and training points, respectively. Then, the critical leverage limit is computed by i and j as follows:

$$H^* = \frac{3(b+1)}{a}. \tag{18}$$

According to William's plot, as shown in Figure 4, we can separate the suspected data obtained as residuals from Hat values. Here, the leverage limits in the green line and two red lines are considered as residuals. So, the data points located outside these lines are considered as suspected data. As you see, for exponential, Matern, squared exponential, and rational quadratic forms, 14, 1, 1, and 1 points are considered as the suspected data, respectively, among 2217 points.

In this study, the suggested GPR algorithms create a relationship between density and inputs. So, sensitivity analysis is utilized to show this relationship affects the output. In this regard, to determine the most efficient variable for density, the relevancy factor, r, in the range of -1 and 1, is used. If the absolute value of r is large, it can affect the density further. The less and more relation to density is shown by negative and positive r values, respectively. The r is calculated as follows [50]:

$$r = \frac{\sum_{i=1}^{n}(x_{K,i} - \bar{x}_k)(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_{K,i} - \bar{x}_k)^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{19}$$

where $X_{k,i}$ and Y_i are inputs and outputs, respectively. The $\overline{X_k}$ and $\bar{Y}$ are the average of inputs and that of outputs, respectively. Figure 5 shows that the higher temperature and melting point lead to less density. Also, the temperature is considered the most efficient parameter for density. In addition, parameters such as pressure and molecular weight have a direct relationship with this target.

4. Conclusion

In this work, four various kernel functions including Matern, rational quadratic, exponential, and square exponential functions have been used for GPR algorithms to compute the density of biodiesels. To prepare and validate these algorithms, a large database containing 2217 actual data is gathered. It is concluded that the proposed models have highly precise to predict real data. The rational quadratic GPR model has shown greater performance compared with other models. In this model, the calculations show that RMSE = 0.47, MSE = 0.22, MRE = 0.04, $R^2 = 1$, and STD is equal to 0.3. Other analyses also confirmed the accuracy of this model, which indicates that this attractive and simple model can be used in biodiesel-related industries.

Data Availability

The data are stated in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper is supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China No. 18KJD110002 and the Scientific Research Foundation of Huaiyin Normal University No. 31WBX00.

References

[1] D. D. Nguyen, R. Daneshfar, A. H. S. Dehaghani, and C. H. Su, "The effect of shear rate on aggregation and breakage of asphaltenes flocs: experimental study and model-based analysis," *Journal of Molecular Liquids*, vol. 325, p. 114861, 2021.

[2] M. Ghadiri, A. Marjani, R. Daneshfar, and S. Shirazian, "Model order reduction of a reservoir simulation by SOD-DEIM," *Journal of Petroleum Science and Engineering*, vol. 200, p. 108137, 2021.

[3] A. H. S. Dehaghani and R. Daneshfar, "How much would silica nanoparticles enhance the performance of low-salinity water flooding?," *Petroleum Science*, vol. 16, no. 3, pp. 591–605, 2019.

[4] N. Nabipour, R. Daneshfar, O. Rezvanjou et al., "Estimating biofuel density via a soft computing approach based on intermolecular interactions," *Renewable Energy*, vol. 152, pp. 1086–1098, 2020.

[5] S. A. Montzka, E. J. Dlugokencky, and J. H. Butler, "Non-CO 2 greenhouse gases and climate change," *Nature*, vol. 476, no. 7358, pp. 43–50, 2011.

[6] L. Cao, G. Bala, K. Caldeira, R. Nemani, and G. Ban-Weiss, "Importance of carbon dioxide physiological forcing to future climate change," *Proceedings of the National Academy of Sciences*, vol. 107, no. 21, pp. 9513–9518, 2010.

[7] T. Sandler, "Environmental cooperation: contrasting international environmental agreements," *Oxford Economic Papers*, vol. 69, no. 2, pp. 345–364, 2017.

[8] S. Jafarmadar and P. Nemati, "Multidimensional modeling of the effect of exhaust gas recirculation on exergy terms in a homogenous charge compression ignition engine fueled by diesel/biodiesel," *Journal of Cleaner Production*, vol. 161, pp. 720–734, 2017.

[9] A. Espinoza, S. Bautista, P. C. Narváez, M. Alfaro, and M. Camargo, "Sustainability assessment to support governmental biodiesel policy in Colombia: a system dynamics model," *Journal of Cleaner Production*, vol. 141, pp. 1145–1163, 2017.

[10] A. Demirbas and S. Karslioglu, "Biodiesel production facilities from vegetable oils and animal fats," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 29, no. 2, pp. 133–141, 2007.

[11] J. J. Torres, N. E. Rodriguez, J. T. Arana, N. A. Ochoa, J. Marchese, and C. Pagliero, "Ultrafiltration polymeric membranes for the purification of biodiesel from ethanol," *Journal of Cleaner Production*, vol. 141, pp. 641–647, 2017.

[12] A. Datta and B. K. Mandal, "Use of Jatropha biodiesel as a future sustainable fuel," *Energy Technology & Policy*, vol. 1, no. 1, pp. 8–14, 2014.

[13] G. Knothe and K. R. Steidley, "Lubricity of components of biodiesel and petrodiesel. The origin of biodiesel lubricity," *Energy & Fuels*, vol. 19, no. 3, pp. 1192–1200, 2005.

[14] L. F. R. Verduzco, "Density and viscosity of biodiesel as a function of temperature: empirical models," *Renewable and Sustainable Energy Reviews*, vol. 19, pp. 652–665, 2013.

[15] E. Sann, M. Anwar, R. Adnan, and M. A. Idris, "Biodiesel for gas turbine application-an atomization characteristics study," *Advances in Internal Combustion Engines and Fuel Technologies*, pp. 213–242, IntechOpen, 2013.

[16] Z. J. West, T. Yamada, C. R. Bruening et al., "Investigation of water interactions with petroleum-derived and synthetic aviation turbine fuels," *Energy & Fuels*, vol. 32, no. 2, pp. 1166–1178, 2018.

[17] M. J. Pratas, S. V. D. Freitas, M. B. Oliveira, S:. C. Monteiro, A:. S. Lima, and J:. A. P. Coutinho, "Biodiesel density: experimental measurements and prediction models," *Energy & Fuels*, vol. 25, no. 5, pp. 2333–2340, 2011.

[18] S. Phankosol, K. Sudaprasert, S. Lilitchan, K. Aryusuk, and K. Krisnangkura, "Estimation of density of biodiesel," *Energy & Fuels*, vol. 28, no. 7, pp. 4633–4641, 2014.

[19] A. T. Hoang, "Prediction of the density and viscosity of biodiesel and the influence of biodiesel properties on a diesel engine fuel supply system," *Journal of Marine Engineering & Technology*, vol. 17, pp. 1–13, 2018.

[20] A. Queimada, L. Rolo, A. Caco, I. Marrucho, E. Stenby, and J. Coutinho, "Prediction of viscosities and surface tensions of fuels using a new corresponding states model," *Fuel*, vol. 85, no. 5-6, pp. 874–877, 2006.

[21] A. Barati-Harooni, A. Soleymanzadeh, A. Tatar et al., "Experimental and modeling studies on the effects of temperature, pressure and brine salinity on interfacial tension in live oil-brine systems," *Journal of Molecular Liquids*, vol. 219, pp. 985–993, 2016.

[22] A. Rostami, H. Ebadi, M. Arabloo, M. K. Meybodi, and A. Bahadori, "Toward genetic programming (GP) approach for estimation of hydrocarbon/water interfacial tension," *Journal of Molecular Liquids*, vol. 230, pp. 175–189, 2017.

[23] A. A. Gakh, E. G. Gakh, B. G. Sumpter, and D. W. Noid, "Neural network-graph theory approach to the prediction of the physical properties of organic compounds," *Journal of Chemical Information and Computer Sciences*, vol. 34, no. 4, pp. 832–839, 1994.

[24] S. M. Miraboutalebi, P. Kazemi, and P. Bahrami, "Fatty acid methyl ester (FAME) composition used for estimation of biodiesel cetane number employing random forest and artificial neural networks: a new approach," *Fuel*, vol. 166, pp. 143–151, 2016.

[25] M. Mostafaei, "Prediction of biodiesel fuel properties from its fatty acids composition using ANFIS approach," *Fuel*, vol. 229, pp. 227–234, 2018.

[26] D. Li, M. Guo, X. Wang, S. Lin, W. Jia, and G. Wang, "Measurement and correlation of density and viscosity of binary mixtures of fatty acid (methyl esters+ methylcyclohexane)," *The Journal of Chemical Thermodynamics*, vol. 137, pp. 86–93, 2019.

[27] R. Aitbelale, Y. Chhiti, F. E. M.'. Alaoui, A. Sahib Eddine, N. Muñoz Rujas, and F. Aguilar, "High-pressure soybean oil biodiesel density: experimental measurements, correlation by Tait equation, and perturbed chain SAFT (PC-SAFT) modeling," *Journal of Chemical & Engineering Data*, vol. 64, no. 9, pp. 3994–4004, 2019.

[28] R. Daneshfar, A. Bemani, M. Hadipoor et al., "Estimating the heat capacity of non-Newtonian ionanofluid systems using ANN, ANFIS, and SGB tree algorithms," *Applied Sciences*, vol. 10, no. 18, p. 6432, 2020.

[29] M. B. Vanani, R. Daneshfar, and E. Khodapanah, "A novel MLP approach for estimating asphaltene content of crude oil," *Petroleum Science and Technology*, vol. 37, no. 22, pp. 2238–2245, 2019.

[30] R. Daneshfar, F. Keivanimehr, M. Mohammadi-Khanaposhtani, and A. Baghban, "A neural computing strategy to estimate dew-point pressure of gas condensate reservoirs," *Petroleum Science and Technology*, vol. 38, no. 10, pp. 706–712, 2020.

[31] R. Setiawan, R. Daneshfar, O. Rezvanjou, S. Ashoori, and M. Naseri, "Surface tension of binary mixtures containing environmentally friendly ionic liquids: insights from artificial intelligence," *Environment, Development and Sustainability*, pp. 1–22, 2021.

[32] F. Mousazadeh, M. H. T. Naeem, R. Daneshfar, B. S. Soulgani, and M. Naseri, "Predicting the condensate viscosity near the wellbore by ELM and ANFIS-PSO strategies," *Journal of Petroleum Science and Engineering*, vol. 204, article 108708, 2021.

[33] S. M. Alizadeh, I. Alruyemi, R. Daneshfar, M. Mohammadi-Khanaposhtani, and M. Naseri, "An insight into the estimation of drilling fluid density at HPHT condition using PSO-, ICA-, and GA-LSSVM strategies," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[34] C. K. Williams, *Prediction with Gaussian processes: from linear regression to linear prediction and beyond, in Learning in graphical models*, Springer, 1998.

[35] L. Uusitalo, "Advantages and challenges of Bayesian networks in environmental modelling," *Ecological Modelling*, vol. 203, no. 3-4, pp. 312–318, 2007.

[36] K. H. Reckhow, "Water quality prediction and probability network models," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 56, no. 7, pp. 1150–1158, 1999.

[37] S. Asante-Okyere, C. Shen, Y. Yevenyo Ziggah, M. Moses Rulegeya, and X. Zhu, "Investigating the predictive performance of Gaussian process regression in evaluating reservoir porosity and permeability," *Energies*, vol. 11, no. 12, p. 3261, 2018.

[38] S. Sniekers and A. van der Vaart, "Adaptive Bayesian credible sets in regression with a Gaussian process prior," *Electronic Journal of Statistics*, vol. 9, no. 2, pp. 2475–2527, 2015.

[39] J. Quinonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[40] Q. Fu, W. Shen, X. Wei, P. Zheng, H. Xin, and C. Zhao, "Prediction of the diet nutrients digestibility of dairy cows using Gaussian process regression," *Information Processing in Agriculture*, vol. 6, no. 3, pp. 396–406, 2019.

[41] M. Zhu, S. Liu, and S. Gu, "Short-term tide level forecasting based on Gaussian process regression," in *The Eleventh ISOPE Pacific/Asia Offshore Mechanics Symposium*, Shanghai, China, 2014.

[42] Y. Zhang, G. Su, and L. Yan, "Gaussian process machine learning model for forecasting of karstic collapse," in *Communications in Computer and Information Science*, Springer, 2011.

[43] Y. Zhang and X. Xu, "Predicting doped MgB_2 superconductor critical temperature from lattice parameters using Gaussian process regression," *Physica C: Superconductivity and Its Applications*, vol. 573, p. 1353633, 2020.

[44] A. S. Alghamdi, K. Polat, A. Alghoson, A. A. Alshdadi, and A. A. Abd el-Latif, "Gaussian process regression (GPR) based non-invasive continuous blood pressure prediction method from cuff oscillometric signals," *Applied Acoustics*, vol. 164, p. 107256, 2020.

[45] S. A. Aye and P. Heyns, "An integrated Gaussian process regression for prediction of remaining useful life of slow speed bearings based on acoustic emission," *Mechanical Systems and Signal Processing*, vol. 84, pp. 485–498, 2017.

[46] W. Gao, M. Karbasi, M. Hasanipanah, X. Zhang, and J. Guo, "Developing GPR model for forecasting the rock fragmentation in surface mines," *Engineering with Computers*, vol. 34, no. 2, pp. 339–345, 2018.

[47] S. Hong, Z. Zhou, C. Lu, B. Wang, and T. Zhao, "Bearing remaining life prediction using Gaussian process regression with composite kernel functions," *Journal of Vibroengineering*, vol. 17, no. 2, pp. 695–704, 2015.

[48] D. Abooali, R. Soleimani, and S. Gholamreza-Ravi, "Characterization of physico-chemical properties of biodiesel components using smart data mining approaches," *Fuel*, vol. 266, p. 117075, 2020.

[49] A. Lekomtsev, A. Keykhosravi, M. B. Moghaddam, R. Daneshfar, and O. Rezvanjou, "On the prediction of filtration volume of drilling fluids containing different types of nanoparticles by ELM and PSO-LSSVM based models," *Petroleum*, 2021.

[50] D. Ahangari, R. Daneshfar, M. Zakeri, S. Ashoori, and B. S. Soulgani, "On the prediction of geochemical parameters (TOC, S1 and S2) by considering well log parameters using ANFIS and LSSVM strategies," *Petroleum*, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 5530093, 10 pages
https://doi.org/10.1155/2021/5530093

Research Article

On the Evaluation of Rhamnolipid Biosurfactant Adsorption Performance on Amberlite XAD-2 Using Machine Learning Techniques

Fengqin Chen,[1] Jinbo Huang,[2] Xianjun Wu,[3] Xiaoli Wu,[4] and Arash Arabmarkadeh[5,6]

[1]*Inspection Department, Maoming People's Hospital, Maoming Guangdong 525000, China*
[2]*Logistics Department, Maoming People's Hospital, Maoming Guangdong 525000, China*
[3]*School of Computer, Guangdong University of Petrochemical Technology, Maoming Guangdong 525000, China*
[4]*Burn Department of Maoming People's Hospital, Maoming Guangdong 525000, China*
[5]*Biotechnology Group, Faculty of Chemical Engineering, Tarbiat Modares University, P.O. Box 14115-143, Tehran, Iran*
[6]*Microbial Biotechnology Department, Agricultural Biotechnology Research Institute of Iran (ABRII), Agricultural Research Education and Extension Organization (AREEO), Karaj, Iran*

Correspondence should be addressed to Xiaoli Wu; xiaoliwu_6859@21cn.com
and Arash Arabmarkadeh; arash.arabmarkade@modares.ac.ir

Received 25 January 2021; Revised 6 March 2021; Accepted 31 March 2021; Published 27 April 2021

Academic Editor: Alireza Baghban

Biosurfactants are a series of organic compounds that are composed of two parts, hydrophobic and hydrophilic, and since they have properties such as less toxicity and biodegradation, they are widely used in the food industry. Important applications include healthy products, oil recycling, and biological refining. In this research, to calculate the curves of rhamnolipid adsorption compared to Amberlite XAD-2, the least-squares vector machine algorithm has been used. Then, the obtained model is formed by 204 adsorption data points. Various graphical and statistical approaches are applied to ensure the correctness of the model output. The findings of this study are compared with studies that have used artificial neural network (ANN) and data group management method (GMDH) models. The model used in this study has a lower percentage of absolute mean deviation than ANN and GMDH models, which is estimated to be 1.71%.The least-squares support vector machine (LSSVM) is very valuable for investigating the breakthrough curve of rhamnolipid, and it can also be used to help chemists working on biosurfactants. Moreover, our graphical interface program can assist everyone to determine easily the curves of rhamnolipid adsorption on Amberlite XAD-2.

1. Introduction

As mentioned above, biosurfactants are organic compounds that are produced by microorganisms and consist of two parts: hydrophilic and hydrophobic. They are often produced by bacteria on living surfaces. One of the reasons for attracting many industrial applications to biosurfactants is due to their amphiphilic properties. Among the usable and outstanding capabilities of biosurfactants used in various industries such as mines, fertilizers, petrochemicals, and petroleum, we can mention the environmental degradability and reduction of surface tension between interstitial and low toxicity. The reduction of the interfacial tension is due to the increase in the solubility of hydrophilic molecules when using biosurfactants. Capabilities such as surface modification and interfacial tension have made surfactants attractive to the industry. Rhamnolipids (RLs) are the most studied type of biosurfactants. According to the literature, rhamnolipids can reduce water surface tension by about 60% [1–3] for different concentrations of RL 50-65 mg/L. The production of RLs usually involves a final product from a dilute solution contaminated with undesirable impurities. There are several ways to increase the concentration and eliminate contaminants in which the adsorption process is widely studied.

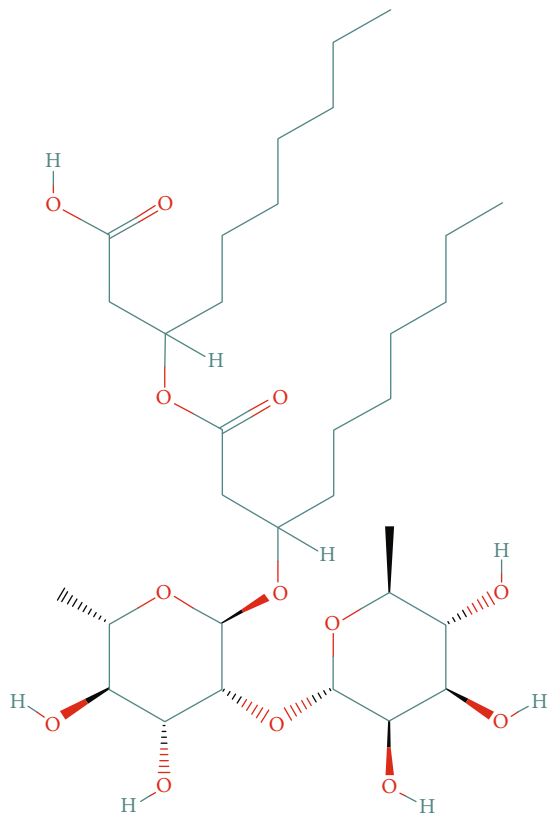


Figure 1: Chemical structure of the first identified rhamnolipid, symbolized as Rha-Rha-C_{10}-C_{10}.

In this research, activated carbon is used for adsorbent in the process. The breakthrough curve of a packed column is a very significant attribute of this system. As a result, determining such curves will be useful for optimizing and understanding the performance of the column. To model the adsorption phenomena, the mass balance in liquid and solid phases is evaluated. It may also include modeling the porous and liquid film resistance and also axial dispersion. Finally, with a suitable software package, a set of differential equations can be solved.

Ill conditions and uncertainty in differential equations make using conventional mathematical models not suitable. Intelligent models have to be a powerful tool in solving process modeling problems. To predict the optimized targets, in various fields such as petroleum and gas fields, methods such as SVM, ANN, group method data manipulation (GMDH), fuzzy logic system, and adaptive fuzzy neural inference system can be used. Interactions between AI neurons are achieved by connecting different units. Artificial neurons' interactions are achieved by connecting different units. Each weighted output is related to the sum of the output from the previous synaptic weight layer, and then it is used as an input for a specific neuron. Backpropagation ANNs are extensively applied, as they have shown to be a capable and powerful tool [4]. The GMDH model is a type of backpropagation ANN that was proposed by Ivakhnenko [5]. Darwin's theory of selection inspired this approach. The prominent feature of this method is the internal process of the elements [6–8]. To process elements in a conventional ANN, log sigmoid, hard limit, linear, and tangent sigmoid transfer functions are considered. On the other hand, the GMDH method constructs simple polynomials, roughly predicting the targeted systems. In the next step, the complexity of the polynomials is further developed so that satisfactory models are achieved. [9, 10].

Due to the importance of predicting a trustworthy estimation of breakthrough curves, this research is aimed at predicting of breakthrough curves utilizing the LSSVM method for rhamnolipid (Figure 1) adsorption over Amberlite XAD-2. Furthermore, results are compared with those of ANN and GMDH models. The investigated model takes into account 204 data points in its network for adsorption over the Amberlite XAD-2. Various graphical and statistical methods are considered to evaluate the accuracy of this strategy.

2. Model Development

In the present research, the LSSVM strategy was applied to calculate the curve to achieve rhamnolipid uptake relative to the Amberlite XAD-2 model resulting in a more simplified way [11, 12]. SVM can be defined as a function as below:

$$f(x) = w^T \varphi(x) + b. \tag{1}$$

The parameters of the above expression are as follows:

w^T denotes the transpose vector corresponding tothe output layer.

b and $\varphi(x)$ represent the bias and the kernel function, respectively.

The input (x) consists of $N \times n$ dimension in which n and N are input parameters and some data points, respectively. The following cost function is optimized to evaluate w^T and b parameters [13]:

$$\text{cost function} = \frac{1}{2} w^T + C \sum_{k=1}^{N} (\xi_k - \xi_k^*), \tag{2}$$

which is constrained by

$$\begin{cases} y_k - w^T \varphi(x_k) - b \le \varepsilon + \xi_k, k = 1, 2, \cdots, N \\ w^T \varphi(x_k) + b - y_k \le \varepsilon + \xi_k^*, k = 1, 2, \cdots, N \\ \xi_k, \xi_k^* \ge 0 \end{cases} \tag{3}$$

y_k is the kth output while x_k is the kth input. ε stands for the fixed precision of the estimation. Also, slack variables (ξ_k, ξ_k^*) are dealing to determine the acceptable error margin. The below lagrangian is applied to minimize the cost function:

$$L(a, a^*) = -\frac{1}{2}\sum_{k,l=1}^{N}(a_k - a_k^*)(a_l - a_l^*)K(x_k, x_l) - \varepsilon\sum_{k=1}^{N}(a_k - a_k^*) + \sum_{k=1}^{N} y_k(a_k - a_k^*),$$
$$\sum_{k=1}^{N}(a_k - a_k^*) = 0,\ a_k, a_k^* \in [0, c],$$
$$K(x_k, x_l) = \varphi(x_k)^T\varphi(x_l),\ k = 1, 2, \cdots, N, \quad (4)$$

where a_k and a_k^* stand for Lagrangian multipliers. In the last step, the SVM is given below:

$$f(x) = \sum_{k,l=1}^{N}(a_k - a_k^*)K(x, x_k) + b. \quad (5)$$

Quadratic programming must be solved to determine the SVM parameters. The LSSVM eliminates deficiencies in the solving process of a quadratic programming problem [11, 12]. LSSVM uses the below equation in the process of model development:

$$\text{cost function} = \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum_{k=1}^{N} e_k^2, \quad (6)$$

where

γ denotes tuning parameter.

e_k is the error variable.

The following constraints are applied to the cost function:

$$y_k = w^T\varphi(x_k) + b + e_k. \quad (7)$$

The Lagrangian of the LSSVM is expressed as

$$l(w, b, e, a) = \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum_{k=1}^{N} e_k^2 - \sum_{k=1}^{N} a_k\left(w^T\varphi(x_k) + b + e_k - y_k\right). \quad (8)$$

In the above phrase, the symbol a_k represents the Lagrangian multipliers. To optimize Eq. (8), its derivatives are set to zero, and as a result, the following equations are achieved:

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \Longrightarrow w = \displaystyle\sum_{k=1}^{N} a_k\varphi(x_k), \\ \dfrac{\partial L}{\partial b} = 0 \Longrightarrow \displaystyle\sum_{k=1}^{N} a_k = 0, \\ \dfrac{\partial L}{\partial e_k} = 0 \Longrightarrow a_k = \gamma e_k,\ k = 1, 2, \cdots, N, \\ \dfrac{\partial L}{\partial a_k} = 0 \Longrightarrow w^T\varphi(x_k) + b + e_k - y_k = 0,\ k = 1, 2, \cdots, N. \end{cases} \quad (9)$$

By solving the aforementioned equations, LSSVM parameters are obtained. LSSVM employs the kernel function in the same way that SVM strategy does. The most common applied kernel function is the radial basis function (RBF) which is given by

$$K(x, x_k) = \exp\left(-\|x_k - x\|^2/\sigma^2\right), \quad (10)$$

where σ^2 stands for the tuning parameter corresponding to the kernel function. As a result, two tuning parameters (σ^2 and γ) are adjustable. The last-mentioned parameters can be determined by minimizing the error between the predicted values and experimental ones through the application of mean square error (MSE):

$$MSE = \frac{1}{N}\sum_{k=1}^{N}\left(y_k^{\text{pred.}} - y_k^{\text{exp.}}\right)^2, \quad (11)$$

where y is the output value, and exp. and pred. subscripts denote experimental and predicted values, respectively. Also, in this paper, we used the particle swarm optimization algorithm for the determination of these tuning parameters. A typical diagram of the proposed LSSVM approach has been shown in Figure 2.

The adjusted parameters are γ and σ^2 in the LSSVM model and based on the identified cost function (Eq. (11)), and these parameters are optimally determined by optimization technique. The values of γ and σ^2 in this study are 984523.52 and 0.246, respectively, through the PSO algorithm with swarm size and iteration of 80 and 1000, respectively.

Different statistical error analyses such as mean absolute error (MAE), coefficient of determination (R^2), and root means square error (RMSE) are implicated to analyze the model's performance.

$$R^2 = 1 - \frac{\sum_{k=1}^{N}\left(y_k^{\text{exp.}} - y_k^{\text{pred.}}\right)^2}{\sum_{k=1}^{N}\left(y_k^{\text{exp.}} - y_{\text{ave.}}\right)^2},$$
$$\text{MAE} = \frac{\sum_{k=1}^{N}\left|y_k^{\text{pred.}} - y_k^{\text{exp.}}\right|}{N}, \quad (12)$$
$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(y_k^{\text{pred.}} - y_k^{\text{exp.}}\right)^2}.$$

3. Identification of Outlying Experimental Data

The outlier is a set of data having a different behavior in comparison with the bulk of data. Finding outliers would improve the accuracy and reliability of a proposed model remarkably. To help to trace outliers, there are two procedures numerical and graphical procedures. One of the most powerful methods is the Leverage method in which the deviation of estimated values from the experimental ones is calculated. It also includes dealing with Hat matrix being made of experimental and predicted data. The equation below is used for calculating the Hat indices [14–16]:

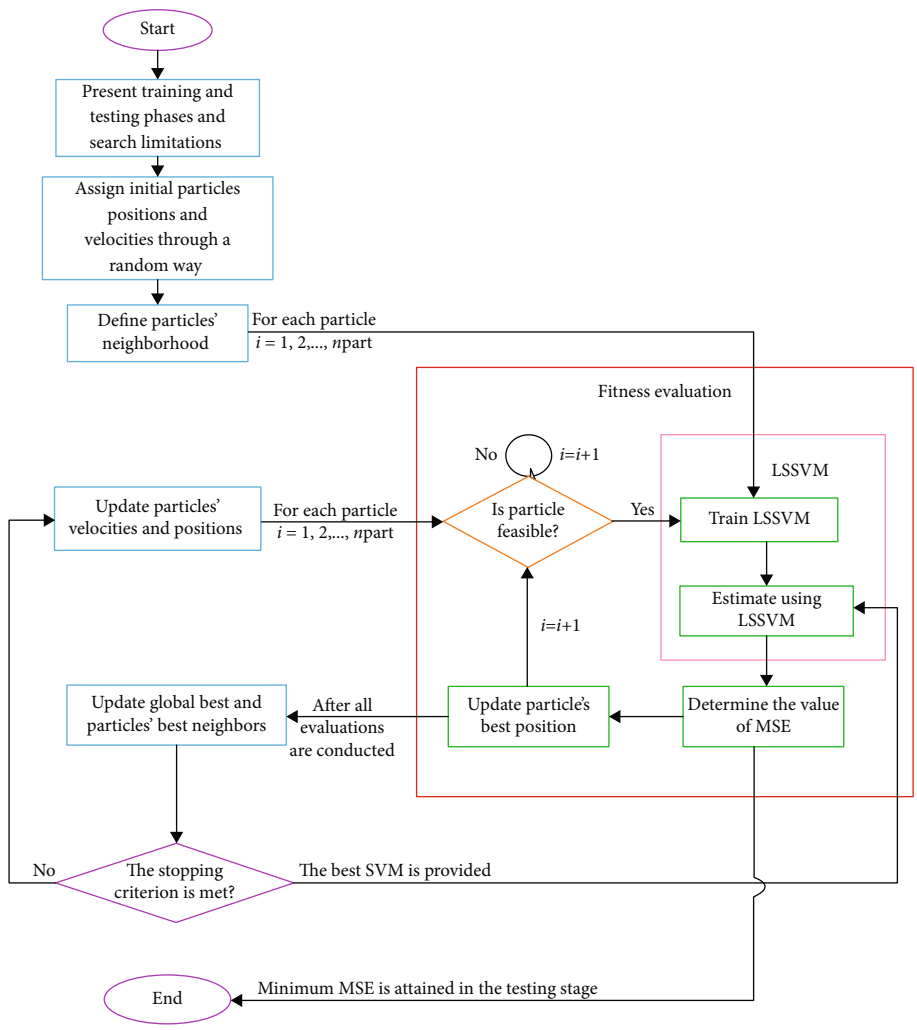


Figure 2: Schematic diagram of PSO-LSSVM strategy.

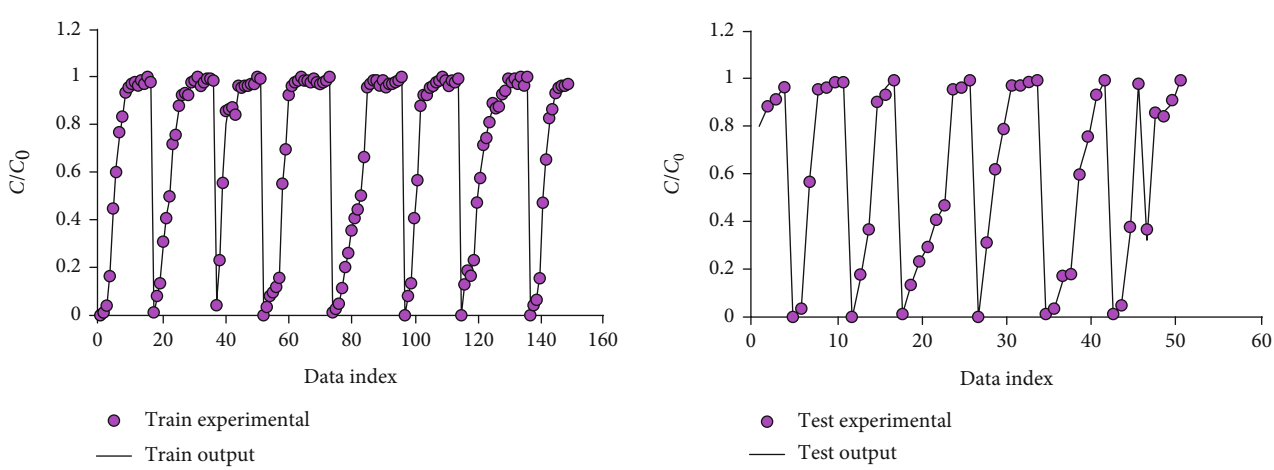


Figure 3: Plot of PSO-LSSVM model's prediction vs. experimental data at training and testing stages.

$$H = X\left(X^t X\right)^{-1} X^t. \quad (13)$$

$X\,(n \times x)$ is a matrix including n data and k parameters of the model, and t denotes the transpose matrix. The diagonal values of the matrix (H) are called H values. H values will aid in the detection process of the possible outliers, utilizing a Williams plot in which the relationship between standardized crossvalidated residuals (R) and Hat indices is shown. The warning leverage is given as follows:

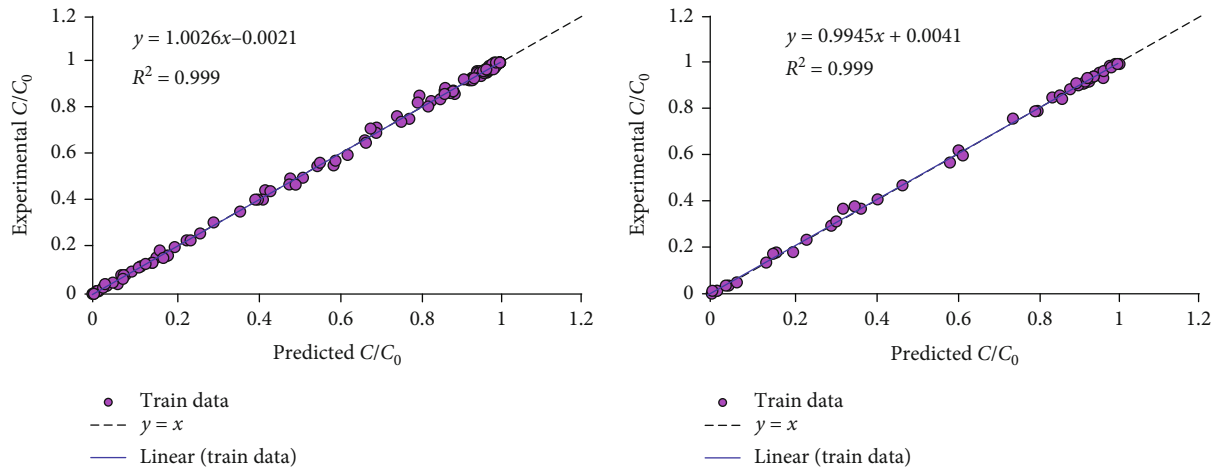


Figure 4: Regression plot of suggested PSO-LSSVM model at training and testing stages.

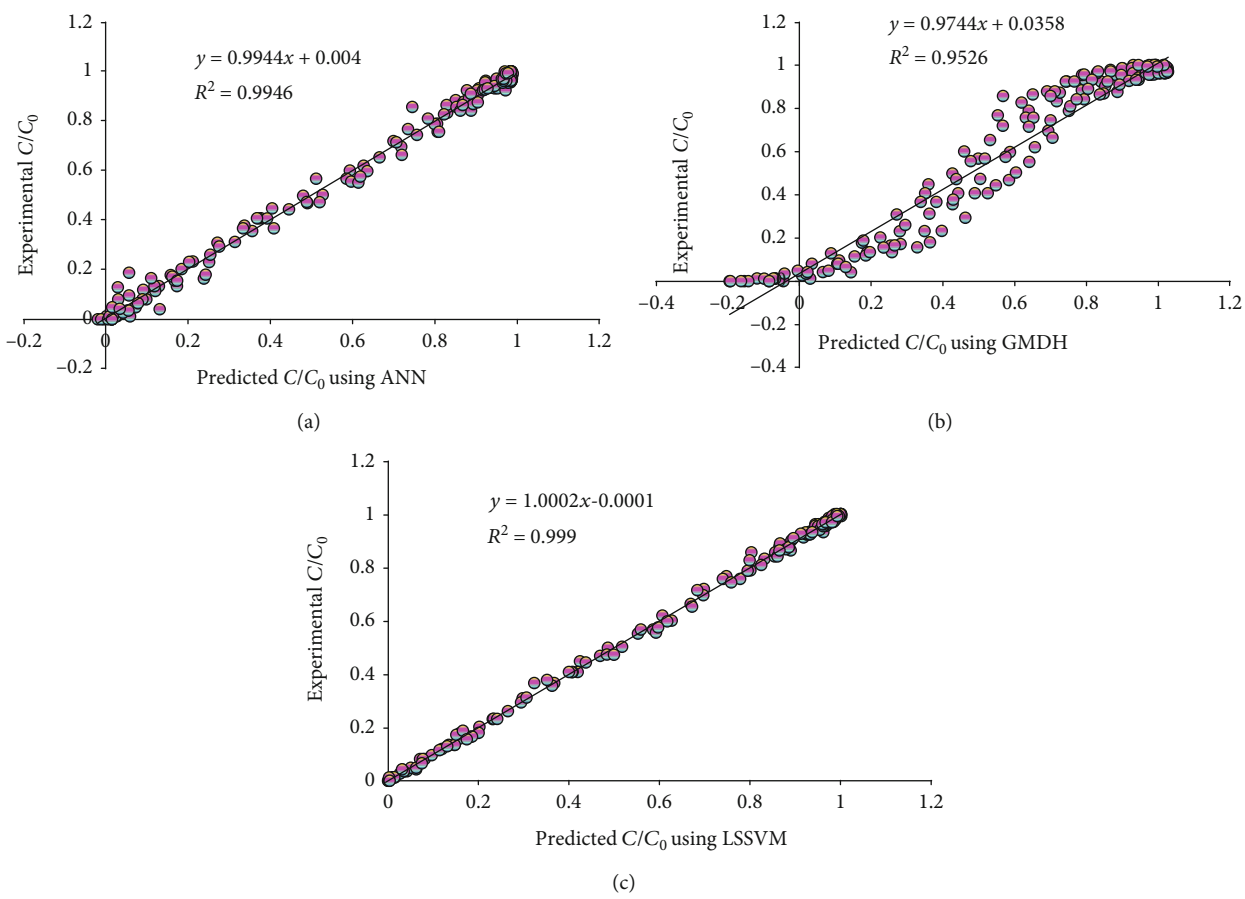


Figure 5: Cross plot of predictions of different models for total data points: (a) ANN, (b) GMDH, and (c) LSSVM.

$$H^* = 3\frac{(f+1)}{p}. \tag{14}$$

p and f stand for numbers of data points and model parameters, respectively.

A reliable model would contain the majority of the predicted values by satisfying the following constraint:

$$R \in [-3, 3], 0 < H < \mathrm{H}^*. \tag{15}$$

Regardless of the value of H, if the value of R for a given

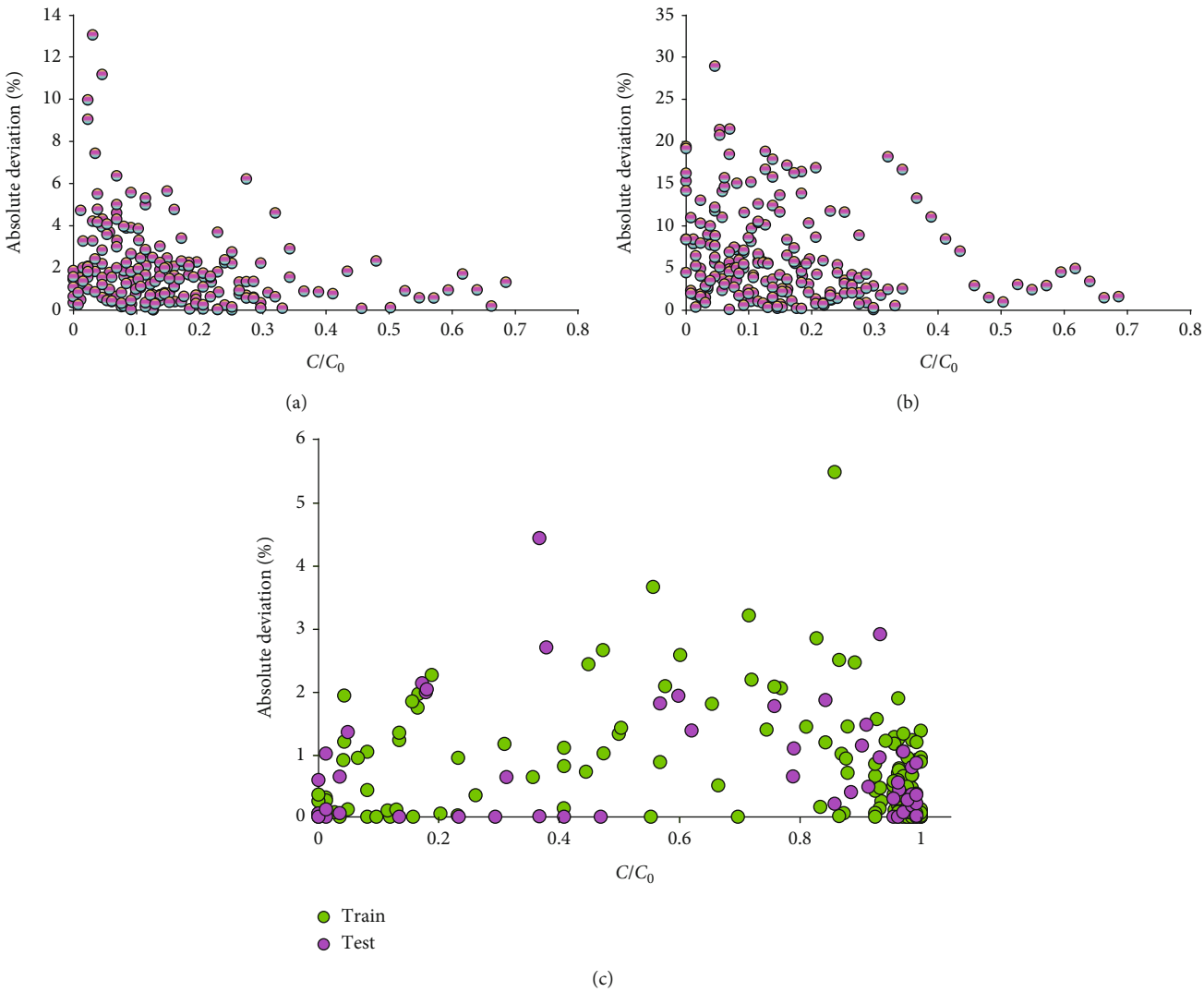


Figure 6: Absolute error for different models' outcomes: (a) ANN, (b) GMDH, and (c) LSSVM.

data is outside the above range, it is considered a possible candidate for being an outlier. The data in this paper are provided in Table S1, and this data set is taken from the previous paper [17]. As discussed, input parameters are initial rhamnolipid concentration, fixed bed height, flow velocity, and run time, while the ratio of final to initial concentration of rhamnolipid (C/C_0) would be the output parameters ones. The 204 data points are divided into two categories: training and testing.

To create the LSSVM model, 75% of data points are considered as learning points, and the rest of them were used to examine the efficiency of the opposed model. Furthermore, data are normalized within the range of [-1,1] applying the equation below:

$$D_N = 2\frac{D - D_{\min}}{D_{\max} - D_{\min}} - 1. \quad (16)$$

Here, D and D_N represent actual and normalized data points, respectively. Also, $D_{\min}$ and $D_{\max}$ stand for minimum and maximum values of data points, respectively.

Table 1: Statistical parameters calculated for three models.

Analysis	LSSVM Train	LSSVM Test	LSSVM Total	ANN Total	GMDH Total
MSE	0.0001	0.0002	0.0001	0.0005	0.0047
AAD	0.7293	0.8043	0.7481	1.9111	6.2395
R^2	0.9990	0.9990	0.9990	0.9946	0.9526
STD	0.0112	0.0122	0.0115	0.0271	0.0808

3.1. Evaluation of the model's Accuracy. The predictive model's accuracy is investigated employing different graphical and statistical methods. Figure 3 represents experimental data points and model estimation by the proposed LSSVM method in the training and testing stages.

Figure 4 shows predicted values against experimental ones. The more it would be close to line $Y = X$, the more appropriate the prediction of the proposed model.

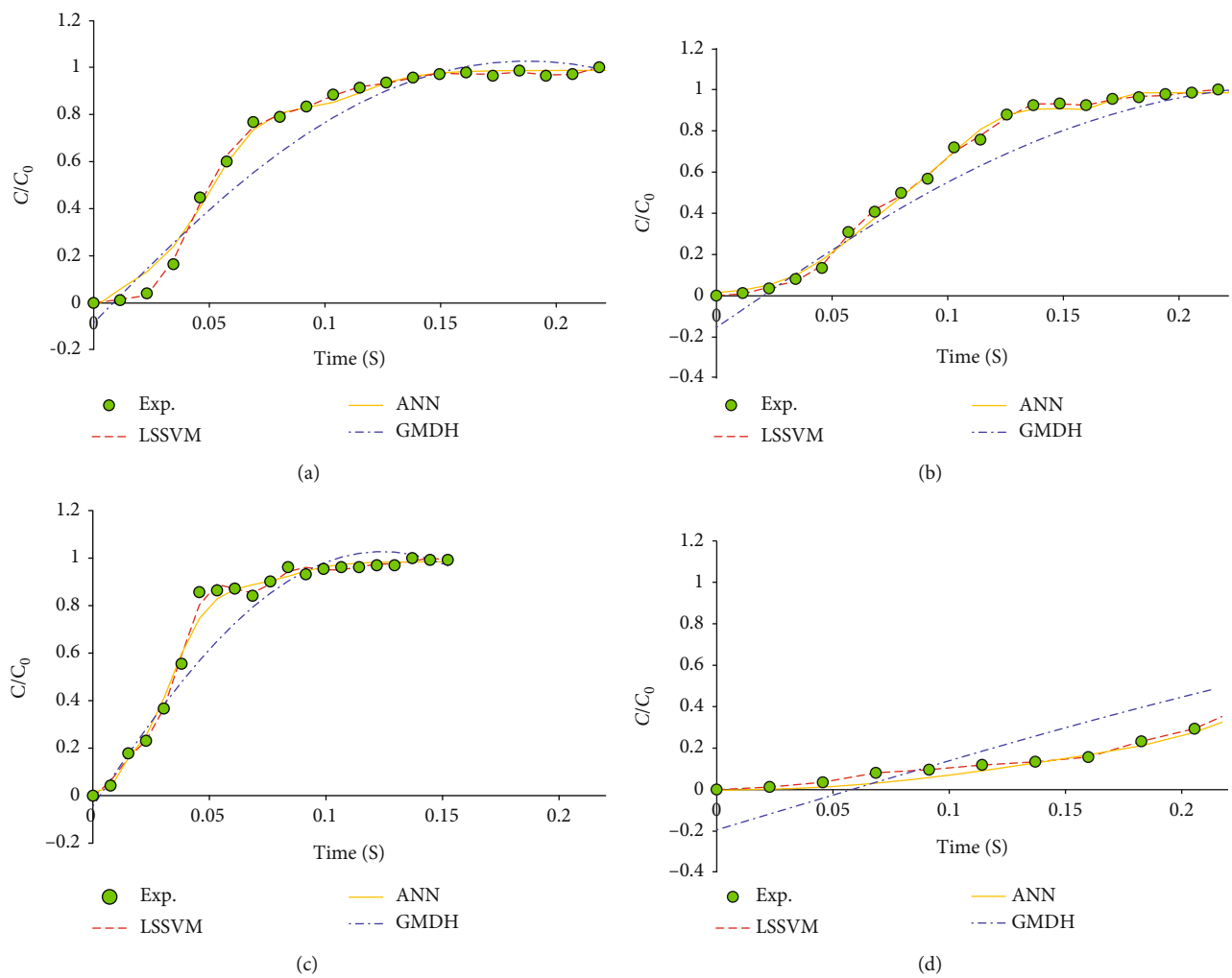


Figure 7: Predicted C/C_0 as a function of time by the three models for (a) $H_0 = 7$, $U = 160$, and $C_0 = 24$, (b) $H_0 = 11$, $U = 160$, and $C_0 = 8$, (c) $H_0 = 7$, $U = 240$, and $C_0 = 24$, and (d) $H_0 = 11$, $U = 80$, and $C_0 = 8$.

Results of the current study are compared to the LSSVM, ANN, and GMDH models [17]. To estimate the curvature of rhamnolipids on activated carbon, the structures of the proposed GMDH model are given below:

(i) First layer:

$$\begin{aligned}
\text{Node 1}: z_1 &= 0.183x_1^2 + 0.149x_2^2 + 0.165x_1x_2 \\
&\quad + -0.242x_1 - 0.048x_2 + 0.678 \\
\text{Node 2}: z_2 &= 0.141x_2^2 - 1.6x_4^2 - 0.21x_2x_4 - 0.028x_2 \\
&\quad + 2.76x_4 - 0.167 \\
\text{Node 3}: z_3 &= 0.182x_3^2 - 1.606x_4^2 - 0.124x_3x_4 \\
&\quad - 0.075x_3 + 2.778x_4 - 0.167 \\
\text{Node 4}: z_4 &= -0.002x_1^2 - 1.64x_4^2 + 0.172x_1x_4 \\
&\quad - 0.169x_1 2.684x_4 - 0.075
\end{aligned} \tag{17}$$

(ii) Second layer:

$$\begin{aligned}
\text{Node 1}: w_1 &= -0.253z_1^2 + 0.248z_2^2 - 0.319z_1z_2 \\
&\quad - 0.05z_1 + 0.948z_2 - 0.044 \\
\text{Node 2}: w_2 &= 4.923z_1^2 + 0.234z_3^2 - 0.057z_1z_3 \\
&\quad - 5.738z_1 + 0.794z_3 + 1.676 \\
\text{Node 3}: w_3 &= 2.915z_1^2 + 0.237z_4^2 - 0.535z_1z_4 \\
&\quad - 2.733z_1 + 1.097z_4 + 0.577
\end{aligned} \tag{18}$$

(iii) Third layer:

$$\begin{aligned}
\text{Node 1}: u_1 &= -0.802w_1^2 - 1.248w_2^2 + 1.972w_1w_2 \\
&\quad + 0.419w_1 + 0.666w_2 - 0.011 \\
\text{Node 2}: u_2 &= 1.433w_2^2 + 1.571w_3^2 - 3.077w_2w_3 \\
&\quad + 0.603w_2 + 0.48w_3 - 0.014
\end{aligned} \tag{19}$$

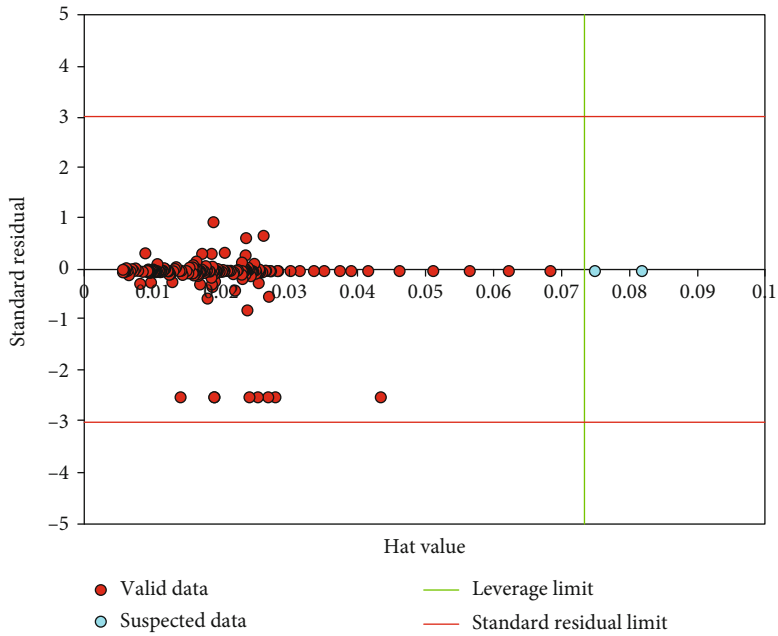


Figure 8: Diagnosis of the probable outlier data and applicability domain of the applied model.

(iv) Genome expression:

$$\text{Node 1}: \frac{C}{C_0} = -0.802u_1^2 - 1.24u_2^2 - 1.972u_1u_2 + 0.419u_1 + 0.666u_2 - 0.011 \quad (20)$$

The ANN model based on these four input variables as mentioned as follows:

(1) input layer

(2) hidden layer including six neurons

(3) output layer

Figure 5 represents the cross plot of the aforementioned strategies. As explained, data points of the LSSVM model are closer to the line $Y = X$, than ANN and GMDH models. Also, the calculation of the determination coefficient shows that the proposed LSSVM approach is superior to ANN and GMDH in terms of accuracy.

Compared to ANN and GMDH models, the less relative error is observed in the proposed LSSVM model. Figure 6 indicates more reliability of the suggested LSSVM model.

Estimation accuracy is also investigated by applying the following statistical methods:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(C_{\text{Pred}}(i) - C_{\text{Exp}}(i)\right)^2}{\sum_{i=1}^{N}\left(C_{\text{Pred}}(i) - \bar{C}_{\text{Exp}}(i)\right)^2},$$

$$\%\text{AAD} = \frac{100}{N}\sum_{i=1}^{N} C_{\text{Pred}}(i) - C_{\text{Exp}}(i),$$

$$\text{MSE} = \left(\frac{\sum_{i=1}^{N}\left(C_{\text{Pred}}(i) - C_{\text{Exp}}(i)\right)^2}{N}\right),$$

$$\text{STD} = \sum_{i=1}^{n}\left(\frac{\left(C_{\text{Pred}}(i) - \bar{C}_{\text{Exp}}(i)\right)^2}{N}\right)^{0.5}. \quad (21)$$

Table 1 presents statistical values of the presented model compared with ANN and GMDH approaches showing the higher value of R^2 and lower values of STD, AAD, and RMSE, and as a result, the LSSVM model possesses higher accuracy and reliability than others. The dependency of the (C/C_0) as an output parameter on input parameters is illustrated in Figure 7.

Four different conditions of $H_0 = 7$, $U = 160$, and $C_0 = 24$, and $H_0 = 11$, $U = 160$, and $C_0 = 8$ and $H_0 = 7$, $U = 240$, and $C_0 = 24$ and $H_0 = 11$, $U = 80$, and $C_0 = 8$ were investigated to measure the prediction ability of LSSVM, ANN, and GMDH models for indicating that the LSSVM model acquired better estimation. As this figure shows, as time goes by, the ratio of C/C_0 increases.

In the last part of this research, the leverage approach is applied to find outliers, employing the Hat matrix, Williams plot, and residuals. As discussed, Eq. (13) is used to calculate H values. Figure 8 also illustrates the Williams plot. All of the H is in the range [-3, +3], and R is in the range [0, 0.08], and then the accuracy of the proposed model is desirable and acceptable; so, the accuracy of the proposed model is satisfactory. There are only two of the data points that are outside of the applicable domain which is shown in the figure by a blue circle. As R values approach zero and H value reduces, the reliability of data points is increased [18–23].

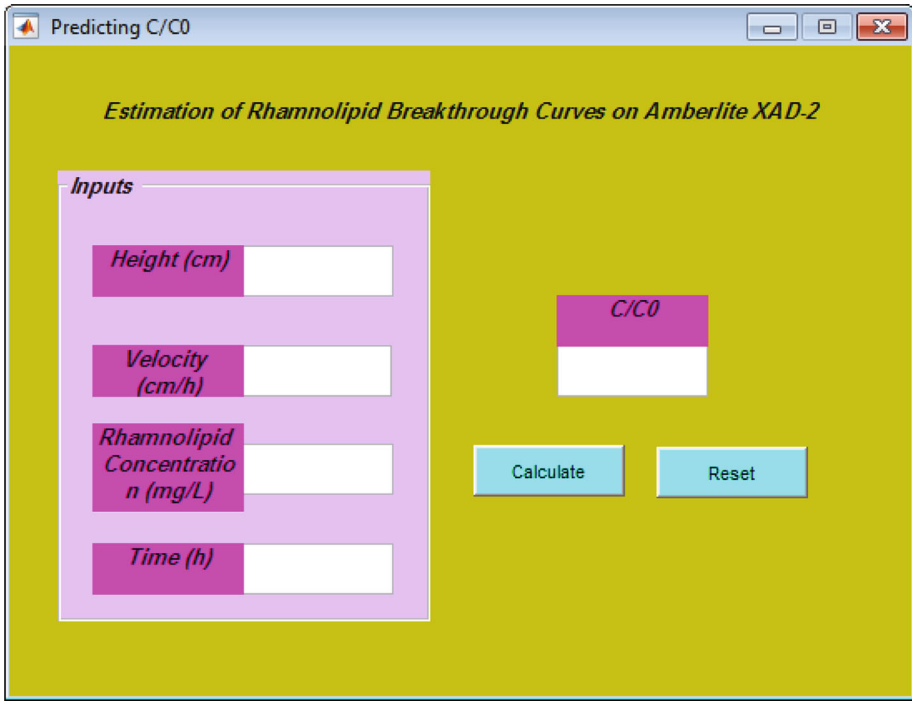


Figure 9: GUI version of the developed LSSVM model.

4. Conclusion

Then LSSVM approach was employed to estimate breakthrough curves of rhamnolipid adsorption over Amberlite XAD-2 as a function of fixed bed height, flow velocity, runtime, and initial rhamnolipid concentration. The particle swarm optimization method was employed for the training process enhancing the accuracy of the proposed model. Various statistical and graphical methods were applied to evaluate the model's reliability showing that the AAD% value for adsorption over activated carbon was 0.75%. For ANN and GMDH models that were developed by Padilha et al., AAD% of activated carbon is reported to be 1.9% and 6.2%. Based on the above evidence, we can find that the proposed LSSVM model is more reliable for the process of predicting the breakthrough curves.

Appendix

Instructions of the Developed Program

A graphical user interface (GUI) version of the model is developed (Figure 9). The code is compiled to an Exe file which is given in the supplementary content. Matlab software must be installed before running the code. It starts by giving four parameters as input; then, to show the output result, it is enough to click on the calculate button.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research presented in this paper was supported by the Funds of High-level Hospital Construction Research Project of Maoming People's Hospital.

Supplementary Materials

Table S1: experimental data points used in this study. *(Supplementary materials)*

References

[1] A. Abalos, A. Pinazo, M. Infante, M. Casals, F. García, and A. Manresa, "Physicochemical and antimicrobial properties of new rhamnolipids produced byPseudomonasaeruginosaAT10 from soybean oil refinery wastes," *Langmuir*, vol. 17, no. 5, pp. 1367–1371, 2001.

[2] A. A. Bodour and R. M. Miller-Maier, "Application of a modified drop-collapse technique for surfactant quantitation and screening of biosurfactant-producing microorganisms," *Journal of Microbiological Methods*, vol. 32, no. 3, pp. 273–280, 1998.

[3] J.-Y. Wu, K.-L. Yeh, W.-B. Lu, C.-L. Lin, and J.-S. Chang, "Rhamnolipid production with indigenous Pseudomonas aeruginosa EM1 isolated from oil-contaminated site," *Bioresource Technology*, vol. 99, no. 5, pp. 1157–1164, 2008.

[4] R. Hecht-Nielsen, "Theory of the backpropagation neural network," *Neural Networks*, vol. 1, pp. 445–448, 1988.

[5] A. G. Ivakhnenko, "Polynomial theory of complex systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. -SMC-1, no. 4, pp. 364–378, 1971.

[6] S. Atashrouz, M. Mozaffarian, and G. Pazuki, "Modeling the thermal conductivity of ionic liquids and ionanofluids based on a group method of data handling and modified Maxwell model," *Industrial & Engineering Chemistry Research*, vol. 54, no. 34, pp. 8600–8610, 2015.

[7] S. Atashrouz, H. Mirshekar, A. Hemmati-Sarapardeh, M. K. Moraveji, and B. Nasernejad, "Implementation of soft computing approaches for prediction of physicochemical properties of ionic liquid mixtures," *Korean Journal of Chemical Engineering*, vol. 34, no. 2, pp. 425–439, 2017.

[8] A. Dargahi-Zarandi, A. Hemmati-Sarapardeh, S. Hajirezaie, B. Dabir, and S. Atashrouz, "Modeling gas/vapor viscosity of hydrocarbon fluids using a hybrid GMDH-type neural network system," *Journal of Molecular Liquids*, vol. 236, pp. 162–171, 2017.

[9] A. Hemmati-Sarapardeh, M. H. Ghazanfari, S. Ayatollahi, and M. Masihi, "Accurate determination of the CO2-crude oil minimum miscibility pressure of pure and impure CO2streams: a robust modelling approach," *The Canadian Journal of Chemical Engineering*, vol. 94, no. 2, pp. 253–261, 2016.

[10] E. Mohagheghian, H. Zafarian-Rigaki, Y. Motamedi-Ghahfarrokhi, and A. Hemmati-Sarapardeh, "Using an artificial neural network to predict carbon dioxide compressibility factor at high pressure and temperature," *Korean Journal of Chemical Engineering*, vol. 32, no. 10, pp. 2087–2096, 2015.

[11] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[12] K. Pelckmans, J. A. K. Suykens, T. Van Gestel et al., *LS-SVMlab: a matlab/c toolbox for least squares support vector machines, Tutorial*, vol. 142, KULeuven-ESAT, Leuven, Belgium, 2002.

[13] J. A. K. Suykens, T. Van Gestel, and J. De Brabanter, *Least Squares Support Vector Machines*, World Scientific, 2002.

[14] M. E. Johnson, "Wiley series in probability and mathematical statistics," in *Multivariate Statistical Simulation*, pp. 231–235, John Wiley & Sons, Inc., 1969.

[15] C. R. Goodall, "13 computation using the QR decomposition," *Handbook of Statistics*, vol. 9, pp. 467–508, 1993.

[16] F. Gharagheizi, A. Eslamimanesh, M. Sattari, B. Tirandazi, A. H. Mohammadi, and D. Richon, "Evaluation of thermal conductivity of gases at atmospheric pressure through a corresponding states method," *Industrial & Engineering Chemistry Research*, vol. 51, no. 9, pp. 3844–3849, 2012.

[17] C. E. . A. Padilha, C. A. . A. Padilha, D. F. . S. Souza, J. A. de Oliveira, G. R. de Macedo, and E. S. dos Santos, "Prediction of rhamnolipid breakthrough curves on activated carbon and Amberlite XAD-2 using artificial neural network and group method data handling models," *Journal of Molecular Liquids*, vol. 206, pp. 293–299, 2015.

[18] A. Hemmati-Sarapardeh, R. Alipour-Yeganeh-Marand, A. Naseri et al., "Asphaltene precipitation due to natural depletion of reservoir: determination using a SARA fraction based intelligent model," *Fluid Phase Equilibria*, vol. 354, pp. 177–184, 2013.

[19] A. Baghban, M. A. Ahmadi, B. Pouladi, and B. Amanna, "Phase equilibrium modeling of semi-clathrate hydrates of seven commonly gases in the presence of TBAB ionic liquid promoter based on a low parameter connectionist technique," *The Journal of Supercritical Fluids*, vol. 101, pp. 184–192, 2015.

[20] A. Baghban, M. A. Ahmadi, and B. Hashemi Shahraki, "Prediction carbon dioxide solubility in presence of various ionic liquids using computational intelligence approaches," *The Journal of Supercritical Fluids*, vol. 98, pp. 50–64, 2015.

[21] A. Baghban, A. Bahadori, A. H. Mohammadi, and A. Behbahaninia, "Prediction of CO 2 loading capacities of aqueous solutions of absorbents using different computational schemes," *International Journal of Greenhouse Gas Control*, vol. 57, pp. 143–161, 2017.

[22] A. Baghban, M. Bahadori, J. Rozyn et al., "Estimation of air dew point temperature using computational intelligence schemes," *Applied Thermal Engineering*, vol. 93, pp. 1043–1052, 2016.

[23] A. Baghban, A. H. Mohammadi, and M. S. Taleghani, "Rigorous modeling of CO2 equilibrium absorption in ionic liquids," *International Journal of Greenhouse Gas Control*, vol. 58, pp. 19–41, 2017.

Hindawi
BioMed Research International
Volume 2021, Article ID 5544742, 16 pages
https://doi.org/10.1155/2021/5544742

Research Article

Lung Infection Segmentation for COVID-19 Pneumonia Based on a Cascade Convolutional Network from CT Images

Ramin Ranjbarzadeh,[1] Saeid Jafarzadeh Ghoushchi,[2] Malika Bendechache,[3] Amir Amirabadi,[4] Mohd Nizam Ab Rahman,[5] Soroush Baseri Saadi,[4] Amirhossein Aghamohammadi,[6] and Mersedeh Kooshki Forooshani[7]

[1]*Department of Telecommunications Engineering, Faculty of Engineering, University of Guilan, Rasht, Iran*
[2]*Faculty of Industrial Engineering, Urmia University of Technology, Urmia, Iran*
[3]*School of Computing, Faculty of Engineering and Computing, Dublin City University, Ireland*
[4]*Department of Electrical Engineering, Islamic Azad University, South Tehran Branch, Tehran, Iran*
[5]*Department of Mechanical and Manufacturing Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 Bangi Selangor, Malaysia*
[6]*Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia*
[7]*Department of Electronics and Telecommunications, Polytechnic University, Turin, Italy*

Correspondence should be addressed to Saeid Jafarzadeh Ghoushchi; s.jafarzadeh@uut.ac.ir

Received 25 January 2021; Revised 18 February 2021; Accepted 31 March 2021; Published 16 April 2021

Academic Editor: Alireza Baghban

The COVID-19 pandemic is a global, national, and local public health concern which has caused a significant outbreak in all countries and regions for both males and females around the world. Automated detection of lung infections and their boundaries from medical images offers a great potential to augment the patient treatment healthcare strategies for tackling COVID-19 and its impacts. Detecting this disease from lung CT scan images is perhaps one of the fastest ways to diagnose patients. However, finding the presence of infected tissues and segment them from CT slices faces numerous challenges, including similar adjacent tissues, vague boundary, and erratic infections. To eliminate these obstacles, we propose a two-route convolutional neural network (CNN) by extracting global and local features for detecting and classifying COVID-19 infection from CT images. Each pixel from the image is classified into the normal and infected tissues. For improving the classification accuracy, we used two different strategies including fuzzy c-means clustering and local directional pattern (LDN) encoding methods to represent the input image differently. This allows us to find more complex pattern from the image. To overcome the overfitting problems due to small samples, an augmentation approach is utilized. The results demonstrated that the proposed framework achieved precision 96%, recall 97%, F score, average surface distance (ASD) of 2.8 ± 0.3 mm, and volume overlap error (VOE) of $5.6 \pm 1.2\%$.

1. Introduction

Since December 2019, the world has been experiencing a new disease caused by SARS-CoV-2, which can cause asthma symptoms, acute respiratory malfunctioning, and even permanent changes to the biology of the lungs in patients regardless of their age limit. This disease was reported for the first time in Wuhan, Hubei province of China, and became a pandemic all over the world [1, 2]. The common symptoms of COVID-19 are shortness of breath, diarrhoea, coughing, sore throat, headaches, and fever. Vanishing of taste, nasal blockage, loss of smell, aches, and tiredness can also be observed in patients. The new infectious disease caused by the virus was named Coronavirus Disease 2019 (COVID-19) by the World Health Organization (WHO), and this coronavirus was named as SARS-CoV-2 by the International Committee on Taxonomy of Viruses (ICTV) [3, 4]. As there are only some definite vaccines available to prevent COVID-19, most of the

unvaccinated people can be easily infected. One of the best ways to prevent the spread of virus infection in healthy persons is isolation and diagnosis of the infected person by any possible legal approach. One of the best methods is through the X-ray or CT images of patients' chest [5–7].

Inflammation growths in the lung can pose significant risks to human health. The increasing occurrence of infected people among the population demands more effective treatments along with a cost-efficient procedure that relies on its primary diagnosis. Providing prompt and precise recognition of the infected tissue plays a key role in effective patient treatment and survival [8, 9].

A CT scan or computed tomography scan as a routine tool and a high sensitivity for the diagnosis of COVID-19 are broadly employed in hospitals and can perform early screening for the defected tissue to recognize them precisely [10–12]. Doctors and specialists are increasingly employing such imaging modality to categorize local injuries and lesions [13]. Also, due to intensity similarity between lesions and normal tissues in CT images, the precise detection and segmentation of the infected area are certainly a cumbersome task, even for experienced radiologists or doctors [14, 15]. The flow of detection and feature extraction of texture information from the lung via manual observation is a time-consuming, tedious, and monotonous process. Computer-aided diagnostic (CAD) approaches are used for such tasks and are based on artificial intelligence and machine learning algorithms to recognize the border differences between two objects. These procedures are standardizable, reproducible, and can be useful in enhancing diagnostic accuracy in a very short time. These procedures act by helping doctors and experts to accomplish accurately sophisticated tasks, employing a combination of diversity classification approaches with a practical running time [3, 16].

Image segmentation is a complex and challenging area of the biomedical engineering task that is affected by numerous aspects, including illumination, low contrast, noise, and irregularity of the objects. Segmentation refers to partitioning an image into different parts or regions based on similar characteristics in neighboring proximity.

Deep learning systems, as a prominent segment of the rising artificial intelligence (AI) technology in recent years, have been reported with significantly improved diagnostic accuracy in medical imaging [11, 17]. These intelligent systems are aiding an accelerated progress in early-stage diagnosis and treatment of many diseases including automatic detection of the liver, lung, and brain diseases [16]. Therefore, the aim of our study is to develop a deep learning model for automatic diagnosis of regions of the lungs infected with the COVID-19 virus using chest CT volumes.

Minaee et al. [18] investigated the application of deep learning structures on chest radiography images to detect COVID-19 patients. For this purpose, they employed four popular convolutional neural networks, including DenseNet-121, ResNet18, SqueezeNet, and ResNet50 to identify COVID-19 disease in the analyzed chest X-ray images. Also, transfer learning on a subset of 2000 radiograms was applied to all networks to overcome the overfitting problem and improve the models' accuracy. Fan et al. [14] applied a lung infection segmentation deep network (Inf-Net) for segmenting the infected tissue in a CT slice automatically. In the first step, a parallel partial decoder is employed for aggregating the high-level features and creates a global map. Then, to increase the accuracy, the implicit reverse attention and explicit edge-attention were incorporated into a model to segmentation the boundaries.

A 3D deep convolutional neural network (DeCoVNet) proposed in [4] for detecting COVID-19 from CT volumes. They used a pretrained UNet model to generate the 3D lung masks. The proposed DeCoVNet was divided into three stages. The first stage is called the network stem, which consisted of a vanilla 3D convolution. A batch normalization layer and a pooling layer with a kernel size of $5 \times 7 \times 7$ were used to preserve rich local visual information based on the ResNet [19] and AlexNet [20]. Also, two 3D residual blocks (ResBlocks) were employed in the second stage. Lastly, a progressive classifier (ProClf) was utilized.

Early-phase detection of Coronavirus proposed by [21] which employed five different feature extraction algorithms. To classify the extracted features, support vector machines (SVM) along with 10-fold cross-validation during the classification process were applied.

To overcome the limitations of previous works, a new hybrid algorithm for finding the location and boundary of the infected tissue from clinical CT images which takes advantage of clustering, local descriptor, and convolutional neural network is introduced. It is broadly considered to be challenging to find the exact location of the lesions inside the lung and extract their borders precisely due to the impact of the COVID-19 which caused the much similar intensity values across the lung. The growing progress of deep learning in all areas of image processing was a great motivation for this study. This work is interested to investigate the power of a CNN model for detecting and segmenting the infected regions inside the lung due to the COVID-19.

2. Methodology

The remaining parts of this paper are organized as follows. In Section 2.1, the Z score normalization technique is represented. In Section 2.2, the fuzzy clustering method is described. In Section 2.3, a local directional number patterns (LDN) encoding approach is proposed. In Section 2.4, the architecture of the convolutional neural network (CNN) is demonstrated. In Section 2.5, our CNN pipeline is represented. The explanation of the dataset, evaluation metrics, and experimental results are clarified in Section 3. Our algorithm is displayed in Figure 1.

2.1. Image Normalization. As indicated in [22], due to the presence of the statistical noise in the computed tomography images (CT images), a deviation in the Hounsfield units (HUs) about a mean can be observed that lead to a high variance in the gray scale or RGB values of all image pixels. These unwanted noises that affect the ability to visualize anatomic structures can be categorized into three main sources: (1) electronic noise that is an unwanted disturbance in an electrical signal caused by electrical equipment in the

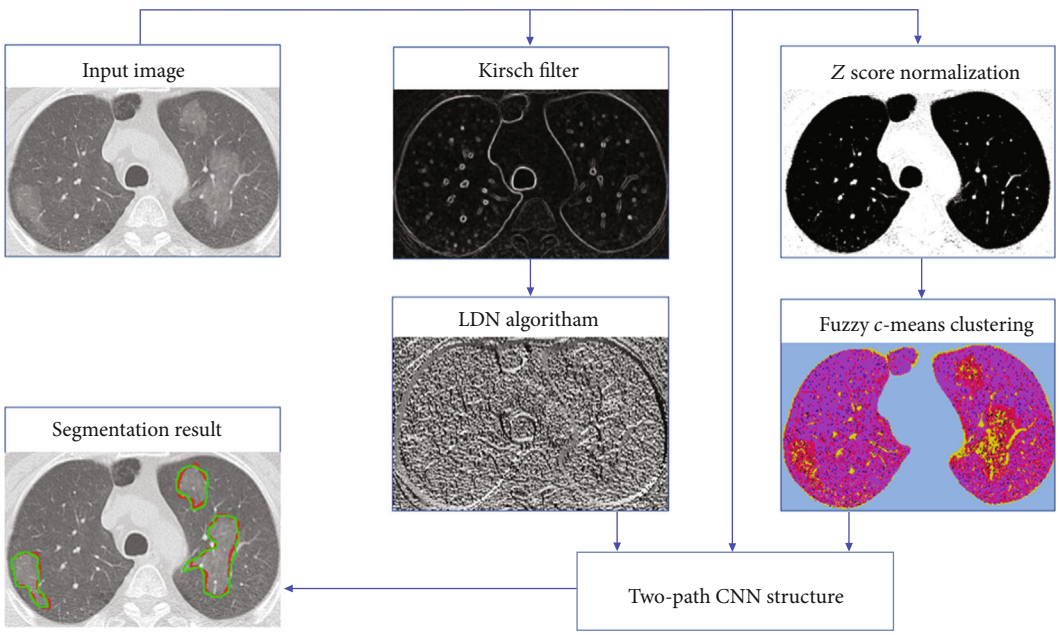


Figure 1: Schematic of the proposed pipeline for segmentation of the infected tissues.

neighborhood, (2) noise of the reconstruction procedure caused by imperfections in the receiver coils, and (3) stochastic noise.

As the stochastic noise is the principal source of noise in these kinds of imaging, the bad effects can be diminished during the imaging procedure by increasing the amount of photons (by considering a tradeoff between radiation risk and image quality). However, in obtained images from any hospital or medical center, a significant amount of noise is observed which needs to be removed before starting the process of the segmentation.

By further investigation, we found out that a normalization approach can be beneficial to create a smooth image along with increasing the contrast of illumination near the border of the organs. So, to overcome the mentioned problems and enhance the result of the segmentation, a Z score normalization technique is employed so that all the nonzero values inside the image have a unit variance and zero mean ([23–25]; Jafarzadeh [26]). Equation (1) outlines how to apply Z score normalization.

$$Z = \frac{(x - \mu)}{\sigma} \tag{1}$$

where σ and μ indicate the standard deviation and mean value of nonzero pixels, respectively. Moreover, x describes the intensity of the current pixel.

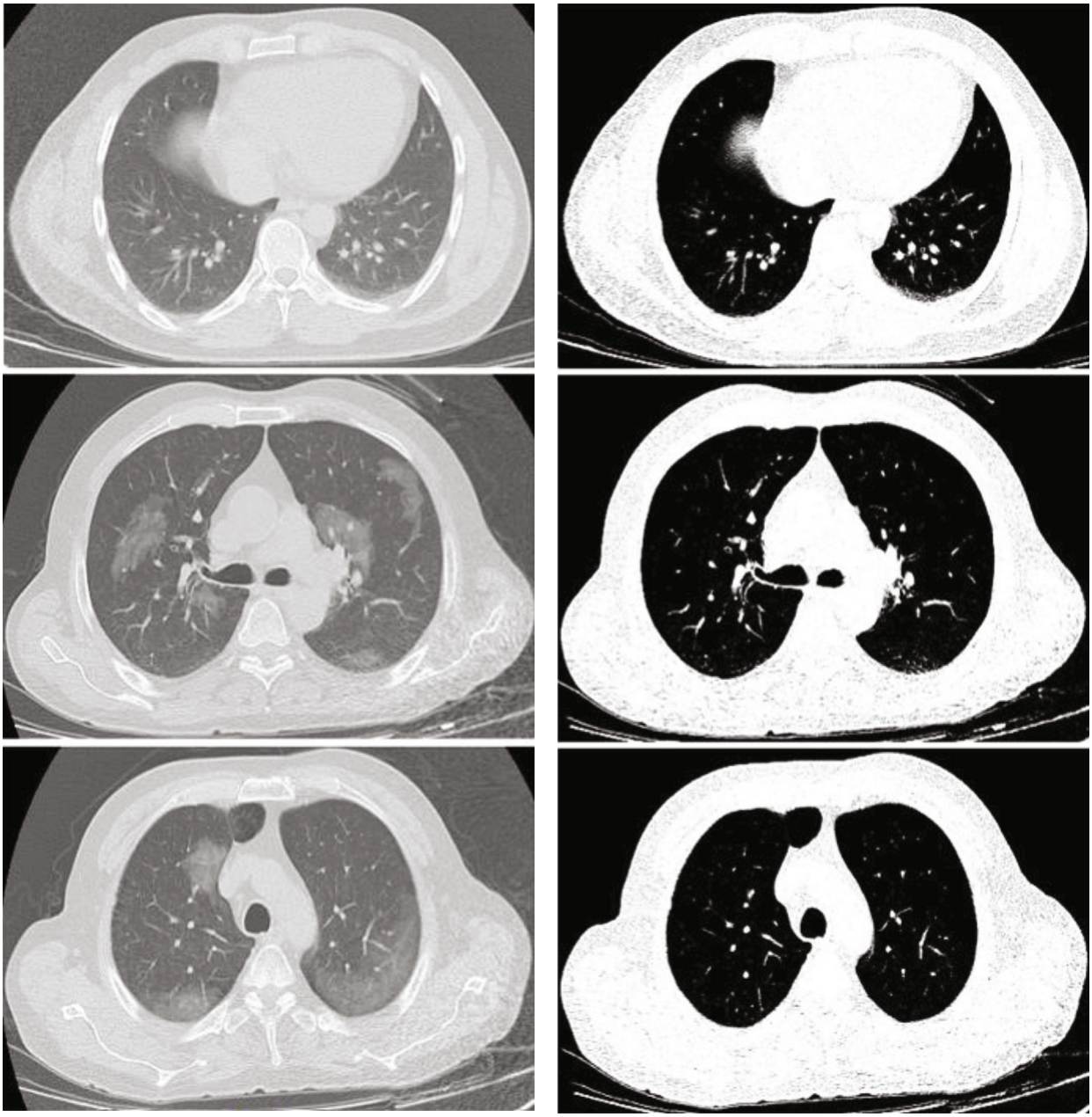
The outcomes of the normalization strategy are depicted in Figure 2. In Figure 2, the first column shows the chest CT images of patients, and their corresponding lesions in the second column demonstrates the Z score output. As illustrated in Figure 2(b), the borders of both the lungs are detected exactly without the effect of the lesions.

2.2. Fuzzy c-Means. After detecting the borders of the lungs with high accuracy, we need to recognize the volume and border of the infected areas inside the lungs more efficiently. The image of the detected lungs achieved from the previous stage has to be clustered to segment the infected areas from the other organs (background tissue). Clustering can be outlined as an unsupervised strategy that is aimed at fragmenting the input data (image or signal etc.) into the predefined segments (such as K-means method) or automated recognize parts (such as mean-shift method) based on certain criteria such as differences in the color, magnitude, and location [27–30]. The fuzzy c-means (FCM) algorithm used in our work is an unsupervised data dividing/splitting strategy. In this method, data is split into n predefined natural groupings, namely, the so-called clusters such that every single pixel in the dataset be owned by at least two clusters with dissimilar weights. In this fuzzy partitioning technique, finding the cluster center of each segment and related pixels are accomplished through an iterative optimization of the objective function [31–33]. This iterative optimization is accomplished by minimizing the following membership cost/objective function:

$$E = \sum_{k=1}^{m} \sum_{i=1}^{N} \mu_{ki} \|\text{pixel}_i - \text{center}_k\|^2, \tag{2}$$

$$\mu_{ki} = 1/\left(\sum_{j=1}^{m} = 1\left(\frac{\text{pixel}_i\text{-center}_k}{\text{pixel}_i\text{-center}_j}\right)\right)^t, \sum_{k=1}^{m} \mu_{ki} = 1, \mu_{ki} \in [0, 1], \tag{3}$$

where center_k shows the center of the kth cluster and pixel_i illustrates the ith sample of I, μ_{ki} outlines the membership value of the ith sample with respect to the kth cluster which is linked inversely to the distance from pixel_i to the cluster center center_k, m defines the number of clusters, t refers to



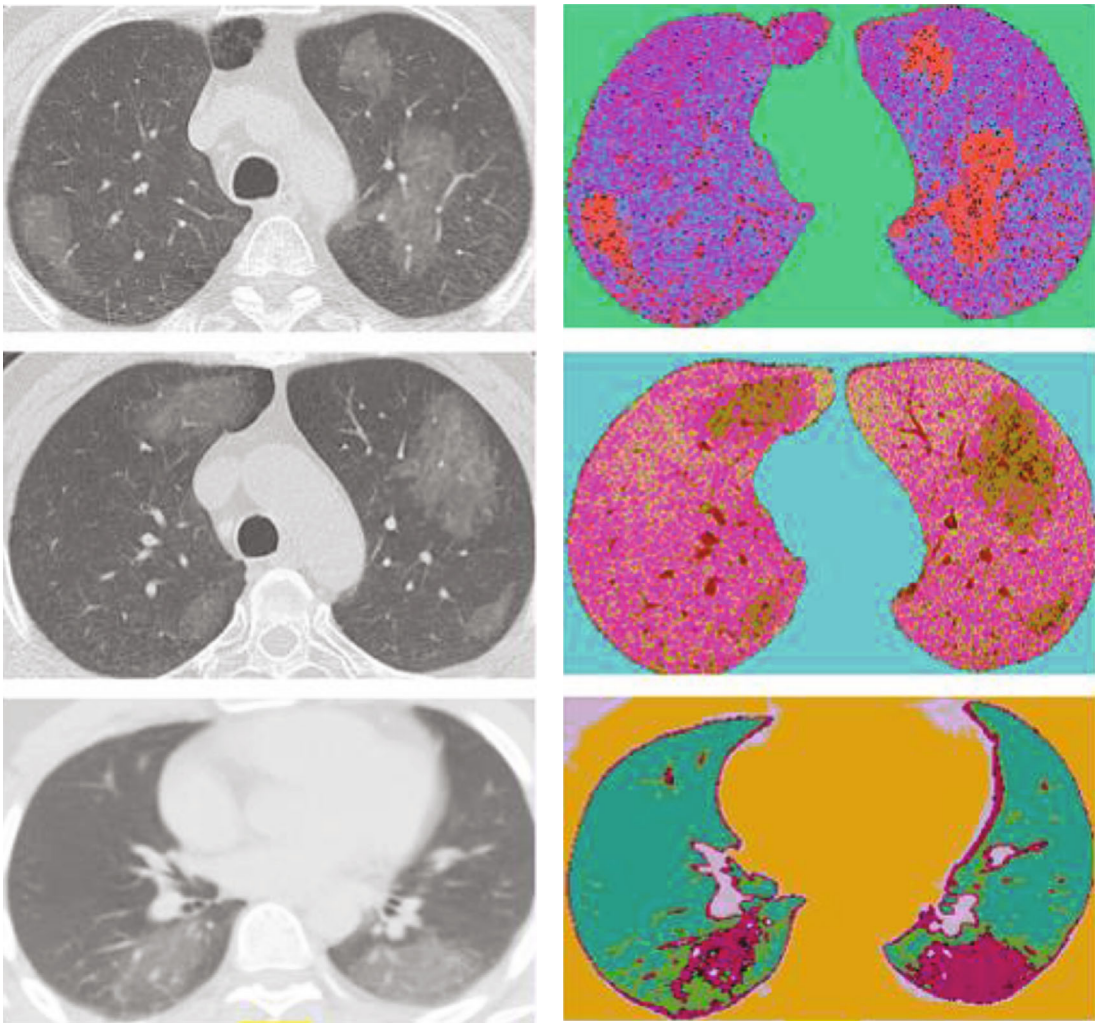
(a) (b)

Figure 2: A demonstration of employing Z score normalization approach. (a) Original images. (b) Z score normalization.

the level of cluster fuzziness, and N denotes the number of pixels image pixels I.

The result of the clustering on the lung's image is represented in Figure 3. For better visualization, we applied a random value to each cluster in the RGB domain. As is illustrated clearly, by defining the number of five for the center of clusters by experimental results, a high distinction between the lesion and normal tissue can be observed in many samples. It means the number of clusters more or less than five cannot obtain an acceptable result. However, as depicted in Figure 4, in some CT images due to much color similarity between the normal and lesion tissues, using only a clustering method to segment the lesions is not optimal. So, in the next step, textural analysis approaches will be employed to improve segmentation accuracy as much as possible.

2.3. Local Directional Number Pattern. Textural analysis of medical and biological images attempts to mine some characterizations of a surface texture such as smoothness, roughness, contrast, colors, and shapes [34]. As presented in many works [35, 36], numerous types of local descriptors are used for converting images into a new representation based on the predefined coding rules or codebook of visual patterns.



(a) (b)

Figure 3: A demonstration of employing fuzzy c-means clustering technique. (a) Original images. (b) Clustered images. For better understanding, the colors of the clusters are in the RGB domain with random values.

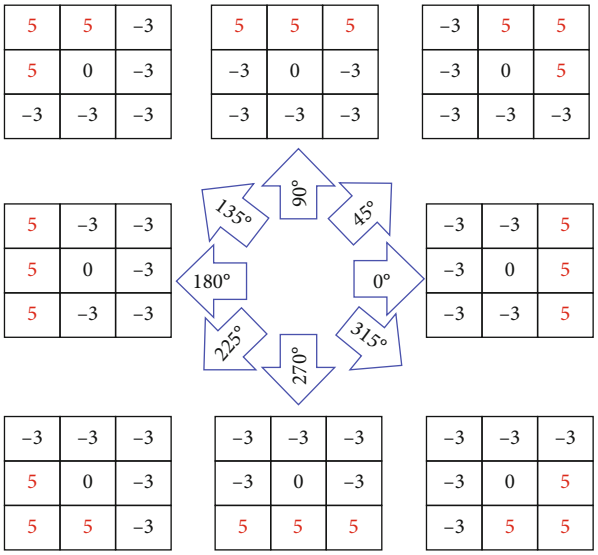
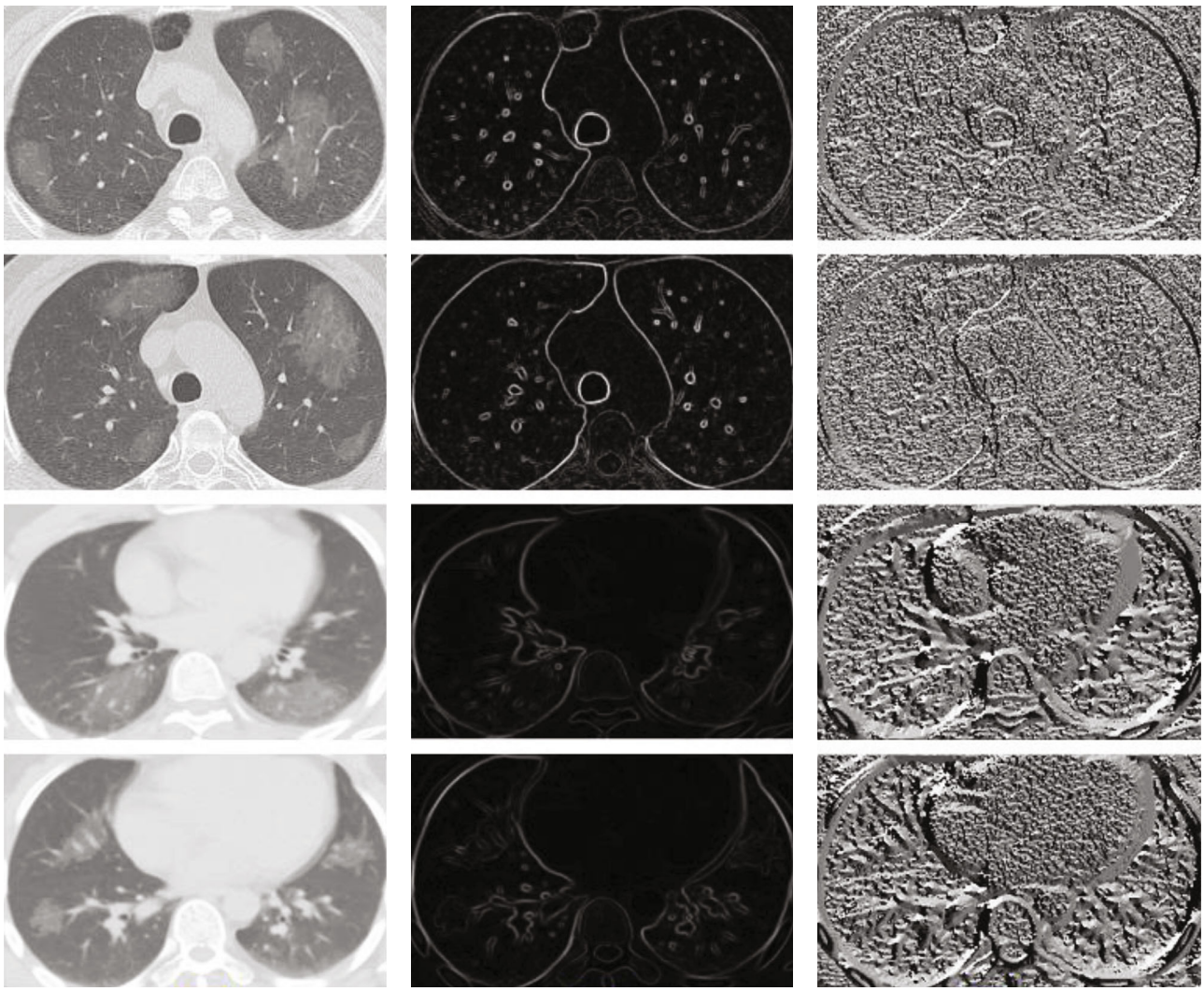


Figure 4: Nonlinear Kirsch kernels in 8 rotations [15].

Local ternary patterns (LTP) and local binary pattern (LBP) feature descriptors are easy to implement and be influenced by the change of the pixel intensity of nearest-neighbor (circular, rectangular, etc. neighborhood) in clockwise or counterclockwise to alter (encoding) the low-level information of a spot, edges, curve, and line inside an image and calculate the outcome as a binary value [37, 38]. Owing to the robustness of the gradient value than a gray level intensity in encoding applications, in recent investigations, some techniques based on the gradient value such as local word directional pattern (LWDP) and local directional number patterns (LDN) have attained much attention [36]. The LDN operates in the gradient domain to create an illumination-invariant representation of the image. It uses directional information for recognizing edge locations that their magnitudes are insensitive to lighting variations.

In our work, the first phase for encoding the chest images is to define the location and value of all significant edges. This is implemented by operating 8 directions of Kirsch kernels (filters) that are rotated by 45° in 8 main compass directions



(a) (b) (c)

Figure 5: The result of applying the Kirsch filter and LDN approach to a chest image. The second column illustrates edge detection using the Kirsch filter. The third column demonstrates the results of the LDN technique.

(Figure 4). These nonlinear edge detector kernels are responsible for identifying the final edges. Each filter produces a feature map, and only the maximum value in each location is selected to create a final edge map [39, 40]. An example of employing the nonlinear Kirsch filter to the chest images is depicted in Figure 5. This section causes a substantial increase in final lesion segmentation, especially when the border of the lesions is vague.

2.4. Convolutional Neural Network Design. Automated recognition of patterns in data by computers based on knowledge already obtained is called pattern recognition. It has applications in image analysis, information retrieval, signal processing, bioinformatics, data compression, statistical data analysis, computer graphics, and machine learning [27, 31, 33, 41–44].

In machine learning approaches and applications, the convolutional neural network (CNN) structures demonstrate a high capability to extract and classify some key features and bridging the gap between the capabilities of machines and humans [45–47]. The structure of a CNN was inspired by the organization of the visual cortex in the human brain and is similar to that of the connectivity pattern of neurons. Every neuron responds to an irritant only in a constrained region of the visual field known as the receptive field. The CNN structure that is originally designed for image analysis largely exploits the low level and high level of the textural features and is used in many applications including action detection and automated lesion segmentation [48, 49].

This neuron-based pipeline that captures temporal and spatial dependencies has a grid-like topology and permits us for extracting characteristics powerfully from the 1D or 2D input data by passing through a stack of convolution layers with the predefined dimension of the filters [36, 50, 51]. This grid-like model is a class of deep learning networks and has numerous trainable biases and weights based on the type of the topology and is applied for feature extraction,

regression, and classification. These trainable weights need to be defined randomly at the beginning.

This structure is able to extract high-level features automatically from raw input features, which are considerably powerful than human-designed features. The core building block of a CNN is outlined as the convolutional layer which calculates the dot product between input data and a set of learnable filters, much like a traditional neural network [49, 52–54]. It should be noticed that the dimension of the filters is smaller than the dimension of the input data [49, 55]. The computed feature maps using the convolutional layer are achieved by stacking the activation maps of all kernels along the depth dimension. The output of one kernel (filter) applied to the previous layer is called the feature map. In the convolving process, for controlling the dimension of the feature maps, padding the input data with zeros around the border can be employed.

Mostly, the spatial-temporal dependencies at various scales are able to be effectively obtained by the convolutional layers. The dimension of the kernel which defines the dimension of the receptive field needs to be selected based on the depth of the applied 1D, 2D, or 3D data. Also, stride defines how much the convolution filter can be moved at each step. Moreover, the bigger strides lead to less overlap between the receptive fields (smaller feature map) [55].

The high-level features are extracted (such as the hand, legs, and, body in pedestrian detection) in the deeper convolutional layers of the model, while the first convolutional layers are responsible for mining the low-level information including curves, edges, and points. It should be mentioned that the numbers of columns and rows for each filter need to be an odd number, for instance, 9×9, 7×7, and 3×3 [54].

It is noteworthy that the dimension of the extracted features in the last convolutional layer is greatly smaller than the input matrix (1D or 2D matrix). The diminution in the width and height of the image relies upon the length of the strides and the filter size employed for the convolution procedure.

The output of the convolution layer is fed to the activation layer in order to help the network learn complex patterns [56]. This layer leaves the size of the applied matrix (data) unchanged. To decrease the consequence of the vanishing gradient in the training process, an activation function is utilized for each feature map to improve the computational effectiveness by inducing sparsity [55, 57].

In this study, the nonlinearity (ReLU) activation function has been employed to shift the negative values to zero. The ReLU act as a linear function for the positive and zero values. As all negative values change to the zero number, it leads some nodes to completely die and not learn anything. It means fewer neurons in the model would activate because of the limitations imposed by this layer.

Some of the most important benefits of the ReLU layer can be expressed as follows [58–60]:

(1) Train deep networks: the architecture with large labeled datasets is able to reach the best performance on purely supervised tasks

(2) Linear behavior: the procedure of decreasing the cost function (optimization) in the CNN is much easier if their behavior could be close to a linear manner

(3) Representational sparsity: as the ReLU layer shift the negative input values to the zero values, it causes some of the neurons in the hidden layers in neural networks to have zero values. In other words, by removing the effect of some neurons with zero weight, an accelerating in the learning process can be achieved which is called a sparse representation

(4) Computational simplicity: dissimilar to the tan h and Sigmoid activation functions, ReLU consists of only simple operations in terms of computation so that computing the exponential function in activations can be eliminated and therefore much more practicable to implement in models

The ReLU layer does not cause the vanishing gradient problem and avoid easy saturation. Also, due to overcoming the vanishing gradient issue, models are permitted to learn faster and perform better. Equation (2) outlines how the ReLU activation function accomplishes [58, 59].

$$f(x) = \max(0, x), \tag{4}$$

where x demonstrates the input value and $f(x)$ is its related output.

Since in object recognition applications, there is evidence that demonstrates the form, dimension, color, or position of the object has no matter, only the spatial variances need to be investigated. In order to accomplish this, a downsampling layer is applied by summarizing the key information in patches of each feature map without losing any details that lead to a good classification. In contrast to the convolution operation, the pooling layer has no parameters and only slides a window over its input, and simply takes the predefined value (mean, max, etc.) in the window. Furthermore, as the quantity of pixels in this layer (in both row and column) is dropped, it leads to shortening the training time and combats overfitting [54, 61–63].

An appropriate technique for dimensionality reduction of feature maps is to reduce the number of parameters and computation in the network so that the model can be robust to alter the high-frequency information (key information) and preserves vital features [55]. This dimension-reduction procedure happens by utilizing a filter along the spatial dimensions (width, height) with a predefined dimension. This layer is regularly incorporated between two sequential convolutional layers. The max pooling layer accomplished in this study first partitions the extracted matrix of features into a set of parts with no overlapping and then takes the maximum number inside each district. The max pooling strategy also employs as a noise suppression technique [53, 64].

In a CNN structure (shallow or deep CNN), since the receptive field in the last convolutional layer does not cover the entire spatial dimension of the image, the generated features by the last convolutional layer correspond to a section

of the input image. Therefore, one or some FC layers are obligatory in such a scenario. A fully connected layer (FC) allows the model to learn the nonlinear combinations of the high-level features in an input image.

Each node in the fully connected layer produces a single output with its learnable corresponding weight that is linked to all the activations in the previous nodes [56]. It is noteworthy that before applying the generated feature matrixes to the fully connected layer, all 2D features have to be changed into a one-dimensional matrix (1D vector) [65–67]. The latest layer for classification tasks in a CNN-based pipeline is the Softmax regression layer which is able to differentiate one from the other. The Softmax regression is also called multinomial logistic, multiclass logistic regression, or just maximum entropy classifier. This single-layer regression tries to normalize an input value into a vector of values to demonstrate how likely the input data belongs to a user-defined class. Also, as the output values are between the range (0, 1), the sum of the output values obtained from the probability distribution procedure is equal to one [52, 53, 67, 68].

For the training step, since we are not working with a big dataset with hundreds of different samples from many patients, it is enormously easy for the CNN-based models to converge or to be specialized according to its reliability level and application area (to be less intelligent). To overcome this issue, there are two main strategies: (1) transform learning and (2) data augmentation.

The transform learning method is utilized to bring some trained biases and weights into any pipeline rather than select them randomly at the first step. Data augmentation is a popular method for artificially boosting the number of training examples [69, 70].

2.5. Our CNN Pipeline. As mentioned before, CNNs are used to explore significant details from an input of raw pixels more efficiently. Hence, in this study, we investigated the probability of the presence of the lesions caused by CVOID-19 using a novel model based on the combination of global and local features. Moreover, to maximize the segmentation accuracy for even small damaged healthy tissue, the proposed approach concludes three distinct input images instead of a single one. The three input images include original image, fuzzy clustered image, and encoded image (LDN). These three different inputs enable our model to handle many types of variability in the raw input pixels. The flowchart of the proposed structure is shown in detail in Figure 6.

When we use CNNs for automatic feature extracting that are effective for various tumor or lesion detection problems, the need for preprocessing and highlighting the suspect regions is significantly reduced. This is due to the fact that the CNN-based structures have millions of parameters that are able to produce the best suited feature maps for expressing the class probability. Although numerous CNN pipelines have been recommended for lesion segmentation in recently published papers, none of them has concentrated on applying the combined the textural encoding algorithm, fuzzy clustered, and raw image pixels as an input to a CNN structure. Since miscellaneous texture or images definitely encompass complementary and detailed information (features), our experimental outcomes for small samples (data) imply that this complex two-path strategy is effective to enhance the score of the evaluation indexes.

While analyzing the complex texture of our input images, due to many similarities of the lesion (infected area) with normal tissue in the margin of the lesion, semiglobal and local features must be taken into account. Moreover, the lesions may appear anywhere on the lung since COVID-19 has a multifocal distribution that for gaining better results need to have knowledge of neighbor information in a little further of each analyzing pixel location. As is clearly indicated in Figure 6, the recommended cascading model is based on investigating key features using two distinct local and global paths.

In contrast to some other recently published methods such as studies by Hu et al. [71], Wang et al. [4], and Fan et al. [14] that employ all pixels inside the image as an input, our method only considers two patches from each applied 2D data (totally 6 patches) as an input to classify each pixel inside the output image. In other words, if there are 1000 pixels inside the image, the number of the produced patches are 1000×2, and due to the use of the three input images, there are $1000 \times 2 \times 3$ patches. This is very interesting that using both local and global patches with a different route for extracting features can get better results compared to using only one of them.

In our model, two distinct routes are employed; the first one (upper path) comprises of the five convolutional layers for extracting the global features. The other path (bottom path) utilizes two convolutional layers for extracting the local features. The local and global investigation windows (patches) are 25×25 and 60×60, respectively.

The semiglobal patches are employed for providing key details about the analogous touching textures with scar tissues, while the local patches are applied more for recognizing inflammation in the tiny air sacs. Moreover, the outcome of our strategy for inflammation detection highly depends on information extracted from the global windows. In Table 1, we exhibit the effect of employing semiglobal and local patches in the ultimate outcome of our approach. As is depicted in Table 1, the best observed Dice score is obtained when the sizes of the local and global patch are 25×25 and 60×60, respectively.

The size of the local region is $25 \times 25 \times 3$, which three implies three distinct input images. The selected regions are convolved using 64 kernels to generate the feature maps based on the 3×3 receptive field. In the next layer, the number of filters is changed to 128 with the same receptive field. After producing feature maps in the first layer, the max pooling layer is not, while after the second layer, max pooling decreases the dimension of the produced feature maps.

Unlike the local features extraction path, in the global feature extraction procedure, five convolutional layers are employed. In this path, only two intermediate layers are employed that are using the max pooling approach. All extracted feature maps with the size of 9×9 at the end of each route are concatenated to create 384 feature maps in order to use in the next convolutional layer. After the concatenation step, 128 kernels are applied to these feature maps, and then, a max-pooling layer changes the all dimensions to the 4×4.

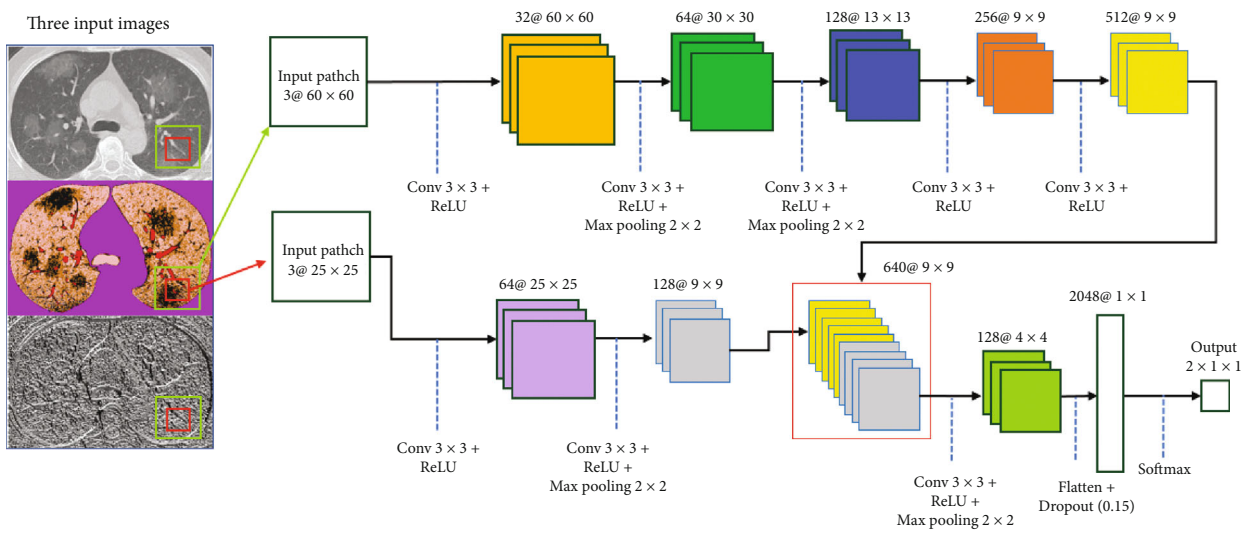


Figure 6: Our implemented two-path CNN model using three distinct inputs.

Table 1: Investigating the accuracy of employing dissimilar dimensions of the regions in the final result of the approach.

Size of the semiglobal patch	Size of the local patch	DICE value for lesion segmentation
40 × 40	11 × 11	24%
50 × 50	11 × 11	31%
60 × 60	11 × 11	33%
70 × 70	11 × 11	40%
80 × 80	11 × 11	41%
40 × 40	15 × 15	61%
50 × 50	15 × 15	70%
60 × 60	15 × 15	72%
70 × 70	15 × 15	73%
80 × 80	15 × 15	81%
40 × 40	21 × 21	74%
50 × 50	21 × 21	76%
60 × 60	21 × 21	81%
70 × 70	21 × 21	88%
80 × 80	21 × 21	91%
40 × 40	25 × 25	56%
50 × 50	25 × 25	73%
60 × 60	25 × 25	92%
70 × 70	25 × 25	89%
80 × 80	25 × 25	87%

Then, all created feature maps are transformed into a 2048 × 1 feature vector. Lastly, by applying a Softmax layer, all extracted data are tagged to one of two expected classes (1 implies the inflammation and 2 shows the normal tissues.).

For minimizing the cross-entropy loss, the proposed CNN structure with two routes was learned through stochastic gradient descent (SGD) in 1000 epochs with a batch size of 128 [72], in Equation (5). Our pipeline calculates the discrepancy between the predicted output and groundtruth for lesion segmentation. The dropout is applied before the FC layer, which is aimed at avoiding "overfitting" and equals to 0.2. For optimization, we applied a weight decay of 0.0001 and a learning rate of 0.01. In the output layer, two logistic units to obtain the probabilities of the given sample belonging to either of the two classes were employed. The backpropagation scheme was applied to generate the derivative of the objective function.

$$\text{loss}_i = -\log\left(\frac{e^U k}{\sum_{d=1}^{L} e^U d}\right), \tag{5}$$

where loss_i implies the loss value for training data i, and U_K demonstrates the raw production score (is not normalized) for the reference class K.

The unnormalized production score is generated by multiplying the outputs from the previous FC layer with the parameters of the corresponding logistic unit. To find the normalized scores for each class between 0 and 1, the denominator aggregates the scores for all the logistic units L. Since two output neurons are presented at the output layer, in the above equation, L is equal to 2.

3. Experiments

3.1. Datasets. The proposed novel technique and three recently published models were investigated on a public chest dataset [73] to evaluate the reliability, validity, and accuracy of experiments. This dataset is available at https://github.com/UCSD-AI4H/COVID-C. To segment the corrupted tissues accurately, four experienced specialists segmented the borders manually. It is noteworthy that by employing an augmentation strategy to increase the number of data, a lot of new samples are generated. Also, 70% of data for training, 10% for validating, and 20% for testing are used. Data augmentations are useful approaches to decrease the validation

and training errors. The augmentation methods artificially inflate the training dataset size by either data oversampling or warping. When in the augmentation process, the labels of the existing images are preserved; this process is called data warping augmentations. This method includes augmentations such as color and geometric transformations, adversarial training, random erasing, and neural style transfer. Oversampling augmentations generate synthetic samples and add them to the training set [74].

Six approaches of data augmentation are utilized in this paper to increase efficiency, namely, flipping, color space, rotation, translation, noise injection, color space transformations, and random erasing.

In flipping, a horizontal axis flipping is used. In the color space, contrast enhancing is employed. In rotation, 180 degrees is selected. In translation, left, right, up, and down are applied. In noise injection, a Gaussian distribution is utilized. In the color space transformations, decreasing and increasing the pixel values by a constant value are applied. In random erasing, an $n \times m$ patch of an image is randomly selected and masking it with zero values.

3.2. Evaluation Metrics. In this study, the following nine measures were calculated by comparing the segmentation results with that of lesions segmented by the experts to appraise the proposed architecture's efficiency. The promising accuracy of the proposed two-path architecture was assessed using recall, precision, F score, ASD (average surface distance), RVD (relative volume difference), RMSD (root mean square symmetric surface distance), MSD (maximum surface distance), VOE (volume overlap error), and DICE (Dice similarity) [15, 75–77]. Some mentioned metrics are defined as follows:

$$\begin{cases} \text{Precision} = \dfrac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%, \\ \text{Recall} = \dfrac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \\ \text{F} = \dfrac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\%, \\ \text{DICE} = \left(2 \times \dfrac{\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}\right) \times 100\%, \\ \text{VOE}(M_{s1}, M_{s2}) = \left(1 - \dfrac{M_{s1} \cap M_{s2}}{M_{s1} \cup M_{s2}}\right) \times 100\%, \\ \text{RVD}(M_{s1}, M_{s2}) = \left(\dfrac{M_{s1} - M_{s2}}{M_{s2}}\right) \times 100\%, \\ \text{ASD} = \dfrac{1}{\left|B_{M_{s1}}\right| + \left|B_{M_{s2}}\right|} \times \left(\sum\limits_{x \epsilon B_{M_{s1}}} d\left(x, B_{M_{s2}}\right) + \sum\limits_{y \epsilon B_{M_{s2}}} d\left(y, B_{M_{s1}}\right)\right), \end{cases} \tag{6}$$

where M_{s1} and M_{s2} denote the result of segmentation using our strategy and ground-truth mask, respectively. Also, $B_{M_{s1}}$ and $B_{M_{s2}}$ imply the borders result of our segmentation technique and ground-truth image, respectively. Moreover, the FN, FP, and TP represent false negative, false positive, and true positive, respectively [37, 78].

Dice similarity coefficient (DSC) is defined as one for a perfect segmentation and is a statistical tool for measuring the similarity between two sets of data. MSD measures the distance between the borders of each segmented object from its corresponding border in the groundruth image. Measuring the difference between the segmented object and related object in the groundtruth image can be calculated by RVD, in which the positive value implies oversegmentation and the negative value represents the undersegmentation result. It means that the best value is zero that indicates the segmented object is equal to the groundtruth image.

3.3. Experimental Results. Our two-path architecture was implemented in Python, and the experiments were run on an Intel(R) Core(TM)i7-3.4 GHz + GEFORCE GTX 1080 Ti GPU+16 gigabytes of RAM under the windows 10 (64-bit) operating system. The results of our pipeline using 3 distinct input images were appraised utilizing the corresponding ground-truths and reported in Tables 2 and 3. In our dataset samples with a large diversity in the volume of the lesions, not well-defined borders (unclear or blurred margin) have the greatest part of the train, validation, and test samples.

For exemplifying the significance of utilizing the grouping of the LDN encoding approach, Z score normalization technique, and CNN framework to accurate estimating borders, Figure 7 demonstrates the outcomes of our structure (drawn by a green line). The results of our method compared to three other recently published methods are shown in Figure 7 on a few slices with the intensity inhomogeneity, ambiguous boundaries, heterogeneous appearances, and various infection shapes. Accordingly, it can noticeably be observed that the intensity inhomogeneity and ambiguous boundaries inside the lung due to the infection cause the infected regions are not suitably extracted when the DenseNet201 [1], weakly supervised deep learning [71], and weakly supervised framework [4] approaches are applied.

As indicated in Figure 7, segmentation by employing the DenseNet201 [1] structure shows the fewest match with the reference data (groundtruth), especially when similar intensity values are encountered near the borders of the infected regions. Weakly supervised deep learning [71] is good to recognize the infection boundary when there is much distance (more than 20 pixels) between two lesions, but when in the small distance (less than 20 pixels), it performs so poorly and the chance of combining two lesions is highly increased. Also, the DenseNet201 [1] method undersegment the infected areas in the most cases, whereas the weakly supervised deep learning [71] and weakly supervised framework [4] models oversegment with equivalent intensity values. Moreover, such pipelines are more prone to boundary leakage, especially when there are unclear borders among the different kinds of infection progress. To solve this issue, we came up with the idea of employing both local and global features when there are three representations of the infected and noninfected tissues. Our model also has not noteworthy boundary leakage, substantial oversegmentation, or undersegmentation, predominantly in particular sections that are near the white objects. By using the Z score normalization and fuzzy clustering methods, our approach is more capable

Table 2: Quantitative comparison of infected tissue segmentation outcomes based on our model and three recently published structures. The evaluations are based on average surface distance (ASD), relative volume difference (RVD), Volume overlap error (VOE), root mean square symmetric surface distance (RMS), and maximum surface distance (MSD).

Approach	ASD (mm)	VOE (%)	RVD (%)	MSD (mm)	RMS (mm)
DenseNet201 [1]	5.4 ± 0.3	11.4 ± 7.3	−4.2 ± 5.9	23.6 ± 7.1	5.9 ± 0.4
Weakly supervised deep learning [71]	5.1 ± 0.4	11 ± 7.3	7.8 ± 10.3	21 ± 6.6	5.5 ± 0.7
Weakly supervised framework [4]	6.1 ± 0.6	11.7 ± 4.2	8.3 ± 6.6	22.7 ± 5.2	5.8 ± 0.5
Proposed CNN	6.3 ± 0.5	11.9 ± 6.8	−5.8 ± 3.5	21.3 ± 6.1	5.7 ± 0.4
Proposed CNN+LDN	5.1 ± 0.1	8.3 ± 4.7	6.5 ± 4.1	15.4 ± 4.8	4.7 ± 0.2
Proposed CNN+fuzzy *c*-means	5.5.3 ± 0.4	8.9 ± 5.2	−6.9 ± 7.3	16.5 ± 4.9	5.2 ± 0.5
Proposed CNN+fuzzy *c*-means+LDN	**2.8 ± 0.3**	**5.6 ± 1.2**	**3.7 ± 5.6**	**7.4 ± 7.3**	**3.6 ± 0.2**

Table 3: Quantitative comparison of infected tissue segmentation outcomes based on our pipeline and three recently published structures. The evaluations are based on recall, precision, and *F* score.

Approach	Precision (%)	Recall (%)	*F* score
DenseNet201 [1]	86%	89%	87%
Weakly supervised deep learning [71]	88%	90%	89%
Weakly supervised framework [4]	91%	89%	90%
Proposed CNN	88%	89%	88%
Proposed CNN+LDN	93%	91%	92%
Proposed CNN+fuzzy *c*-means	92%	94%	93%
Proposed CNN+fuzzy *c*-means +LDN	**96%**	**97%**	**97%**

of enhancing the contrast near the border of the lung to obtain more accuracy in the distinction of an infected region and vague border of the lung. Considering the heterogeneous textures, opaque appearance of the infected tissue, misalignment of the infection boundaries, unclear borders, and different dimensions of the infection regions, it is more evident that our pipeline suitably finds a pattern most similar to the infected area, which demonstrates its robust performance under realistic scenarios on countless infection outlines. It worth mentioning that in all methods, the white tissue (pulmonary nodules) near the infected area cannot be properly recognized due to much similarity between both tissue values. The results may get better if the amount of training data is increased.

The proposed two-path CNN structure achieved a higher segmentation performance than the other three evaluated methods when other representations of the lung images are applied; meaning more substantial features are available to achieve the best distinction between classes. The efficiency of our technique on different CT infected lungs was assessed using the Dice similarity index, as illustrated in Figure 8. The Dice score averages for the segmented infection areas with diverse appearance varied from 80% to 94%. As is shown, the worst result belongs to the DenseNet201 approach with an average of 84%. The result of our approach implies that the appearance, intensity values, and outline of the infected tissue cannot significantly affect the segmentation performance and efficiency.

Tables 2 and 3 indicate the comprehensive evaluation of our complex strategy for lesions segmenting and compare it with the results of other mentioned methods on our dataset.

Table 2 implies a quantitative comparison, in practice, between the automated lesion segmentation outcomes of the novel proposed two-patch model over the other three mentioned approaches. For each index in Tables 2 and 3, the highest values of RVD, ASD, RMS, MSD, VOE, recall, precision, and *F* score are highlighted in bold. The outcomes of every first five assessment criteria are demonstrated by standard deviation and mean values in Table 2. The proposed two-route segmentation model gains a smaller mean in mentioned assessment criteria. The obtained VOE is meaningfully altered between all appraised architectures, while the outcomes of RMS and ASD imply the lowest variance. The RVD score for DenseNet201, proposed CNN, and proposed CNN+fuzzy *c*-means algorithms are less than 0. Also, adding the LDN method to the proposed CNN model leads to observe the positive value in the RVD result. The RMS score imply that the proposed CNN+fuzzy *c*-means+LDN and proposed CNN+LDN methods produced the best outcomes among the seven structures. Also, the DenseNet201 technique gains the highest mean score of RMS.

In addition, the mean value of MSD and VOE of the models employed by DenseNet201 and weakly supervised framework were outstandingly higher as compared to our outcomes. Moreover, both the weakly supervised deep learning and the weakly supervised framework models show a large standard deviation in the RVD; however, a major standard deviation in MSD score is obtained in the DenseNet201 method. The observed results in the ASD and VOE indicate that adding LDN and fuzzy clustering methods to our CNN model can significantly improve our model accuracy.

The results in Table 3 indicate the measurements for differentiating the objects inside the lung, including normal and infected tissues. As can be observed in Table 3, our technique, CNN+fuzzy *c*-means+LDN, consistently performs the best among all approaches. The *F* score, precision, and recall of the DenseNet201 and weakly supervised deep learning structures are highly similar to the proposed CNN algorithm; however, by adding the LDN or fuzzy clustering approach, these three criteria are highly increased. Also, the

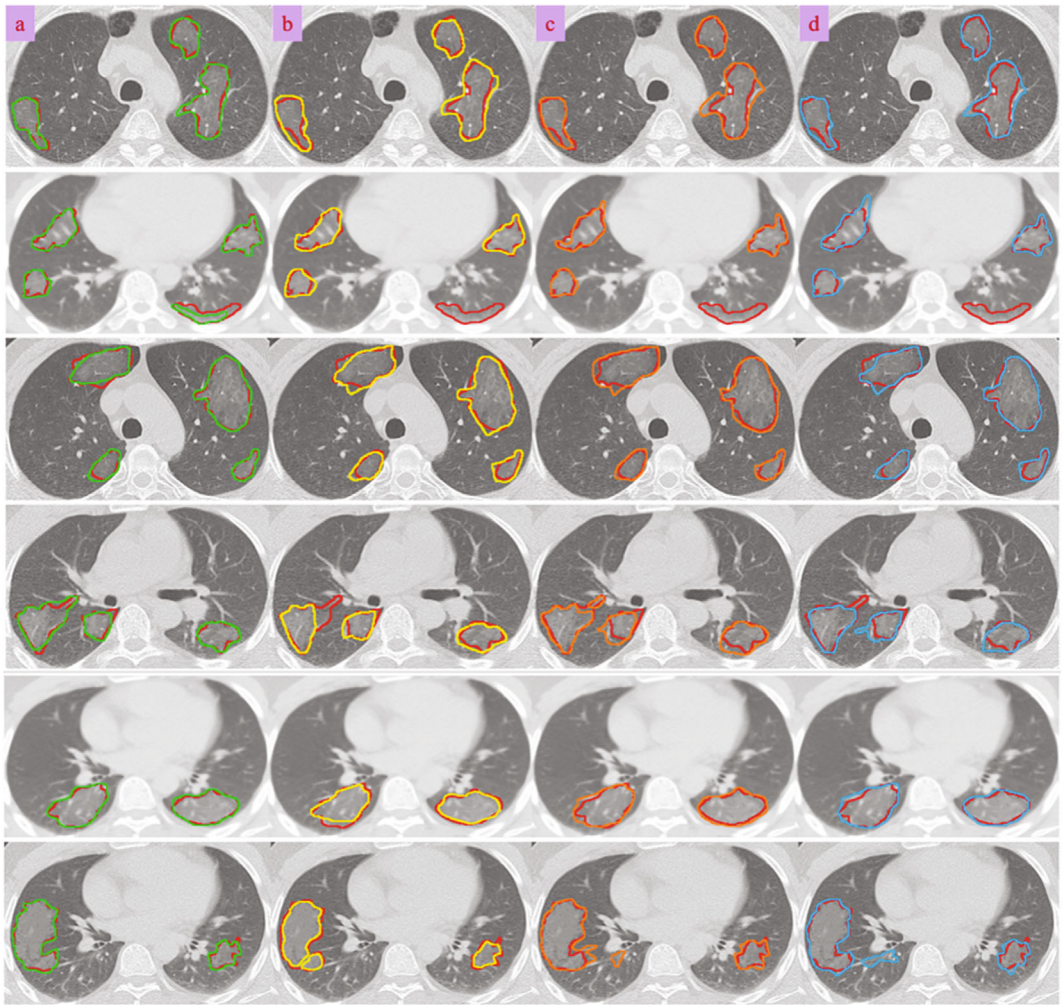


Figure 7: Comparisons between four different kinds of strategies for COVID-19 infection detection. The red contours indicate the reference border (groundtruth). Segmentation based on the (a) proposed strategy, (b) DenseNet201 [1], (c) weakly supervised deep learning [71], and (e) weakly supervised framework [4].

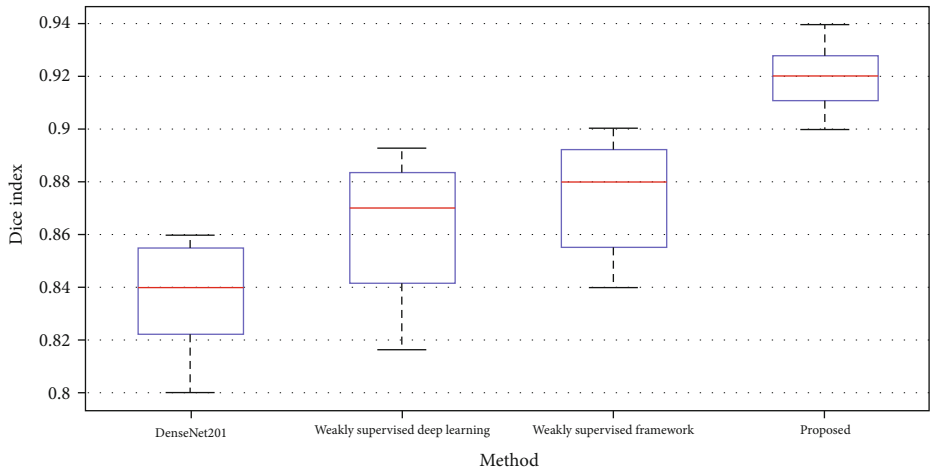


Figure 8: Comparison between the Dice scores of the four models employed for lung infection segmentation in CT images.

DenseNet201 approach gains the worst results and our architecture obtains the competitive performance on lesions segmentation in all evaluation metrics.

4. Discussion and Conclusions

In this study, we implemented a two-path CNN pipeline that incorporates the three distinct input images, to automatically segment the infected tissues inside the lung caused due to the COVID-19 from CT images. For a better demonstration of the tissues to extract more key features inside the CNN model, we showed the input CT image represented in the two other different ways which each of them includes some unique information. Due to inflammation inside the lung because of COVID-19, infected areas near the border of the lung are highly difficult to segment. So, our algorithm first employed a Z score normalization technique to obtain a more distinguishable lung border from the original image. Then, by using a fuzzy clustering method, all tissues in the image are clustered and obtain a distinct pixel value for all pixels corresponding to each cluster. This approach helps the CNN pipeline for decreasing the convolutional layers for extracting some key features and leads to a drop in the training time of the pipeline and increase the final efficiency.

Then, an LDN encoding approach was implemented for representing the information of the images in another form to extract more essential details from the input image. This strategy roots in the fact that sometimes by changing the representation domain (like frequency domain rather than the time domain) some other substantial features can be observed.

We also represented a new two-route CNN model that considered semiglobal and local information to categorize each pixel in the input image to one of the two normal and infected tissues. The number of the convolutional layers in the global route is more than the local route, while the kernel size for all convolutional layers is the same. To overcome the overfitting problems and boost efficiency, using data augmentation methods, the number of samples has been increased. Lastly, using the CT image and two obtained images, our CNN structure was trained.

The suggested two-route segmentation pipeline was appraised on a public dataset which 70% of data for training, 10% for validating, and 20% for testing were used. Our significant findings demonstrate that our CNN pipeline and three distinct input images gained the following: (1) acceptable performance even if the infected area shared an extended border with touching tissues, (2) appropriately robust as indicated by the negligible standard deviations which show the uniformity of the values for all the nine criteria, and (3) accomplished well in the detection and segmentation process even for the intricate cases with numerous unalike categories of the infection, which had the amoeboid shapes and analogous thicknesses.

The proposed architecture satisfactorily overcomes the difficulty of failing in accurate detection of the lesions at the presence of the similar adjacent tissues and identification of an uneven border where it seemed to not properly appear to exist with an aim to reach superior outcomes. In addition, the employed technique does not require more extra parameters for feeding into the algorithm apart from one CT image to define the position of the lesions and border detection. But the functional limitation of this architecture is that the white matter (pulmonary nodules) inside the normal lung near the border of a lesion cannot properly be recognized from the infected tissue. We think that by increasing the training samples this problem can be solved.

Tables 2 and 3 approve that our technique divides erratic and wide infections and irregular shapes. Most of the segmentation strategies that merely rely on measuring the illumination, energy, thickness, location, and shape could fail when the infected tissue and other touching objects have an analogous density and intensity levels. Under such specific circumstances, applying additional distinguishable features from different kinds of images may result in improving the ability of segmentation and fulfilled a leading role in gently separating infections associated with the abovementioned problems. Our unique pipeline could potentially be more advantageous when encountering diverse infections with the blurred boundaries and wide-ranging lesion sizes. The implemented procedure proposed herein yields a more classification efficiency in terms of simplicity, stability, and time consumption compared to the baseline models.

Data Availability

The data used to support the findings of this study are included within the article (https://github.com/UCSD-AI4H/COVID-CT).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning," *Journal of Biomolecular Structure and Dynamics*, vol. 2, pp. 1–8, 2020.

[2] F. Shan, Y. Gao, J. Wang et al., "Lung infection quantification of COVID-19 in CT images with deep learning," 2020, arXiv preprint arXiv:2003.04655.

[3] M. Ahmadi, A. Sharifi, S. Dorosti, S. Jafarzadeh Ghoushchi, and N. Ghanbari, "Investigation of effective climatology parameters on COVID-19 outbreak in Iran," *Science of the Total Environment*, vol. 729, p. 138705, 2020.

[4] X. Wang, X. Deng, Q. Fu et al., "A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2615–2625, 2020.

[5] S. Dorosti, S. Jafarzadeh Ghoushchi, E. Sobhrakhshankhah, M. Ahmadi, and A. Sharifi, "Application of gene expression programming and sensitivity analyses in analyzing effective parameters in gastric cancer tumor size and location," *Soft Computing*, vol. 24, no. 13, pp. 9943–9964, 2020.

[6] B. Kamble, S. P. Sahu, and R. Doriya, "A review on lung and nodule segmentation techniques," in *Advances in Data and*

Information Sciences, Lecture Notes in Networks and Systems, pp. 555–565, Springer, Singapore, 2020.

[7] L. Zhou, Z. Li, J. Zhou et al., "A rapid, accurate and machine-agnostic segmentation and quantification method for CT-based COVID-19 diagnosis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2638–2652, 2020.

[8] J. Chen, L. Wu, J. Zhang et al., "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020, medRxiv: 2020.02.25.20021568.

[9] A. Hamzenejad, S. J. Ghoushchi, V. Baradaran, and A. Mardani, "A robust algorithm for classification and diagnosis of brain disease using local linear approximation and generalized autoregressive conditional heteroscedasticity model," *Mathematics*, vol. 8, no. 8, p. 1268, 2020.

[10] G. D. Rubin, C. J. Ryerson, L. B. Haramati et al., "The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society," *Chest*, vol. 158, no. 1, pp. 106–116, 2020.

[11] F. Shi, J. Wang, J. Shi et al., "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19," *IEEE Reviews in Biomedical Engineering*, vol. 14, no. 1, pp. 4–15, 2020.

[12] G. Wang, X. Liu, C. Li et al., "A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2653–2663, 2020.

[13] A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem, and A. Mohammadi, "Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks," *Computers in Biology and Medicine*, vol. 121, p. 103795, 2020.

[14] D. P. Fan, T. Zhou, G. P. Ji et al., "Inf-Net: automatic COVID-19 lung infection segmentation from CT images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[15] R. Ranjbarzadeh and S. B. Saadi, "Automated liver and tumor segmentation based on concave and convex points using fuzzy _c_ -means and mean shift clustering," *Measurement: Journal of the International Measurement Confederation*, vol. 150, p. 107086, 2020.

[16] X. Ouyang, J. Huo, L. Xia et al., "Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2595–2605, 2020.

[17] V. Rajinikanth, N. Dey, A. N. J. Raj, A. E. Hassanien, K. C. Santosh, and N. S. M. Raja, "Harmony-search and Otsu based system for coronavirus disease (COVID-19) detection using lung CT scan images," 2020, arXiv preprint arXiv:2004.03431.

[18] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Jamalipour Soufi, "Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning," *Medical Image Analysis*, vol. 65, p. 101794, 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[21] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) classification using CT images by machine learning methods," 2020, arXiv preprint arXiv:2003.09424.

[22] M. Willner, G. Fior, M. Marschner et al., "Phase-contrast Hounsfield units of fixated and non-fixated soft-tissue samples," *PLoS One*, vol. 10, no. 8, article e0137016, 2015.

[23] L. Friedman and O. V. Komogortsev, "Assessment of the effectiveness of seven biometric feature normalization techniques," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2528–2536, 2019.

[24] S. Jafarzadeh Ghoushchi, M. N. Ab Rahman, D. Raeisi, E. Osgooei, and M. Jafarzadeh Ghoushji, "Integrated decision-making approach based on SWARA and GRA methods for the prioritization of failures in solar panel systems under Z-information," *Symmetry*, vol. 12, no. 2, p. 310, 2020.

[25] S. V. Khond, "Effect of data normalization on accuracy and error of fault classification for an electrical distribution system," *Smart Science*, vol. 8, no. 3, pp. 117–124, 2020.

[26] S. J. Ghoushchi, K. Gharibi, E. Osgooei, M. N. Ab Rahman, and M. Khazaeili, "Risk prioritization in failure mode and effects analysis with extended SWARA and MOORA methods based on Z-numbers theory," *Informatica*, vol. 32, no. 1, pp. 41–67, 2020.

[27] M. Bendechache, *Study of distributed dynamic clustering framework for spatial data mining [Ph.D. thesis]*, University College Dublin. School of Computer Science, 2019.

[28] M. Bendechache, M. T. Kechadi, and N. A. Le-Khac, "Efficient large scale clustering based on data partitioning," in *Proceedings -3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, hlm*, pp. 612–621, Fuzhou, China, 2016.

[29] M. Bendechache, N. A. Le-Khac, and M. T. Kechadi, "Hierarchical aggregation approach for distributed clustering of spatial datasets," in *IEEE International Conference on Data Mining Workshops, ICDMW, hlm*, pp. 1098–1103, Barcelona, Spain, 2016.

[30] R. Ranjbarzadeh and S. Baseri Saadi, "Corrigendum to "Automated liver and tumor segmentation based on concave and convex points using fuzzy c-means and mean shift clustering" [Measurement 150 (2020) 107086]," *Measurement: Journal of the International Measurement Confederation*, vol. 151, p. 107230, 2020.

[31] M. Bendechache and M. T. Kechadi, "Distributed clustering algorithm for spatial data mining," in *ICSDM 2015- Proceedings 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, pp. 60–65, Fuzhou, China, 2015.

[32] N. Dhanachandra and Y. J. Chanu, "An image segmentation approach based on fuzzy c-means and dynamic particle swarm optimization algorithm," *Multimedia Tools and Applications*, vol. 79, no. 25–26, pp. 18839–18858, 2020.

[33] R. Ranjbarzadeh, S. B. Saadi, and A. Amirabadi, "LNPSS: SAR image despeckling based on local and non-local features using patch shape selection and edges linking," *Measurement: Journal of the International Measurement Confederation*, vol. 164, p. 107989, 2020.

[34] T. Tuncer, S. Dogan, and F. Ozyurt, "An automated residual exemplar local binary pattern and iterative ReliefF based COVID-19 detection method using chest X-ray image," *Chemometrics and Intelligent Laboratory Systems*, vol. 203, p. 104054, 2020.

[35] C. Leng, H. Zhang, B. Li, G. Cai, Z. Pei, and L. He, "Local feature descriptor for image matching: a survey," *IEEE Access*, vol. 7, pp. 6424–6434, 2019.

[36] F. Naiemi, V. Ghods, and H. Khalesi, "A novel pipeline framework for multi oriented scene text image detection and recognition," *Expert Systems with Applications*, vol. 170, p. 114549, 2021.

[37] N. Karimi, R. Ranjbarzadeh Kondrood, and T. Alizadeh, "An intelligent system for quality measurement of golden bleached raisins using two comparative machine learning algorithms," *Measurement: Journal of the International Measurement Confederation*, vol. 107, pp. 68–76, 2017.

[38] L. Liu, S. Lao, P. W. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Median robust extended local binary pattern for texture classification," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1368–1381, 2016.

[39] Y. T. Luo, L. Y. Zhao, B. Zhang et al., "Local line directional pattern for palmprint recognition," *Pattern Recognition*, vol. 50, pp. 26–44, 2016.

[40] M. Z. Uddin, M. M. Hassan, A. Almogren, M. Zuair, G. Fortino, and J. Torresen, "A facial expression recognition system using robust face features from depth videos and deep learning," *Computers and Electrical Engineering*, vol. 63, pp. 114–125, 2017.

[41] W. A. Ali, K. N. Manasa, M. Bendechache, M. F. Aljunaid, and P. Sandhya, *A review of current machine learning approaches for anomaly detection in network traffic*, Telecommunications Association Inc, 2020.

[42] S. R. de Assis Neto, G. L. Santos, E. da Silva Rocha et al., "Detecting human activities based on a multimodal sensor data set using a bidirectional long short-term memory model: a case study," in *Challenges and Trends in Multimodal Fall Detection for Healthcare*, pp. 31–51, Springer, Cham, 2020.

[43] H. Azary and M. Abdoos, "A semi-supervised method for tumor segmentation in mammogram images," *Journal of Medical Signals and Sensors*, vol. 10, no. 1, pp. 12–18, 2020.

[44] T. Nasir, M. Asmael, Q. Zeeshan, and D. Solyali, "Applications of machine learning to friction stir welding process optimization," *Jurnal Kejuruteraan*, vol. 32, no. 1, pp. 171–186, 2020.

[45] M. Z. Islam, M. M. Islam, and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," *Informatics in Medicine Unlocked*, vol. 20, article 100412, 2020.

[46] S. Jafarzadeh-Ghoushchi and M. N. A. Rahman, "Performance study of artificial neural network modelling to predict carried weight in the transportation system," *International Journal of Logistics Systems and Management*, vol. 24, no. 2, pp. 200–212, 2016.

[47] A. Waleed Salehi, P. Baglat, and G. Gupta, "Review on machine and deep learning models for the detection and prediction of coronavirus," *Materials Today: Proceedings*, vol. 33, pp. 3896–3901, 2020.

[48] S. Hassantabar, M. Ahmadi, and A. Sharifi, "Diagnosis and detection of infected tissue of COVID-19 patients based on lung X-ray image using convolutional neural network approaches," *Chaos, Solitons and Fractals*, vol. 140, p. 110170, 2020.

[49] A. Mahmood, M. Bennamoun, S. An et al., "Deep learning for coral classification," in *Handbook of Neural Computation*, pp. 383–401, Elsevier Inc, 2017.

[50] M. Nour, Z. Cömert, and K. Polat, "A novel medical diagnosis model for COVID-19 infection detection based on deep features and Bayesian optimization," *Applied Soft Computing Journal*, vol. 97, article 106580, 2020.

[51] F. Ucar and D. Korkmaz, "COVIDiagnosis-Net: deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images," *Medical Hypotheses*, vol. 140, p. 109761, 2020.

[52] J. Chen, Z. Liu, H. Wang, A. Nunez, and Z. Han, "Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 2, pp. 257–269, 2018.

[53] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "ABCNN: attention-based convolutional neural network for modeling sentence pairs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 259–272, 2016.

[54] J. Zhong, Z. Liu, Z. Han, Y. Han, and W. Zhang, "A CNN-based defect inspection method for catenary split pins in high-speed railway," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 8, pp. 2849–2860, 2019.

[55] Y. Bengio, *Practical recommendations for gradient-based training of deep architectures, hlm*, Springer, Berlin, Heidelberg, 2012.

[56] A. D. Torres, H. Yan, A. H. Aboutalebi, A. Das, L. Duan, and P. Rad, "Patient facial emotion recognition and sentiment analysis using secure cloud with hardware acceleration," in *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications*, pp. 61–89, Elsevier, 2018.

[57] J. Dolz, C. Desrosiers, and I. Ben Ayed, "3D fully convolutional networks for subcortical segmentation in MRI: a large- scale study," *NeuroImage*, vol. 170, pp. 456–470, 2018.

[58] N. Calik, M. A. Belen, and P. Mahouti, "Deep learning base modified MLP model for precise scattering parameter prediction of capacitive feed antenna," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 33, no. 2, 2020.

[59] A. Dureja and P. Pahwa, "Analysis of non-linear activation functions for classification tasks using convolutional neural networks," *Recent Patents on Computer Science*, vol. 12, no. 3, pp. 156–161, 2018.

[60] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *Annals of Statistics*, vol. 48, no. 4, pp. 1875–1897, 2020.

[61] S. Di Cataldo and E. Ficarra, *Mining textural knowledge in biological images: applications, methods and trends*, Elsevier B.V, 2017.

[62] N. Dong, M. Kampffmeyer, X. Liang, Z. Wang, W. Dai, and E. Xing, "Reinforced auto-zoom net: towards accurate and fast breast cancer segmentation in whole-slide images," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 317–325, Springer Verlag, Cham, 2018.

[63] Z. Liu, Y. Q. Song, V. S. Sheng et al., "Liver CT sequence segmentation based with improved U-Net and graph cut," *Expert Systems with Applications*, vol. 126, pp. 54–63, 2019.

[64] A. Doğantekin, F. Özyurt, E. Avcı, and M. Koç, "A novel approach for liver image classification: PH-C-ELM,"

Measurement: Journal of the International Measurement Confederation, vol. 137, pp. 332–338, 2019.

[65] F. Ettensperger, "Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field," *Quality and Quantity*, vol. 54, no. 2, pp. 567–601, 2020.

[66] F. Liu, G. Lin, and C. Shen, "CRF learning with CNN features for image segmentation," *Pattern Recognition*, vol. 48, no. 10, pp. 2983–2992, 2015.

[67] R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian, "Benign and malignant breast tumors classification based on region growing and CNN segmentation," *Expert Systems with Applications*, vol. 42, no. 3, pp. 990–1002, 2015.

[68] M. Havaei, A. Davy, D. Warde-Farley et al., "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[69] N. Dvornik, J. Mairal, and C. Schmid, "On the importance of visual context for data augmentation in scene understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 1, pp. 1–15, 2019.

[70] T. He, W. Huang, Y. Qiao, and J. Yao, "Text-Attentional convolutional neural network for scene text detection," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2529–2541, 2016.

[71] S. Hu, Y. Gao, Z. Niu et al., "Weakly supervised deep learning for COVID-19 infection detection and classification from CT images," *IEEE Access*, vol. 8, pp. 118869–118883, 2020.

[72] N. Wahab, A. Khan, and Y. S. Lee, "Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection," *Computers in Biology and Medicine*, vol. 85, pp. 86–97, 2017.

[73] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-dataset: a CT scan dataset about COVID-19," 2020, arXiv preprint arXiv:2003.13865.

[74] T. Bahadur Chandra, K. Verma, B. Kumar Singh, D. Jain, and S. Singh Netam, "Coronavirus disease (COVID-19) detection in chest X-ray images using majority voting based classifier ensemble," *Expert Systems with Applications*, vol. 165, article 113909, 2021.

[75] M. Liao, Y. Q. Zhao, W. Wang et al., "Efficient liver segmentation in CT images based on graph cuts and bottleneck detection," *Physica Medica*, vol. 32, no. 11, pp. 1383–1396, 2016.

[76] X. Lu, J. Wu, X. Ren, B. Zhang, and Y. Li, "The study and application of the improved region growing algorithm for liver segmentation," *Optik*, vol. 125, no. 9, pp. 2142–2147, 2014.

[77] R. Suresh, A. N. Rao, and B. E. Reddy, "Detection and classification of normal and abnormal patterns in mammograms using deep neural network," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 14, 2019.

[78] X. Xu, X. Jiang, C. Ma et al., "A deep learning system to screen novel coronavirus disease 2019 pneumonia," *Engineering*, vol. 6, no. 10, pp. 1122–1129, 2020.