**Indonesian Journal of Electrical Engineering and Computer Science**
Vol. 22, No. 1, April 2021, pp. 241~251
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i1.pp241-251


# Efficient intelligent system for diagnosis pneumonia (SARS-COVID19) in X-ray images empowered with initial clustering

**Salam Saad Mohamed Ali[1], Ali Hakem Alsaeedi[2], Dhiah Al-Shammary[3], Hassan Hakem Alsaeedi[4], Hadeel Wajeeh Abid[5]**
[1]Babylon Education Directorate, Ministry of Education, Iraq
[2,3]College of Computer Science and Information Technology University of Al- Qadisiyah, Iraq
[4,5]Al-Diwaniyah Teaching Hospital, Directorate of Al-Diwaniyah Health, Iraq

**Article Info**

***Article history:***

Received Oct 10, 2020
Revised Dec 20, 2020
Accepted Jan 14, 2021

***Keywords:***

COVID-19
Features extraction
Features selection
Machine learning
Metaheuristic optimization

**ABSTRACT**

This paper proposes efficient models to help diagnose respiratory (SARS-COVID19) infections by developing new data descriptors for standard machine learning algorithms using X-Ray images. As COVID-19 is a significantly serious respiratory infection that might lead to losing life, artificial intelligence plays a main role through machine learning algorithms in developing new potential data classification. Data clustering by K-Means is applied in the proposed system advanced to the training process to cluster input records into two clusters with high harmony. Principle component analysis (PCA), histogram of orientated gradients (HOG) and hybrid PCA and HOG are developed as potential data descriptors. The wrapper model is proposed for detecting the optimal features and applied on both clusters individually. This paper proposes new preprocessed X-Ray images for dataset featurization by PCA and HOG to effectively extract X-Ray image features. The proposed systems have potentially empowered machine learning algorithms to diagnose pneumonia (SARS-COVID19) with accuracy up to %97.



***Corresponding Author:***

Ali Hakem Alsaeedi
College of Computer Science and Information technology
University of Al-Qadisiyah, Iraq
Email: ali.alsaeedi@qu.edu.iq

## 1. INTRODUCTION

Prognostication and cure of respiratory infections of depending highly on early detection by clinical examination and radiological image (chest x-ray) [1, 2]. Determining the respiratory infection in the x-ray image is based on consolidation (abnormal white color area) inside the thoracic cage. Moreover, extraction interest features of characteristics of the thoracic cage might be accompanied by many misleading features [3]. Therefore, the proper feature selection for each category in a dataset would usually increase the opportunity to make the right decisions of the machine learning algorithms. Generally, feature selection enhances the performance of the machine learning algorithm in three aspects: reduces overfitting, improves accuracy, and reduces processing time [4]. Reducing overfitting works on scaling down the redundancies in the dataset because the machine learning algorithm has less opportunity to make the right decisions with highly noisy data. The misleading features in data perform passively on accurate prediction of data mining models. The wrapper model is a powerful feature selection technique that selects features based on testing different groups of subset features [5]. A metaheuristic technique is a popular methodology that could feed the wrapper model with an optimal subset of features. It is mainly based on randomness for searching for the

optimal solution to a given problem. Metaheuristic has been recognized as fast, flexible, easy to implement, and successful in optimizing different fields [6-9]. Particle swarm optimization (PSO) is an efficient metaheuristic algorithm that has been applied successfully to optimize the performance of machine learning algorithms [5, 10]. PSO is highly featured for simplicity, ease of development and potentials in selecting optimum features [11].

COVID-19 is a world epidemic disease of respiratory infections due to the wide propagation among people. X-ray examination is widely used in medical test to determine respiratory infections [2, 3, 12]. Designing an application that identifies cases infected with respiratory infections based only on X-ray would significantly reduce the momentum on hospitals and helps patients to know their health state remotely.

With the aim to increase the machine precision in classifying chest X-ray image into positive and negative would require both features extracted from images and select high effective features. Wrapper model is a popular feature selection technique that selects features based on testing several groups of features and pick the best one. Metaheuristic algorithms can potentially enhance wrapper model performance for selecting an optimal group of features efficiently [11]. Choosing the same features would be inappropriate for all categories. Applying Wrapper model on data as one block often suffers from several drawbacks reflected on their technical efficiency:

a) High time complexity: Applying the wrapper model on data as one block significantly have high time complexity that potentially effects on classification tile of machine learning.
b) Inefficient prediction: The generalization in features selection often results in an inefficient classification of machine learning algorithms. Moreover, selecting features without having a high relative with dataset classes would negatively affect the system performance
c) Inaccurate-inefficient features selection: The wrapper model essentially starts with an initial group of features that are selected randomly, this would usually result in inaccurate and inefficient classification for query messages. For example, a dataset has two class A and B, selected effective features from class A would not necessarily be effective for class B. Therefore, applying feature selection to all classes might result in features that are not effective for all classes.

In this paper, a novel multi-feature selection empowered with initial clustering (MFSC) is proposed which potentially enhances the Wrapper model for selecting optimal features. We have clustered the dataset into two groups by K-mean and applied the wrapper model individually on each cluster. The proposed method has resulted in several promising achievements:

a) Efficient prediction: Most features selection models are introduced to find optimal features for all dataset categories that are applied based on the same analysis process overall categories. The proposed technique is capable to select features for each group of data that have high similarity by segmenting data (clustering) into two independent clusters based on the similarity of their characteristics. Wrapper model is applied to each cluster individually. Experimental results show significant enhancing of machine learning predictions when using them with the proposed technique as feature selection comparison with traditional feature selections.
b) Low time complexity: Proposed MFSC model enhances the required time of machine learning algorithms to find a class of query data compared with other models. Technically, technically, selecting optimal feature for each part of data individually is better than selecting features same for the whole dataset.
c) Accurate-efficient features selection: The proposed model trains wrapper model to select typical suitable features for each class in the dataset. The data are broken up into two clusters which have high similarity data based on clusters in the dataset. The proposed features selection has tremendously improved the performance of the machine learning algorithms.

With the aim to develop and optimize features selection techniques, several studies have proposed the different features of selection models. Most of these models are based on selecting the same features overall used data. This will negatively affect the performance of machine learning algorithms

Yang *et.al.* [12] have proposed a new model to classify x-ray images of oesophagal cancer. They extracted features in four-level discrete wavelet decomposition (DWT). The proposed system selects optimal features based on two methods sequential forward selection (SFC) and principal component analysis (PCA). The authors have applied support vector machine (SVM) and K nearest neighbors (KNN) as Classifiers and performance evaluation. They did not take into account the required time, as the SFC algorithm updates feature groups iteratively by adding one and deleting the other in succession, which increases the processing time. Li *et. al.* [13] the authors have used both particle swarm optimization (PSO) and hybrid self-adaptive bat-inspired (HSBAT) to select optimal features of x-ray images. The proposed method selects an optimal feature of a high-dimensional dataset for enhancing the performance of classification models that achieved high accuracy and reliability. The authors optimize several feature selections without the model that feeding features to machine learning algorithms. Asuntha *et. al.* [14] have proposed a model using PSO to select

optimal features from the x-ray image and applying the SVM algorithm as a classifier. They detect lung cancer in x-ray image results of feature extraction and feature selection after segmentation. Zhu *et al.* [15] have employed the SVM to make a distinction within a class of Src kinase inhibitors. The sequential forward selection and sequential backward selection methods were used to remove redundant variables. The results showed that the proposed method could be employed to structure-activity relationship modelling with much-improved quality and pre-disability.

## 2. RESEARCH METHOD

This section discusses the most important mechanism used in the proposed system. It includes X-ray images collection and preprocessing, feature extraction, features selection, machine learning techniques, and the last part illustrates the proposed multi-feature selection empowered with initial clustering (MFSC).

### 2.1. Dataset collection and integration

In this paper, we have been collecting 292 samples of the chest x-ray images, it available in [16]. They were normalized with size 700×800 pixels and adjust the intensity by histogram equalization. Every image is checked by a respiratory doctors in the Ad Diwaniya consulate education hospital in Iraq to diagnose positive (has pneumonia disease) or negative (normal). Table1 shows the description of the chest x-ray dataset.

Table 1. Chest x-ray dataset details

| Class | Type | Number of samples | Total |
|---|---|---|---|
| Negative | Normal | 126 | 125 |
| Positive | COVID | 125 | 166 |
| | E-Coli | 4 | |
| | ARDS | 4 | |
| | Chlamydia | 3 | |
| | SARS | 11 | |
| | Streptococcus | 17 | |

### 2.2. Image preprocessing

Technically, image preprocessing enhances images in two aspects: image size and pixel components [13]. Furthermore, it has increased the performance of the machine learning algorithm. Appling feature extracting algorithms on images with equal size would approximately produce the same number of features. Generally, images are scanned in different illumination and though they may reflect different contrast values. Therefore, extracted features would be imprecision to represent the same features in the same image category. The image preprocessing phase reduces differentials among images that are in the same class [14]. Nearest-neighbor interpolation algorithm is used to normalize the size of the image to be 700x800 pixels. Nearest neighbor interpolation is the image resize approach by replication. It technically resizes images based on interpolation and resampling. In (1) generates image $g(m,n)$ from image $f(m,n)$ by factor $c$ in direction $m$ and $d$ in direction $n$ [17].

$$g(m,n) = \begin{cases} \frac{1}{cd}, & -\frac{c}{2} \le m < \frac{c}{2}, \ -\frac{d}{2} \le n < \frac{d}{2} \\ 0 & otherwise \end{cases} \quad (1)$$

In order to enhance x-ray images before using them in the proposed model, the Nearest neighbor interpolation is used to resize examining images (unify the size of the image).

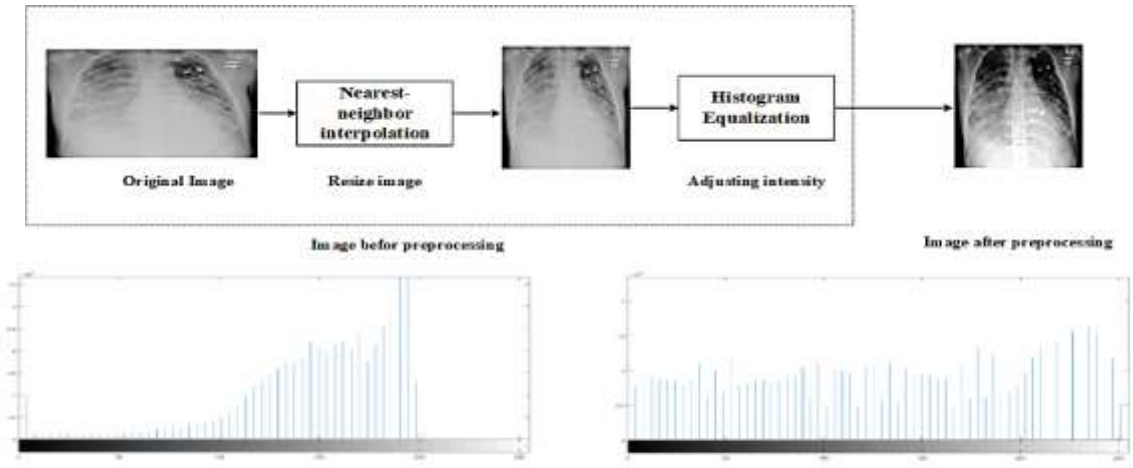
Histogram equalization algorithm (HEQ) is applied to diminishing the effects of over-brightness and over darkness in an image. With the aim for contrast adjustment color for the image has $m \times n$ pixels by HEQ, (2) calculates new value of pixels that have intensity $v$.

$$h(v) = round\left(\frac{cdf(v) - cdh_{min}}{(m \times n) - cdh_{min}} \times (L-1)\right) \quad (2)$$

where: $cdf$ is cumulative distribution function is calculated in (3)

$$cdf(v) = \sum_{i=1}^{k} \Pr(V = v_i) \quad (3)$$

where: $v_i$ is the largest possible value of $V$ that is less than or equal to $v$. Images after preprocessing are again been observed by a respiratory doctor to ensure that their quality is useful for diagnosis. Figure 1 illustrates the preprocessing phase in the proposed MFSC system.




(a) Histogram of the original image (b) Histogram of the processed image

Figure 1. Preprocessing image

### 2.3. Features extraction

Features extraction in fact is only extracting interest features from data meant to describe the most important information in data. Technically, raw data is converted into a features vector to be fed information about data into a machine learning algorithm [11]. The classification of chest x-ray images is basically based on color blocks within the thoracic cage [1, 2]. Three features extraction methods are applied to extract interest features: histogram of oriented gradients (HOG), principal component analysis (PCA), and principal component analysis of HOG (PCA-HOG).

#### 2.3.1. Histogram of oriented gradients (HOG)

HOG is an efficient image descriptor that converts a 2D image into a vector of features [12, 17, 18]. Technically, it extracts features from an image based on the oriented gradient of colors in localized portions of an image. HOG divides the image into small regions (windows). The dimensions of windows determine the number of features extracted from each cell in windows [8]. The features of $cell_{x,y}$ are determined based on gradient magnitude $m_{x,y}$ and orientation $\theta_{x,y}$. In (4) and (5) calculate the magnitude and orientation of $cell_{x,y}$ using $x_{directional}$ and $y_{directional}$ gradients $dx_{x,y}$ and $dy_{x,y}$ [9].

$$m_{x,y} = \sqrt[2]{dx(x,y)^2 - dy(x,y)^2} \tag{4}$$

$$\theta_{x,y} = \begin{cases} tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) - \pi & if\ dx(x,y)\ and\ dy(x,y) > 0 \\ tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) + \pi & if\ dx(x,y)\ and\ dy(x,y) < 0 \\ tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) & otherwise \end{cases} \tag{5}$$

#### 2.3.2. Principal component analysis (PCA)

Principle component analysis (PCA) is an orthogonal linear transformation technique that transfers raw data into a new form of equal or fewer dimensions of original data [19]. Furthermore, It reduces the required computations for features extraction method in data science [20]. Technically, PCA calculates the eigenvectors of a covariance matrix where the highest eigenvalues represent the significant features.

#### 2.3.3. Principal component analysis of HOG (PCA-HOG)

Features are extracted by HOG based on magnitude and orientation color in windows individually without count the interest window or unquiet characteristics in the image. Meanwhile, the obtained features

by this descriptor are over ten thousand features. Therefore, the resultant features vector may have several redundancies and garbage features [9]. PCA is a successful technique in reducing and omitting redundancy in data [4, 8]. PCA and HOG are applied sequentially to extract high significant features of x-ray image with low redundancies.

### 2.4. Machine learning algorithms

#### 2.4.1. K-mean

k-means is a clustering model partitioning $n$ data in $k$ part (clusters). Technically, it partitions data mainly based on two steps [21]: first, it finds centroids ($m$) that equal number of clusters and second aggregates data points ($x$) with closest centroid. In (6) finds nearest centroid for data point $\boldsymbol{x}$ K

$$S_i^{(t)} = \{x_p : \| x_p - m_i^{(t)} \|^2 \leq \| x_p - m_j^{(t)} \|^2 \ \forall j, 1 \leq j \leq k\} \quad (6)$$

where clusters are represented as sets $S_1, S_2, \ldots, S_i$.

Data is groped with a close centroid based on the nearest distance. In (7) calculates Euclidean distance $\boldsymbol{d}$ between two object $\boldsymbol{x}$ and $\boldsymbol{y}$ that have $\boldsymbol{n}$ diminutions.

$$d_{x,y} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (7)$$

In the initial step of k-mean, the centroid point is selected randomly from the dataset. Next, the centroid point of each cluster is calculated as the mean of cluster points (8)

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \quad (8)$$

#### 2.4.2. K nearest neighbor (KNN)

K nearest neighbors (KNN) is a non-parametric technique used in supervised machine learning to classify new object based on the high-density class of the nearest available cases. Generally, the distance between objects is calculated by one of minkowski distances (manhattan, euclidean distance, and distance), in this work the euclidean distance (7) has been applied to find the distance.

#### 2.4.3. Support vector machine (SVM)

A support vector machine (SVM) is a discriminative supervised machine learning introduced by [22] as both regression and classifier model. Technically, it classifies a new object based on hyperplane and support vectors. The hyperplane is multiple lines detected boundaries of classes that help to determine class data objects easily. SVM model sets the diminution of hyperplane based on present features in the dataset [18]. In (8) calculates the hyperplane.

$$w.x + b = 0 \quad (8)$$

where: $x$ is input, $w$ is weights vector, and $b$ is bias. The dataset that has few numbers features often is linearly separable, therefore, The SVM uses (9) to classify a new object $x$ [23].

$$f(x) = sign(w.x + b) \quad (9)$$

The high dimensional data potentially is not in every case linearly separable. As a result, the nonlinear decision function (10) is used to classify a new object $x$.

$$f(x) = \text{sign}(\sum_{i=1}^{N} a_i y_i K(x_i \cdot x) + b) \quad (10)$$

where: $K(x_i \cdot x)$ is the Kernel function, $y_i$ class data.

#### 2.4.4. Decision tree (DT)

A decision tree (DT) is a non-parametric supervised machine learning model that predicts the class of query data based on a sequences series of decision rules [23]. The root of each decision rule is a feature of data has the highest information gained than others. In (11) calculates the information gain of feature $A$ for dataset $D$.

$$Gain(A) = \sum_{i=1}^{n} A_{p_i} \, log_2(A_{p_i}) - \sum_{i=1}^{m} d_{p_i} \, log_2 \, (d_{p_i}) \quad (11)$$

where $d_{p_i}$ is the probability of each class in the dataset, $A_{p_i}$ probability of each distinct entity in a feature.

#### 2.4.5. Naïve bayes (NB)

Naive Bayes is a probabilistic supervised machine learning model that predicts the category of query data based on core concepts Bayes' theorem. Technically, it calculates the probability $P(y/x)$ of query data $x$ with all training classes $y$. The high $P(y/x)$ is determined by the new class of $x$. In (12) calculates the class probability.

$$P(y|x_1, \dots, x_n) = \frac{P(x_1|y)P(x_2|y) \dots P(x_n|y)P(y)}{P(x_1)P(x_2) \dots P(x_n)} \quad (12)$$

$P(x_1|y)$ is calculated by Gaussian Naive Bayes (13)

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}) \quad (13)$$

where $\mu_y$ , $\sigma_y$ is mean and standard deviation of train class $y$

#### 2.4.6. Random forest (RF)

Random forest is a supervised learning algorithm that is used for both classifications as well as regression. However, it is mainly used for classification problems. Technically, it builds its model based on four major steps: select random sample from data, construct a decision tree for every sample, voting will be performed for every predicted result, at last select the most voted prediction result as the final prediction result.

### 2.5. Feature selection

The reduction of data dimensions and important features selection is a necessity to enhance the performance of a machine learning algorithm. Several techniques have been applied to select the best group of features that have a high relative with a data objective. For example, the filter method, wrapper model, and embedded method [11]. In this paper, the Wrapper model is used for feature selection. Wrapper model is a feature selection method testing different groups of features and selecting a group that satisfying the best result [21]. It has applied features selection based on randomness. Metaheuristic techniques have succeeded in applying randomness to search for an optimal solution [24]. Therefore, it has significantly optimized features selection that feeds into the wrapper model [5]. Figure 2 illustrates the principles of the wrapper model with a metaheuristic for selecting the optimum feature.

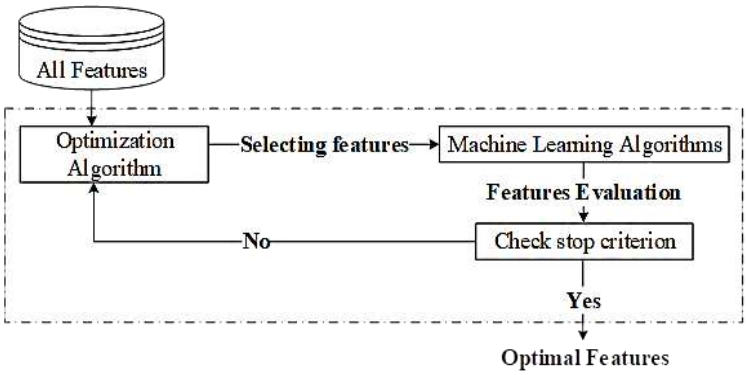



Figure 2. Features selection based on wrapper model and metaheuristic

Generally, Wrapper model is based on binary feature selection - only features that corresponding 1 in the optimization algorithm vector are selected-which restricts the metaheuristic algorithms to search in

limited boundaries either 0 or 1 [21]. The extending search space has enhanced the performance of metaheuristic algorithms. In the proposed system, the space searching interval range of -5, +5x is applied and the same range for both PSO and BA. The threshold for feature selection is determined by the user for each experiment [11, 25].

### 2.6. Particle swarm optimization (PSO)

Particle swarm optimization (PSO) is an evolutionary search algorithm (EA) inspired by the social behavior of animals that lives within groups [26]. PSO has been favored over other EA because it has a simple model with few parameters. The PSO is different from the other EA by the structure of the population (particles) where each particle of PSO has two positions: current position and best position finds by itself($p_{best}$) [27]. It picks up best $p_{best}$g among particles to be global best position ($g_{best}$ ). Technically, PSO updates particles position by adding the current position and new velocity of particle as shown in (13). PSO has in calculating particles velocity both exploration and exploitation of search processes [24]. It modifies particles velocity dynamically during the search process. The factors that govern particles movements are inertia component, the cognitive component, and the social component [28]. In (14) calculates the new velocity of particles.

$$v_i' = \underbrace{wv_i}_{\text{inertia component}} + \underbrace{c_1 r_1 (p_{best_i} - x_i)}_{\text{cognitive component}} + \underbrace{c_2 r_2 (g_{best} - x_i)}_{\text{social component}} \quad (14)$$

where: $w$ is weight inertia, $c_1, c_2$ are coefficient constant, $r_1, r_2$ random value within interval -1,+ 1x. In (15) calculates the new position of particles.

$$x_i' = x_i + v_i' \quad (15)$$

where: $x_i$ current particle position. $v_i'$ , $x_i'$ new particle velocity and position.

### 2.7. Proposed multi-feature selection empowered with initial clustering (MFSC)

The generalization in feature selection is often inappropriate to all categories of dataset though features of importance for one class in dataset are not necessarily important for others. Consequently, applying feature selection on each category of the dataset would provide best feature groups for each one. Technically, clustering is a suitable technique to partition a dataset into more than one part and group mostly identical in terms of characteristics. K-mean algorithm is applied in the proposed system to cluster the dataset. The proposed system mainly consists of three-phase as shown in Figure 3 preprocessing, training, and diagnosis (test).

#### 2.7.1. Preprocessing

The proposed system normalizes images to be in unifying size 800x700 pixels. The Histogram equalization algorithm (HEQ) is applied to the contrast adjustment color for mages. The proposed MFSC extract features from x-ray images based on HOG32x32 followed by PCA to extract 20 components.

#### 2.7.2. Training phase

The proposed system builds the training model by implementing the following series of procedures: Initially, the proposed system has applied the K-mean algorithm to partition X-ray COVID19 dataset into two clusters $C_1$ and $C_2$. The original version of the dataset has two categories positive and negative therefore it is divided into two clusters. The proposed system finds a rank for each cluster ($CR_i$ ) with keeping clustered data ($CD_i$) together. Cluster rank is the centroid point of each cluster. Wrapper model with PSO is applied to each cluster to select optimal features ($CF_i$).

#### 2.7.3. Test phase

In the diagnosis phases, the system finds nearest cluster rank to the query image. The Optimal feature of the close cluster is applied on the query image to select optimal features. Finally, machine learning is applied to predict the type of query image. Figure 3 illustrates the main steps of proposed MFSC system.

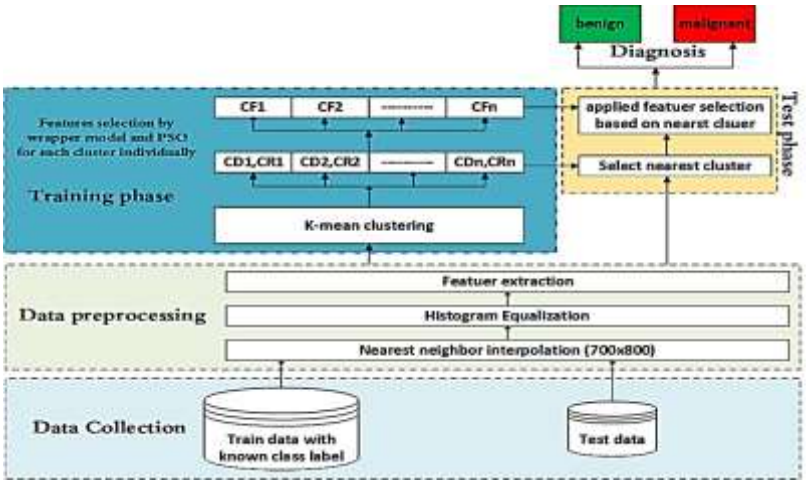



Figure 3. Proposed MFSC system

## 3. RESULTS AND DISCUSSION

This section discusses the experimental results of the proposed system and at the same time is given the comprehensive with other models. The evaluation methodology is based on four efficiency metrics: accuracy, precision, recall, F1_score. These metrics determine the correct rate of data identified for the same class from overall data retrieval. In (14-17) show computations of the evaluation metrics accuracy, precision, Recall, F1_score respectively. In addition, the processing time is calculated per case to investigate system performance.

$$Accuracy = \frac{true\ posiative + true\ nagative}{true\ posiative + true\ nagative + false\ posiative + false\ nagative} \quad (14)$$

$$precision = \frac{true\ posiative}{true\ posiative + true\ nagative} \quad (15)$$

$$Recall = \frac{true\ posiative}{true\ posiative + fals\ nagative} \quad (16)$$

$$F1\_score = 2 * \frac{precision * Recall}{precision + Recall} \quad (17)$$

In principle, the main strategy of the achieved evaluation is to find out how accurate is the proposed system to detect respiratory infection from x-ray images. The experimental evaluation has considered 292 x-ray respiratory cases in [16]. Features are extracted in three different methods, PCA, $HOG_{16x16}$ the size of sliding windows 16x16 pixel, and $HOG_{32x32}$. Features are reduced using PCA. Optimal components extracted by PCA are 20. Four popular prediction algorithms: KNN, SVM, NB, and DT are applied on all cases of features extraction and reduction. Five evaluation criterions: Precision, Recall, f1_scor, Accuracy, and processing time are calculated for comparing the proposed model efficiency with other standard methods. PCA has achieved less accuracy compering with other descriptors (PCA, $HOG_{16x16}$, and $HOG_{32x32}$) when applied for machine learning algorithms. Medically, determining the respiratory infection in x-ray images is based on white color blocks inside the thoracic cage [24]. therefore, localization and color orientation is needed in diagnosis. PCA extracts features by only value localization though it would not be an optimal descriptor for x-ray respiratory image. Technically, HOG extracts feature based on localization and color orientation. As a result, it provides high relative features for describing x-ray respiratory images. In addition, the windows size 32x32 is better than windows size 16x16 for HOG descriptor.

Table 2 shows the effect of the proposed system on the performance of KNN. Obviously, the proposed MFSC enhances the performance of KNN with all descriptors. The best accuracy is achieved by KNN as shown in Table 3 with a descriptor ($HOG_{32x32}$+PCA) reaches to 93.06%. Furthermore, the proposed model reduces the processing time of KNN approximately 60.7%. Table 2 illustrates the proposed system effects on the prediction capabilities of SVM, where it has an accurate prediction reach to 97.13%. Moreover, the required time of the prediction process is reduced by 62.5%. The best results achieved by SVM was with descriptor $HOG_{32x32}$+PCA. NB classifier has peak performance when features are extracted by HOG with

windows size of 32. It has an accuracy of 90.72 with both full features and features selected in the traditional wrapper model while the accuracy reaches to 95.45% with the proposed features selection. Table 2 and 3 show the performance of NB with full features, traditional features selection (FS), and MFSC. RF classifier has high performance when applying HOG with windows size of 32 with PCA. It has an accuracy of 93.02 with both full features. Table 2 shows the performance of RF with full features, traditional features selection (FS), and MFSC. The proposed model has shown a clear improvement in the performance of the DT algorithm in terms of accuracy and processing time. Table 2 and 3 show the comparison of the performance of DT algorithm with five descriptors. DT algorithm has shown the lowest results when compared to SVM, DT, KNN, and RF algorithm.

Table 2. The performance of the machine learning algorithms over three features extraction (PCA, HOG16x16, and HOG32x32) and selection (full, FS, MFSC)

| ML | FS method | PCA | | | | | $HOG_{16x16}$ | | | | | $HOG_{32x23}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | f1scor | Accuracy | time | Precision | Recall | f1scor | Accuracy | time | Precision | Recall | f1scor | Accuracy | time |
| KNN | full | 74.29 | 78.79 | 76.47 | 83.51 | 0.002 | 72.09 | 93.94 | 81.58 | 85.57 | 0.17 | 67.39 | 93.94 | 78.48 | 82.47 | 0.05 |
| | FS | 80 | 84.85 | 82.35 | 87.63 | 0.002 | 75.61 | 93.94 | 83.78 | 87.63 | 0.15 | 73.81 | 93.94 | 82.67 | 86.6 | 0.05 |
| | MFSC | **85.45** | **90.04** | **87.69** | **91.09** | **0.002** | **91.18** | **90** | **89.61** | **91.06** | **0.11** | **89.47** | **100** | **94.44** | **90.91** | **0.04** |
| SVM | full | 83.33 | 75.76 | 79.37 | 86.6 | 0.002 | 85.29 | 87.88 | 86.57 | 90.72 | 0.16 | 87.88 | 87.88 | 87.88 | 91.75 | 0.05 |
| | FS | 83.33 | 75.76 | 79.37 | 86.6 | 0.002 | 85.29 | 87.88 | 86.57 | 90.72 | 0.15 | 96.67 | 87.88 | 92.06 | 94.85 | 0.05 |
| | MFSC | **87.06** | **80.52** | **83.28** | **88.26** | **0.002** | **91.18** | **90** | **89.61** | **91.06** | **0.09** | **100** | **80** | **88.89** | **95.45** | **0.03** |
| NB | full | 76.67 | 69.7 | 73.02 | 82.47 | 0.002 | 76.47 | 78.79 | 77.61 | 84.54 | 0.14 | 90 | 81.82 | 85.71 | 90.72 | 0.05 |
| | FS | 79.31 | 69.7 | 74.19 | 83.51 | 0.002 | 78.79 | 78.79 | 78.79 | 85.57 | 0.14 | 90 | 81.82 | 85.71 | 90.72 | 0.05 |
| | MFSC | **82.11** | **80.52** | **81.18** | **86.41** | **0.001** | **91.67** | **84.64** | **87.55** | **88.84** | **0.1** | **100** | **94.12** | **96.97** | **95.45** | **0.03** |
| RF | full | 81.62 | 80.74 | 79.99 | 83.50 | 0.002 | 79.84 | 77.86 | 78.92 | 81.44 | 0.13 | 74.58 | 81.39 | 76.45 | 80.14 | 0.05 |
| | FS | 82.47 | 81.62 | 81.08 | 84.92 | 0.002 | 73.17 | 85.90 | 81.01 | 84.50 | 0.13 | 84.94 | 84.44 | 85.19 | 86.59 | 0.03 |
| | MFSC | 82.47 | 81.62 | 81.08 | 84.92 | 0.001 | 80.03 | **88.23** | **82.44** | **87.01** | **0.09** | **90.00** | **81.18** | **85.17** | **88.72** | **0.03** |
| DT | full | 56.1 | 69.7 | 62.16 | 71.13 | 0.002 | 60 | 54.55 | 57.14 | 72.16 | 0.14 | 62.5 | 75.76 | 68.49 | 76.29 | 0.05 |
| | FS | 57.5 | 69.7 | 63.01 | 72.16 | 0.002 | 72.97 | 81.82 | 77.14 | 83.51 | 0.14 | 69.05 | 87.88 | 77.33 | 82.47 | 0.05 |
| | MFSC | **69.33** | **80.36** | **73.89** | **81.65** | **0.001** | **88.89** | **85.71** | **87.27** | **84.44** | **0.09** | **81.27** | **96.97** | **88.4** | **86.27** | **0.03** |

Table 3. The performance of the machine learning algorithms over two features extraction (HOG16x16+PCA and HOG32x32 + PCA) and selection (full, FS, MFSC)

| ML | FS method | $HOG_{16x16}$ +PCA | | | | | $HOG_{32x32}$ +PCA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | f1scor | Accuracy | time | Precision | Recall | f1scor | Accuracy | time |
| KNN | full | 82.86 | 87.88 | 85.29 | 89.69 | 0.13 | 78.95 | 90.91 | 84.51 | 88.66 | 0.03 |
| | FS | 82.86 | 87.88 | 85.29 | 89.69 | 0.13 | 85.29 | 87.88 | 86.57 | 90.72 | 0.03 |
| | MFSC | **90.91** | **86.96** | **88.89** | **90.74** | **0.11** | **93.1** | **84.64** | **88.3** | **93.06** | **0.03** |
| SVM | full | 93.33 | 84.85 | 88.89 | 92.78 | 0.13 | 96.55 | 90.32 | 84.85 | 93.81 | 0.03 |
| | FS | 96.43 | 81.82 | 88.52 | 92.78 | 0.13 | 87.88 | 92.06 | 92.06 | 94.85 | 0.03 |
| | MFSC | **95.65** | **85.65** | **90.1** | **94.02** | **0.08** | **95.44** | **100** | **97.86** | **97.13** | **0.02** |
| NB | full | 83.33 | 75.76 | 79.37 | 86.6 | 0.13 | 77.14 | 81.82 | 79.41 | 85.57 | 0.03 |
| | FS | 83.87 | 78.79 | 81.25 | 87.63 | 0.13 | 80 | 84.85 | 82.35 | 87.63 | 0.03 |
| | MFSC | **90.89** | **86.96** | **88.83** | **90.74** | **0.08** | **100** | **56.25** | **72** | **90.67** | **0.02** |
| RF | full | 89.23 | 79.18 | 85.09 | 88.75 | 0.12 | 87.19 | 82.50 | 84.78 | 85.95 | 0.03 |
| | FS | 91.01 | 87.87 | 89.23 | 90.78 | 0.07 | 90.23 | 84.89 | 87.50 | 91.75 | 0.03 |
| | MFSC | **90.62** | **88.44** | **89.23** | **91.99** | **0.07** | **91.15** | **86.03** | **89.72** | **93.02** | **0.02** |
| DT | full | 75.61 | 93.94 | 83.78 | 87.63 | 0.13 | 67.5 | 81.82 | 73.97 | 80.41 | 0.03 |
| | FS | 75 | 81.82 | 78.26 | 84.54 | 0.13 | 81.58 | 93.94 | 87.32 | 90.72 | 0.03 |
| | MFSC | **86.36** | **82.61** | **84.44** | **87.04** | **0.08** | **90.79** | **86.97** | **87.32** | **93.09** | **0.02** |

Generally, the overall results have proven an interesting success and potential for the proposed technique as it outperformed other models in all the testing cases reflecting a significant achievement. In other words, the proposed MFSC features selection has consistently required fewer features to deliver accurate classification results in high-dimensional datasets and demonstrated its value as a potential alternative to traditional features selection.

## 4. CONCLUSION

Coronavirus (COVID-19) is diagnosed medically as the second most common cause of a cold. The chest x-ray is first required to determine the negative or positive cases of the severe acute respiratory syndrome such as COVID-19. Medically, respiratory infection white clusters occur within the thoracic cage. HOG is proposed to extract features from images based on color gradient orientation without reducing redundancy in extracted features. Consequently, this negatively effects on the accuracy of machine learning algorithms and the fairness of features selected by Wrapper model. Moreover, PCA can extract features from images based on color convergence without taking into account the color gradient that is significant for detecting respiratory infections in x-ray images. Therefore, both HOG and PCA for features extraction have significantly increased the accuracy of machine learning algorithms. In order to reduce the wide range of features and support features selection and diagnosis, a novel machine learning model is proposed by introducing clustering of data in advance. K-Means clustering is applied to cluster X-Ray images into two clusters. Moreover, Wrapper model is applied to detect Pneumonia cases from both clusters. The evaluation of the proposed system has shown significant performance and efficiency with potential accuracy.

## REFERENCES


[1] N. S. Lingayat and M. R. Tarambale, "A Computer Based Feature Extraction of Lung Nodule in Chest X-Ray Image," *Int. J. Biosci. Biochem. Bioinforma.*, vol. 3, no. 6, pp. 624-629, 2013, doi: 10.7763/ijbbb.2013.v3.289.

[2] F. R. Hirsch, W. A. Franklin, A. F. Gazdar, and P. A. Bunn, "Early detection of lung cancer: Clinical perspectives of recent advances in biology and radiology," *Clin. Cancer Res.*, vol. 7, no. 1, pp. 5-22, 2001.

[3] S. A. Patil and V. R. Udupi, "Chest X-ray features extraction for lung cancer classification," *J. Sci. Ind. Res. (India).*, vol. 69, no. 4, pp. 271-277, 2010.

[4] R. T. C and R. Sivramakrishnan, "Fuzzy Neuro-Genetic Approach for Feature Selection and Image Classification in Augmented Reality Systems," vol. 8, no. 3, pp. 194-204, 2019, doi: 10.11591/ijra.v8i3.pp194-204.

[5] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection - A comparative study," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4881 LNCS, pp. 178-187, 2007, doi: 10.1007/978-3-540-77226-2_19.

[6] F. Lü, C. Qin, and Yunpeng, "Particle swarm optimization-based BP neural network for UHV DC insulator pollution forecasting," *J. Eng. Sci. Technol. Rev.*, vol. 7, no. 1, pp. 132–136, 2014, doi: 10.25103/jestr.071.21.

[7] S. Ibrahim, N. A. Wahab, F. S. Ismail, and Y. M. Sam, "Optimization of artificial neural network topology for membrane bioreactor filtration using response surface methodology," *IAES Int. J. Artif. Intell.*, vol. 9, no. 1, pp. 117-125, 2020, doi: 10.11591/ijai.v9.i1.pp117-125.

[8] N. F. Fadzail, S. M. Zali, M. A. Khairudin, and N. H. Hanafi, "Stator winding fault detection of induction generator based wind turbine using ANN," vol. 19, no. 1, pp. 126-133, 2020, doi: 10.11591/ijeecs.v19.i1.pp126-133.

[9] N. Z. Mohd Ali, I. Musirin, and H. Mohamad, "Clonal evolutionary particle swarm optimization for congestion management and compensation scheme in power system," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 2, pp. 591-598, 2019, doi: 10.11591/ijeecs.v16.i2.pp591-598.

[10] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognit. Lett.*, vol. 28, no. 4, pp. 459-471, 2007, doi: 10.1016/j.patrec.2006.09.003.

[11] A. H. Jabor and A. H. Ali, "Dual Heuristic Feature Selection Based on Genetic Algorithm and Binary Particle Swarm Optimization," *J. Univ. BABYLON pure Appl. Sci.*, vol. 27, no. 1, pp. 171-183, 2019, doi: 10.29196/jubpas.v27i1.2106.

[12] F. Yang *et al.*, "Feature extraction and classification on esophageal x-ray images of xinjiang kazak nationality," *J. Healthc. Eng.*, vol. 2017, 2017, doi: 10.1155/2017/4620732.

[13] J. Li, S. Fong, L. Liu, N. Dey, and A. S. Ashour, "Dual feature selection and rebalancing strategy using metaheuristic optimization algorithms in X-ray image datasets," vol. 78, pp. 20913-20933, 2019, doi: 10.1007/s11042-019-7354-5.

[14] J. Too, A. R. Abdullah, N. M. Saad, N. M. Ali, and T. N. S. Tengku Zawawi, "Featureless EMG pattern recognition based on convolutional neural network," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1291-1297, 2019, doi: 10.11591/ijeecs.v14.i3.pp1291-1297.

[15] J. Zhu, W. Lu, L. Liu, T. Gu, and B. Niu, "Classification of Src Kinase Inhibitors Based on Support Vector Machine," pp. 719-727, 2009, doi: 10.1002/qsar.200860105.

[16] Pneumonia-SARS-COVID19-in-X-Ray-Images-Datasets-2020. Available https://github.com/I-SOFT-developer/Pneumonia-SARS-COVID19-in-X-Ray-Images-

[17] A. Hidaka and T. Kurita, "Selection of Histograms of Oriented Gradients Features for Pedestrian Detection Selection of Histograms of Oriented Gradients," no. May 2014, 2007, doi: 10.1007/978-3-540-69162-4.

[18] P. Carcagnì, M. Del Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients : a comprehensive study," *Springerplus*, no. November, 2015, doi: 10.1186/s40064-015-1427-3.

[19] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016, doi: 10.1098/rsta.2015.0202.

[20] A. Sophian, G. Y. Tian, D. Taylor, and J. Rudlin, "A feature extraction technique based on principal component analysis for pulsed Eddy current NDT," *NDT E Int.*, vol. 36, no. 1, pp. 37-41, 2003, doi: 10.1016/S0963-

8695(02)00069-5.
[21] R. A, S. P. S, K. V Rangarao, and A. Saranya, "Efficient datamining model for prediction of chronic kidney disease using wrapper methods," *Int. J. Informatics Commun. Technol.*, vol. 8, no. 2, p. 63, 2019, doi: 10.11591/ijict.v8i2.pp63-70.
[22] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16-28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.
[23] Y. Zhang, S. Wang, P. Phillips, and G. Ji, "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection," *Knowledge-Based Syst.*, vol. 64, pp. 22-31, 2014, doi: 10.1016/j.knosys.2014.03.015.
[24] A. Mansoor *et al.*, "Segmentation and image analysis of abnormal lungs at CT: Current approaches, challenges, and future trends," *Radiographics*, vol. 35, no. 4, pp. 1056-1076, 2015, doi: 10.1148/rg.2015140232.
[25] A. H. Al-saeedi, "Binary Mean-Variance Mapping Optimization Algorithm (BMVMO)," *J. Appl. Phys. Sci.*, vol. 2, no. 2, pp. 42-47, 2016, doi: 10.20474/japs-2.2.3.
[26] R. N. Khushaba, A. Al-Ani, and A. Al-Jumaily, "Differential Evolution based feature subset selection," *Proc. - Int. Conf. Pattern Recognit.*, 2008, doi: 10.1109/icpr.2008.4761255.
[27] A. H. Alsaeedi, A. H. Aljanabi, M. E. Manna, and A. L. Albukhnefis, "A proactive metaheuristic model for optimizing weights of artificial neural network," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 20, no. 2, pp. 976-984, 2020, doi: 10.11591/ijeecs.v20.i2.pp976-984.
[28] H. Dhrif, L. G. S. Giraldo, M. Kubat, and S. Wuchty, "A Stable Combinatorial Particle Swarm Optimization for Scalable Feature Selection in Gene Expression Data," pp. 1-13, 2019, [Online]. Available: http://arxiv.org/abs/1901.08619.

## BIOGRAPHIES OF AUTHORS

**Salam Saad Alkafagi** received the M.Sc. degree in Information technology from BAMU University, India, 2013. He is currently a working teacher in Al-Waeli High School for the distinguished, Babylon, Iraq. His research interests in optimization, cloud computing security, machine learning and deep learning.

**Ali Hakem Alsaeedi** is completed B.Sc. in Computer Sciences in 2006 from the college of sciences at the University of Al-Qadisiyah, Ad Diwaniya, Iraq. Received his M.Sc. (master) in computer sciences in the year 2016 from the college of computer sciences at the Yildiz Technical University (YTU), Istanbul, Turkey. He has worked as a lecturer at several Iraqi universities in the areas of Artificial Intelligent, Data mining, and signal processing. He currently works as a lecturer at the University of Al-Qadisiyah. His research interests machine learning, smart optimization algorithms, and optimization of Big Data.

**Dhiah Al-Shammary** received his M.Sc. (Masters) in Computer Science as the top student in his department for the year 2005 from the college of science at Al-Nahrain University, Baghdad, Iraq. In 2002.the PhD degree in computer science at RMIT University, Melbourne, Australia. Dhiah was awarded as the top student in computer science in Iraq after he participated in the annual scientific competition exam for the top bachelor students. He is currently an associate professor in the College of Computer Science and Information Technology, University of Al-Qadisiyah. He has worked as a lecturer at several Iraqi universities in the areas of software engineering, computer systems and machine language.

**Hassan Alsaeedi** is completed B.Sc. in Medicine Sciences in 2015 from the College of Medicine at the University of Al-Qadisiyah, Ad Diwaniya, Iraq. He has been worked as a doctor in several hospitals in Iraq.

**Hadeel Wajeeh Abid** is completed B.Sc. in Medicine Sciences in 2013 from the College of Medicine at the University of Al-Qadisiyah, Ad Diwaniya, Iraq. She has been worked as a doctor in several hospitals in Iraq.