Cluster Computing (2020) 23:1781–1794
https://doi.org/10.1007/s10586-020-03127-w

Data mining techniques for analyzing healthcare conditions of urban space-person lung using meta-heuristic optimized neural networks

Ahed Abugabah[1] · Ahmad Ali AlZubi[2] · Feras Al-Obeidat[1] · Abdulaziz Alarifi[2] · Ayed Alwadain[2]

Received: 13 March 2020 / Revised: 17 April 2020 / Accepted: 6 May 2020 / Published online: 25 May 2020

Abstract
Urban computing is one of the effective fields that have ability to collect the large volume of data, integrate and analyze the data in urban space. The urban space faces several issues such as traffic congestion, more energy consumption, air pollution and so on. Among the several problems, air pollution is one of the major issues because it creates several health issues. So, this paper introduces the meta-heuristic optimized neural network to analyze patient health to predict different diseases. Initially, patient data are collected, normalized by applying a min–max normalization process. Then different features are extracted and Hilbert–Schmidt Independence Criterion based features are selected. Further patient's health condition is analyzed and classified into a normal and abnormal person. The classification process is done by applying the harmony optimized modular neural network. Here the system efficiency is evaluated using simulation results, which ensures maximum accuracy of 98.9% -ELT-COPD and 98% -NIH clinical dataset.

Keywords Urban computing · Air pollution · Health issues · Meta-heuristic optimized neural networks · Hilbert–Schmidt independence criterion · Harmony optimized modular neural network

1 Introduction

Urban computing [1] is one of the interdisciplinary fields that provides enormous computing technologies to build an effective urban area. The urban computing process enhances the urban quality, computational power in the populated areas. The urban computing has ability to collect, integrate and examine the large volume of data [2] in urban space. Although the urban space faces several issues such as traffic congestion, more energy consumption, air pollution and so on. Among the several issues, air pollution [3] is one of the crucial issues because it affects people's health seriously. The air pollution caused by a mixture of man-made and natural substance in the breathing air. The urban space polluted due to the gases (radon, carbon monoxide), household chemicals, products, building materials, tobacco smoke, pollen, and mold. In addition to this, the outdoor pollutants [4] called fine particles, fossil burning, noxious gases, tobacco usage, and ground-level ozone are the main reasons for polluting the urban space. This air pollution creates health issues over the past 30 years. There are several health issues [5] such as wheezing, asthma, breathing, coughing, worsening, changes in lung function, adverse pregnancy outcomes, cardiovascular disease, and death also happened due to air pollution. According to the world health organization report in 2013, which concluded that outdoor air pollutants are mostly affecting human health. Among the several health issues, asthma, chronic obstructive pulmonary disease (COPD) and respiratory-related issues are the most common air pollution disease [6] in urban space. The urban space normally haves several traffic-related issues that created heavy air pollution. Due to the serious impact of urban space air pollution [7], people's lung health conditions are continuously analyzed to minimize the mortality rate.

The lung disease consists of around 200 chronic inflammations in lung tissues [8] that creates severe breathing problem. The lung disease has specific

✉ Ahed Abugabah
ahed.abugabah@zu.ac.ae

[1] College of Technological Innovation, Zayed University, Dubai, United Arab Emirates

[2] Computer Science Department, King Saud University, Riyadh, Saudi Arabia

characteristics that need to identify clearly for making effective clinical decisions. To predicting the lung disease [9], earlier detection system continuously captures the patient lung images as it creates ambiguity and 50% of complexity in radiological prediction. More ever, the radiological process consumes more time, requires huge manpower to examine the number of patients in urban space. To overcome these issues, the automation image processing [10] and classification techniques [11] are used in many medical applications. Further, an automatic recognition system is created by incorporating several pattern analysis and recognition techniques. The developed automatic system utilizes effective techniques to examine the large volume of lung images [12] and respective image features.

Nowadays artificial intelligence and deep learning techniques [13] are used to resolve pattern recognition [14] and computer vision problems effectively. The intelligent techniques can examine the captured images with robustness manner and also to ensure the maximum accuracy results [15]. In addition to intelligent learning techniques [16], image preprocessing, region segmentation, features extraction the classification techniques are also used to predict the changes in the normal patterns. Especially, the recent convolution neural network have several deep neural networks such as VGG [17], Alexnet [18] and google net [19] which is successfully analyzed and provides the deep learning to study image features. This learning process effectively gives the label to the images that helps to match the features while performing the testing process. Even though, the method consumes more time, computation complexity [20] while processing the massive amount of data. In addition to this, the system faces overfitting issues and data starvation problem while processing the lung data. These issues need to be resolved before performing the lung health condition identification else the system efficiency is less To ensure better performance of automatic lung disease identification process effectively, optimized and meta-heuristic approaches need to be introduced to analyzed further.

In this manuscript, the automatic lung health condition analysis system is developed by applying the meta-heuristic optimized neural networks. To maximize the system efficiency, the intelligent feature selection technique called Hilbert–Schmidt Independence Criterion based features selection [21] process is applied. The feature selection approach reduces the data overfitting problem that selects the optimized and most relevant features to the system. This feature selection process reduces the data issues and reduces time consumption for performing the classification process. In addition to this, the effective learning process reduces confusion while recognizing the lung health condition successfully. Based on the discussion, the general working process is depicted in Fig. 1. In our process, the large volume of information is processed by a meta-heuristic approach and excellent features are derived which determines the target lung patterns. During the analysis process, two different datasets such as Expression data in lung tissue from moderate COPD patients [22], healthy smokers and nonsmokers (ELT-COPD) and NIH Clinical dataset information [23] is used for evaluating the efficiency of the introduced system. Here the contribution of the work is mentioned as follows.

1. The Meta-heuristic optimized technique is introduced to analyze the urban space people's lung health condition identification process. Furthermore, the optimized feature selection technique is applied to reduce the data overfitting problem.
2. To our knowledge, meaningful and effective features are derived from the lung database information to investigating the people's lung health status in an accurate manner.
3. The effective learning process is applied to process people's health information and classifies their health condition accurately.
4. To demonstrate the effective optimized technique to provide better results while solving the discussed issues.

2 Related works

(Huang et al. 2020) [24] Created the deep convolution neural network architecture for analyzing and predicting the interstitial lung disease. During this process, radiological information are collected from patients who are processed with the help of two-stage transfer learning process. This convolution network extracts the various features from lung image and classifies the lung disease features effectively. The two-stage training process successfully processes both supervised and unsupervised manner which maximizes the lung disease pattern recognition efficiency. This introduced system ensures the feasibility as well as robustness while recognizing the lung disease patterns. (Selvanambi 2018) [25] Developed an effective lung cancer identification system with the help of higher-order recurrent neural network and glowworm swarm optimization algorithm. Initially, the lung images are gathered and processed with the help of optimized techniques to detect cancer in an earlier stage. This system analyzes the lung images to predict the multimodal disease which is done by the levenberg Marquardt model. The efficiency of the system is evaluated using the benchmark dataset and the system ensures 98% accuracy while classifying the lung tumor.

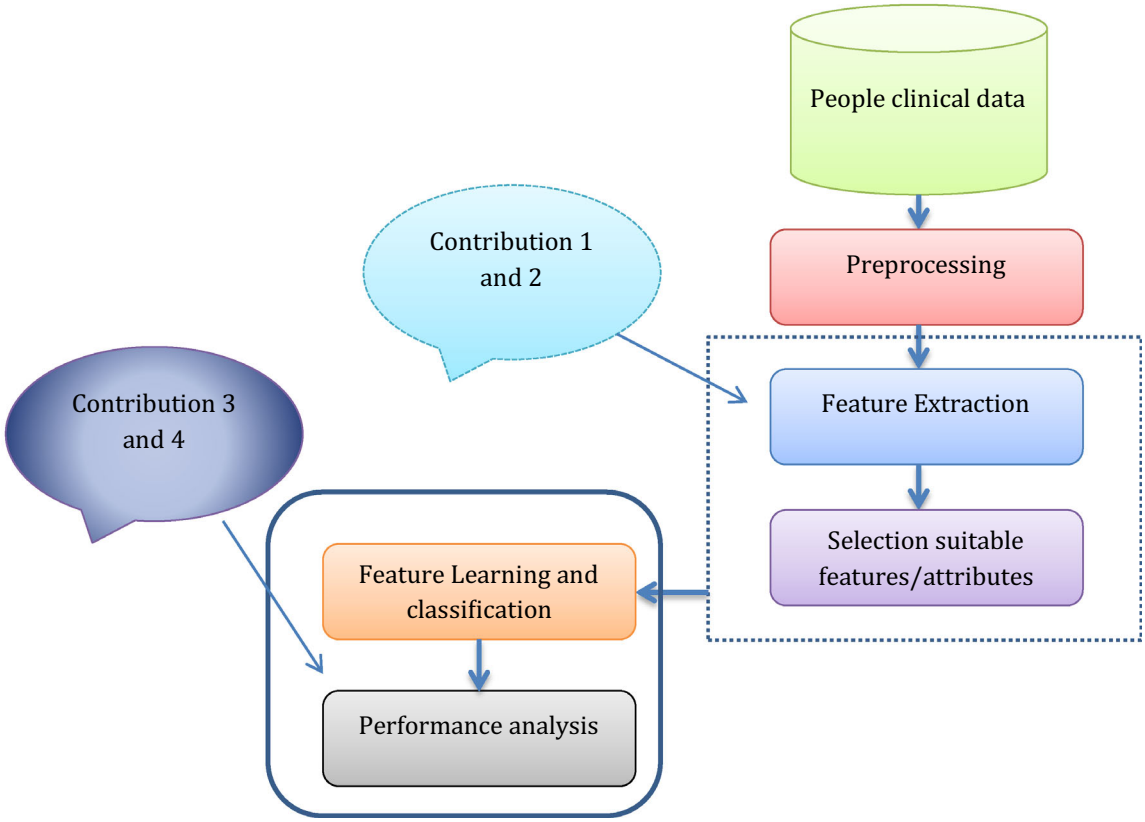

Fig. 1 Urban space-people lung health condition analysis process architecture

(ALzubi et al. 2019) [26] Maximizing the lung cancer prediction accuracy using ensemble weight-optimized neural network along with maximum likelihood boosting approach. Initially, the lung images are gathered and important features are extracted from the image. Afterward, optimized features are selected with the help of integrated newton Raphson maximum likelihood and minimum redundancy approach. Further classification process is performed using the ensemble and boosting approach. The introduced approach effectively recognizes the lung cancer with maximum accuracy and minimizes the recognition delay compared to conventional techniques. (Hankey et al. 2017) [27] Discussing the urban space air pollution condition and people's health status to improve the urban people's lifestyle. During the analysis process, the relationship between urban people's lifestyle, health status, and air pollution details are continuously monitored. In addition to this, their physical activities, noise level, and green space information also collected to get changes in their urban relationship. From the analysis, urban people transportation criteria, air quality improvement, and compact growth also examined to provide the proper guidelines to the people for improving the health condition.

(Zhang et al. 2015) [28] Discussing urban air pollution and the impact of the health condition in developing countries. The main intention of the paper is to examine outdoor air pollution because most of the developing countries generate pollution in huge amounts. This air pollution leads to create a different kind of diseases in developing countries. Based on the analysis, the respective recommendation is provided to minimize air pollution and reduces health disease effectively. (Jiao et al. 2018) [29] Analyzing the socio-economic impact in china due to urban air pollution and health impacts. Initially, air pollution and the respective impact of health effects are investigated because this affects the socioeconomic factors. The data is collected in China in 2014, which is examined using the multi-level mixed effect model. The model successfully predicts the air pollution level, ranked the pollutants according to the health status. Based on the rank level respective actions are carried out to minimize the problem in china's socio-economic factor. Depending on the various research analyses, urban space requires more attention to minimize air pollution because it creates several health diseases. From the analysis, lung disease is the most commonly affected one because the pollutants directly affect the respiratory system. So, this manuscript focuses on examining the lung health condition using the meta-heuristic optimization algorithm. The introduced algorithm helps to resolve the above Sect. 1discussed issues. Finally,

the efficiency of the system is evaluated using experimental results.

The rest of the manuscript is arranged as follows, Sect. 3 examines the meta-heuristic algorithm based on lung disease identification, Sect. 4 evaluate the efficiency of the lung disease identification process and conclusion is discussed in Sect. 5.

3 Urban space people lung health condition identification process

This section discusses the urban space of people's health conditions due to the air pollution problem. During the analysis, two different datasets such as Expression data in lung tissue from moderate COPD patients, healthy smokers and nonsmokers (ELT-COPD) and NIH Clinical dataset information are used.

3.1 Expression data in lung tissue from moderate COPD patients, healthy smokers, and nonsmokers (ELT-COPD)

The dataset consists of chronic obstructive pulmonary disease information because air pollution directly affects people's lung function. In addition to this, the data set consists of lung changes information because most of the people's lung is affected due to the e smoking habit. The peoples are continuously monitored, lung resection information, RNA information is extracted. The dataset consists of a healthy patient, smoker information and moderated Chronic obstructive pulmonary disease (COPD) related gene information [22]. According to the dataset information, respective dataset details are depicted in Fig. 2.

Figure 2 describes the three criteria people and their habits distribution level. In addition to this information, the NIH Clinical dataset also used in this work to identify the efficiency of the system. More ever, this introduced system not only analysis the attributes, further it examines the people's lung image to predict the lung health condition.

3.2 NIH clinical dataset

The national institution health clinical center [23] analyzed more than 30,000 people and over 100,000 chest X-ray images are captured to examine the changes in their lung function. During the image capturing process, patients are instructed to perform the clinical trial and respective images are captured. The captured images are processed and unwanted (preprocessed) information is removed successfully. The X-ray images are easy to analyze by a radiologist, to detect different diseases successfully the effective and optimized techniques are introduced. The dataset includes several X-ray images and respective attributes are listed. The sample images and attributes with the label are depicted in Fig. 3.

Based on the above discussion, the people's lung information is collected from two datasets which are examined by applying the meta-heuristic optimized neural networks. After collecting the lung image, it should be normalized because it converts the data into the processing format. Here min–max normalization [30] process is used to normalize the data which improves lung data processing. This normalization process changes the data from 0 to 1 or 1 to − 1. Then the normalization is performed using Eq. (1).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

Here, x is the original lung value,

x' is the normalized value.

Based on Eq. (1), data is normalized and various statistical features such as mean, standard deviation, correlation, entropy, energy, and different features are extracted. As discussed already, the system utilizes more volume of data which is time consuming to predict people's health conditions. So, the system must be optimized by selecting relevant features from the collection of extracted features.

3.3 Hilbert–Schmidt independence criterion based features selection process

The important step of the work in selecting the optimized features which are done by applying the Hilbert–Schmidt Independence [31] Criterion based features selection method. The method works on both small feature set ($< 10^3$) and large dataset ($> 10^5$) to eliminate the irrelevant features and select the optimized features successfully. More ever, this approach resolves the Hilbert Schmidt independence criterion lasso problem that is defined as follows,

$$HSIC_{Lasso} : min_x \frac{1}{2}\sum_{k,l=1}^{n} x_k x_l HSIC(f_k, f_l) - \sum_{k=1}^{n} x_k HSIC(f_k, c) + \lambda \|x\|_1 \tag{2}$$

Here, $x_1, x_2, \ldots . x_n \geq 0$

Kernel-based independency feature measure is denoted as $HSIC(f_k, c) = tr(\bar{\mathrm{k}}^{(k)}\bar{\mathrm{L}})$

HSIC is Hilbert Schmidt's independence criterion (HSIC) [32].

Trace is denoted as tr.

Regularization parameter is represented as λ.

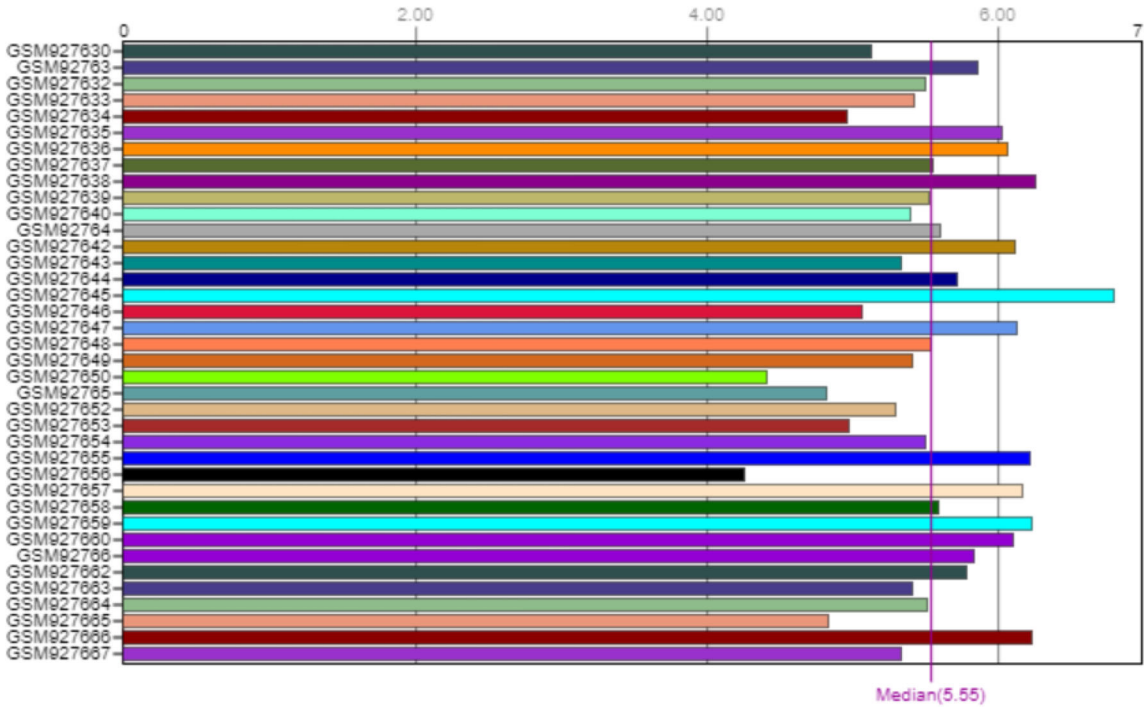

Fig. 2 ELT-COPD dataset attributed information

The input gram matrices are denoted as $\bar{\mathrm{k}}^{(k)} = \Gamma K^{(k)}$ output gram matrices $\bar{\mathrm{L}} = \Gamma L \Gamma$.

A kernel function is denoted as $K(u, u^{'})$, $L(c, c^{'})$.

Gram matrices are represented as $K_{i,j}^{(k)} = K(u_{k,i}, u_{k,j})$ and $L_{i,j} = L(c_i, c_j)$.

$\Gamma = I_m - \frac{1}{m} 1_m 1_m^T$ is denoted as the centering matrix, m dimensionality identity matrix is denoted as I_m with m number of samples.

All one of m dimensionality is represented as 1_m and l_1 is the norm.

Based on the feature dependency value the features are analyzed continuously. If the value is statistically independent which has 0 value else negative value. During the computation process. Gaussian kernel value is used to get the optimized features. Further, the HSIC is written as follows,

$$HSIC_{Lasso} : min_x \frac{1}{2} \left\| \overline{L} - \sum_{k=1}^{n} x_k \overline{K}^{(k)} \right\|_F^2 + \lambda \|x\|_1 \quad (3)$$
$$x_1, x_2, \ldots . x_n \geq 0$$

In Frobenius norm is denoted as $\|.\|$. After solving the lasso problem, the correlated features are computed using Eq. (4).

$$S_k = \frac{k\overline{rcf}}{\sqrt{k + k(k-1)\overline{rff}}} \quad (4)$$

In Eq. (4), feature correlation in average value is denoted as $\bar{\mathrm{r}}_{cf}$.

Feature and feature correlation value [33] is represented as, $\bar{\mathrm{r}}_{ff}$.

Correlation feature defined as follows,

$$CFS = max_{S_k} \left[\frac{r_{cf1} + r_{cf2} + \ldots . \ldots + r_{cfk}}{\sqrt{k + 2(rf_1f_2 + \ldots + rf_if_j + \ldots + rf_kf_{k-1}}} \right] \quad (5)$$

Correlation between the variables are denoted as $r_{cfi} and rf_if_j$.

According to the feature correlation [34], the optimized and independent features are selected for recognizing the lung disease-related information successfully (Table 1).

Based on the feature selection process, overfitting data is eliminated from the feature list and relevant features are selected. These selected features are used to reduces the computation time and maximize identification efficiency.

3.4 Lung feature learning process

The next important step is feature learning which is done with the help of convolution network [35]. During the training process, 6 network layers are used which implies the network size is 32*32 which extracts the features from the captured X-ray images. The 6 layers of network successfully train the features which improve the overall lung

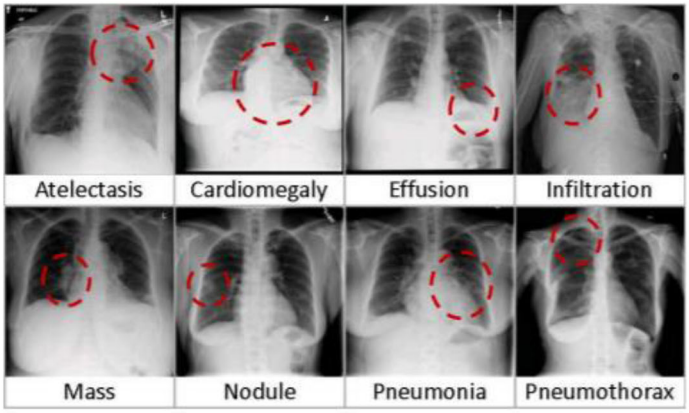

Image Index	Finding Labels	Follow-up #	Patient ID	Patient Age	Patient Gender	View Position	OriginalImage[Width	Height]	OriginalImagePixelSpacing[x	y]
00000001_000.png	Cardiomegaly	0	1	58	M	PA	2682	2749	0.143	0.143
00000001_001.png	Cardiomegaly|Emphysema	1	1	58	M	PA	2894	2729	0.143	0.143
00000001_002.png	Cardiomegaly|Effusion	2	1	58	M	PA	2500	2048	0.168	0.168
00000002_000.png	No Finding	0	2	81	M	PA	2500	2048	0.171	0.171
00000003_000.png	Hernia	0	3	81	F	PA	2582	2991	0.143	0.143
00000003_001.png	Hernia	1	3	74	F	PA	2500	2048	0.168	0.168
00000003_002.png	Hernia	2	3	75	F	PA	2048	2500	0.168	0.168
00000003_003.png	Hernia|Infiltration	3	3	76	F	PA	2698	2991	0.143	0.143
00000003_004.png	Hernia	4	3	77	F	PA	2500	2048	0.168	0.168
00000003_005.png	Hernia	5	3	78	F	PA	2686	2991	0.143	0.143

Fig. 3 NIH dataset information

Table 1 Algorithm for feature selection

Input : f_i*// extracted features, i=1,2,.....n*
Output: f_s*// feature subset*
Define feature range, $x_1, x_2, \ldots . x_n \geq 0$
For i= compute (feature dependency)
$HSIC(f_k, c) = tr(\overline{k}^{(k)}\overline{L})$
$K_{i,j}^{(k)} = K(u_{k,i}, u_{k,j})$ *and* $L_{i,j} = L(c_i, c_j)$
Then compute the correlation between features
$$S_k = \frac{k\overline{rcf}}{\sqrt{k + k(k-1)\overline{rff}}}$$
Afterward, find maximum correlated feature values.
$$CFS = max_{S_k}\left[\frac{r_{cf1} + r_{cf2} + \cdots \ldots . + r_{cfk}}{\sqrt{k + 2(rf_1f_2 + \cdots + rf_if_j + \cdots + rf_kf_{k-1}}}\right]$$
Choose maximum values as the best feature and form feature subset

disease recognition accuracy. In every layer, the convolution network uses the 2*2 kernel value that trains the selected low-level features. More ever, 32 to 192 kernel value for four layers and 1, 2 is set as last two layers kernel value. Finally, the softmax activation function is used to train the features. The softmax activation [36] function is defined using Eq. (6).

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_i}} \tag{6}$$

Then the activation function is optimized further using the cross-entropy loss function which is done Eq. (7).

$$Loss_{cls} = -\sum_{i=1}^{c} 1\{y = 1\} log \frac{e^{z_i}}{\sum_{i=1}^{c} e^{z_i}} \tag{7}$$

In Eq. (7), c represented as the number of class,

z_i is denoted as output value in the network layer.

Y is the respective label value for the input.

During the feature training process, the error value is reconstructed as follows,

$$Loss_{recon} = \|\hat{x}_i - x_i\|_2^2 \tag{8}$$

This computed error value should be minimized and the reconstructed error valuesare computed as follows.

$$J = Loss_{recon} + Loss_{cls} \tag{9}$$

Based on the above process, the features are continuously trained and labeled stored in a database which used to match the lung disease pattern effectively. After that rest of the people, lung information is processed, features are extracted which are classified with the help of introduced neural networks.

3.5 Lung disease identification

The final step of this work is lung disease identification using harmony optimized modular neural network. This neural network [37] is one of the effective independent artificial neural networks and changed by a few intermediary values. The incoming features are analyzed by a modular network that decomposes the features into sub-features and processes the network independently. The discussed intermediary process consumes each network output as input and produces the final output. The intermediary process only takes each module output and performs the respective actions that do not work on any signals and not interact with each other. This modular neural network reduces the computation complexity and also improves system performance and robustness. Due to this reason, in this work optimized modular network is used. According to the discussion, the general structure of a modular neural network is depicted in Fig. 4.

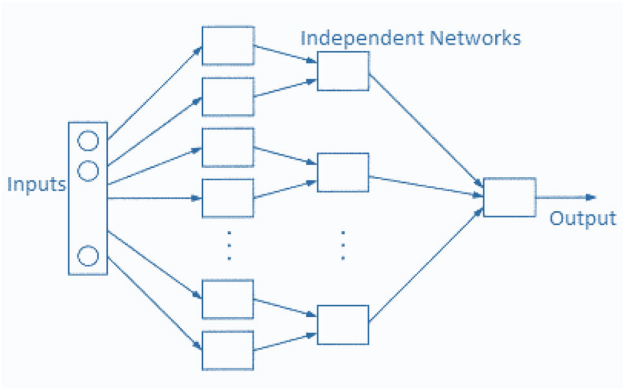

Fig. 4 Modular neural network structure

As similar to other neural networks, each module works independently and produce the output value. The output computation is performed using Eq. (10).

$$S = \left| \sum_{i=1}^{P} (V_i - V)(E_i - E) \right| \tag{10}$$

Here, S needs to be maximized because it is the output value of the network.

Ei is the error value in theith pattern, E is the network mean error value.

Vi is the hidden unit pattern output and V is the training pattern average output.

During this output computation process, the learning process needs to be improved by minimizing the error value. The network error value is optimized by selecting the effective network weight and a bias value. These values are chosen according to the optimization algorithm called harmony search [38]. It is one of the effective multi-objective algorithms, simple implementation, and little operands. This algorithm selects the best values by solving the multi-objective optimization problem (min/max). The multi-objective problem is resolved by applying the harmony memory initialization $((HM = x_i \in \Omega, i \in (1, \ldots HMS)$, improvisation and memory update process. After initializing the weight values, respective memory need to be created, $x_i^{new} = x_i^k, k \in (1, \ldots HMS)$. During the memory creation process, harmony parameters are updated continuously for adjusting the pitch value. The maximum value is selected as the best value and updated in memory. Based on this process network error values are adjusted. The computed output value is compared with the trained features, and the respective class labels are identified. This process is continuously performed to recognize urban space-people lung health status. The efficiency of the system is evaluated using experimental results.

4 Results and discussions

This section examines the excellence of meta-heuristic optimized modular neural network-based urban people's lung health status identification process. As mentioned in Sect. 3, this work uses the Expression data in lung tissue from moderate COPD patients, healthy smokers and non-smokers (ELT-COPD) and NIH Clinical dataset information. The collected information is processed by the above-defined algorithm procedures which effectively extracted and selected the optimized features. Due to the optimized feature selection, the data overfitting and high computation time issues are successfully resolved. More ever, the training process and meta-heuristic classifier rectify the prediction accuracy results on both small and huge volumes of the dataset.

4.1 Evaluation

The efficiency of the system is evaluated using the F1-score metric. The metric is computed by taking the harmonic mean value of precision and recall. The F1 score [39] is calculated using Eq. (11).

$$F_1 Score = \frac{1}{n}\sum_{c=1}^{n} 2 * \left(\frac{recall_c * Precision_c}{recall_c + Precision_c}\right) \tag{11}$$

In Eq. (11), c denotes as class value (normal and abnormal class), samples that correctly identified in class c is denoted as $recall_c$

Total test number of attributes present in the class is successfully classified is denoted as $Precision_c$.

Here, the computation is repeated up to 10 times of iteration to minimize the deviation of estimated and exact output value. More ever, the F1score value is applied to feature selection, feature training and classification process to examine the entire system efficiency. Here the collected dataset information is listed. The first dataset consists of 38 samples which include, nonsmoker, healthy smoker and moderate COPD.

These above tables clearly show that the efficiency of the system works on both high and small datasets. By consideration, the utilized feature selection accuracy is applied to Table 2 information and the obtained results are shown in Tables 3 and 4.

Table 2 ELT-COPD dataset information

Scheme	Training sample	Testing sample	Class
A	6	3	Nonsmoker
B	6	5	Healthy smoker
C	10	8	Moderate COPD

Table 3 NIH clinical dataset

Information	Images
Front view X-ray images	108,949
More than one pathology image	24,636
Normal images	84,312
Training	70%
Validation	10%
Testing	20%

From Table 4, it clearly depicted that the introduced algorithm meta-heuristic modular neural network successfully classifies the scheme A, B, and C with maximum results due to the effective feature selection algorithm. Comparing to the without feature selection accuracy and with feature selection process ensures the maximum results on scheme A, B, and C. Respective graphical analysis is shown in Fig. 5.

In addition to this, the efficiency of lung health condition prediction system efficiency is analyzed on the NIH health dataset. The efficiency of the system is evaluated using various images. According to the discussion, the excellence of the system with the feature selection algorithm is depicted in Fig. 6.

From Fig. 6, it clearly depicted that the Metaheuristic optimized modular neural network ensures good result on lung disease features. But the feature selection process improves the further efficiency of the system compared to other classifiers. Further, the excellence of the system is evaluated using the learning process. In this work effective learning algorithm called convolution neural network is used to train lung information. The convolution learning process automatically extracts information from the previous analysis and helps to recognize the lung disease successfully. The efficiency of the learning process is evaluated using the confusion matrix. The result obtained on the ELT-COPD database confusion matrix is depicted in Table 5.

Table 5 depicted the confusion matrix value on the ELT-COPD dataset information. The obtained results are drawn from predicted and actual class value. According to the results, the introduced optimized neural network effectively classifies the different schemes such as A, B, and C with a high rate. In addition to this, the method predicts the exact class also eliminates the wrong decision successfully. Further, the excellence of the system is analyzed on the NIH clinical dataset. The obtained confusion matrix is depicted in Table 6.

Table 6 demonstrates confusion matrix of the NIH clinical dataset information. From the result, it clearly depicted that the introduced approach successfully works

Table 4 Feature selection accuracy on ELT-COPD dataset

Network	Feature selection method	A	B	C
Convolution neural networks (CNN)	With feature selection	0.987	0.9765	0.973
	Without feature selection	0.982	0.984	0.983
Extreme learning neural network (ELNN)	With feature selection	0.9738	0.978	0.97
	Without feature selection	0.9647	0.947	0.965
meta-heuristic optimized modular neural network (MHOMNN)	With feature selection	0.993	0.993	0.986
	Without feature selection	0.976	0.965	0.943

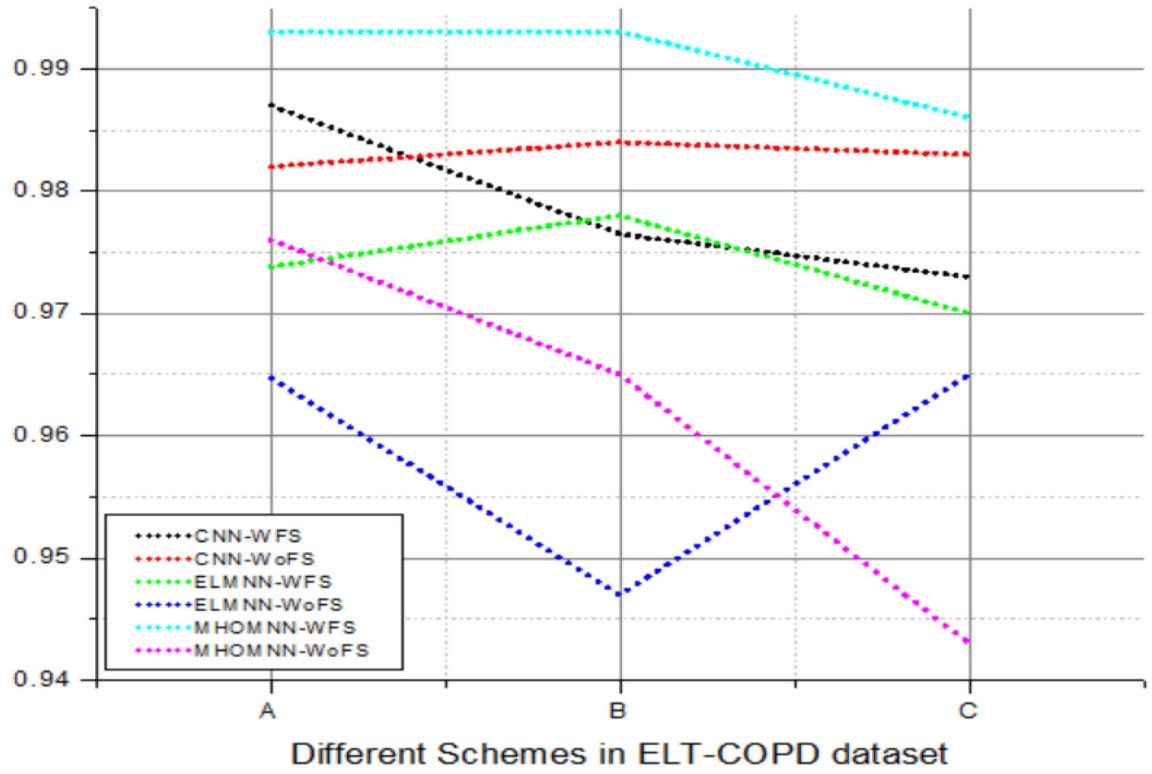

Fig. 5 Efficiency of ELT-COPD dataset

on both large and small datasets. More ever, the method recognizes the right lung disease features and eliminates the false decision very accurately due to the effective learning process. Further, the excellence of the system is compared with the literature reviewed by the authors' methods. The obtained results are depicted in Table 7.

Table 7 depicted the efficiency of introduced Meta-heuristic optimized modular neural network approach efficiency of ELT-COPD dataset and NIH clinical dataset values. From the analysis, it clearly depicted the introduced system ensures high accuracy (98.9% -ELT-COPD and 98% -NIH clinical dataset). The obtained accuracy is maximum compared to other research studies. The related graphical analysis is shown in Fig. 7.

5 Discussion

In this study the main aim to improve the urban people's health condition by diagnosing the disease in an earlier stage. For this purpose, the people's health information [43] is collected from two different dataset Expression data in lung tissue from moderate COPD patients, healthy smokers and nonsmokers (ELT-COPD) and NIH Clinical dataset. This information is used to analyze and detect urban people's lung condition due to air pollution. The first aim is to reduce the data overfitting because it completely reduces the efficiency of the system. To achieve this goal system uses the effective feature selection technique called Hilbert–Schmidt Independence Criterion based features selection approach. This method recognizes the Hilbert

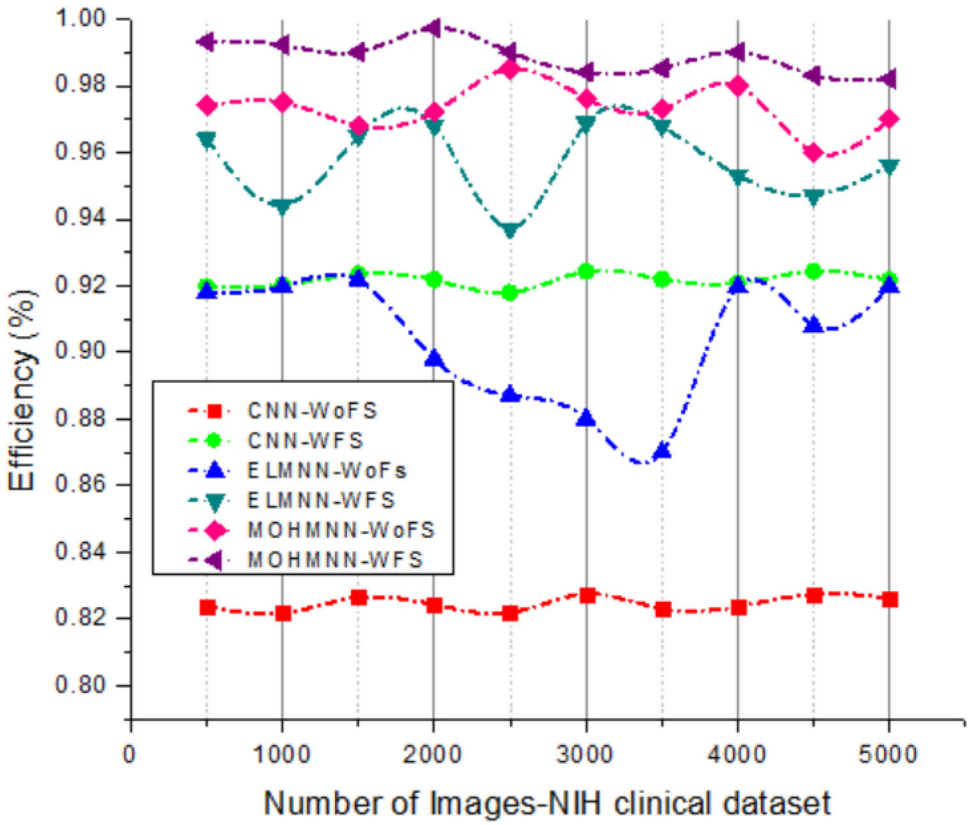

Fig. 6 Efficiency of NIH clinical dataset

Table 5 Confusion matrix

Ground truth	Prediction		
	A	B	C
A	0.97	0.03	0.00
B	0.00	0.96	0.04
C	0.01	0.02	0.97

Schmidt problem by computing the feature independency

$$HSIC_{Lasso} : min_x \frac{1}{2}\sum_{k,l=1}^{n} x_k x_l HSIC(f_k, f_l) - \sum_{k=1}^{n} x_k HSIC(f_k, c) + \lambda \|x\|_1$$

and the relationship, correlation $S_k = \frac{k\overline{rcf}}{\sqrt{k+k(k-1)\overline{rff}}}$ between the features are computed continuously. From the identified feature correlation, maximum correlated features

Table 6 NIH Clinical dataset –confusion matrix

Ground truth	Prediction							
	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumothorax	Pneumonia
Atelectasis	0.97	0.01	0.00	0.00	0.02	0.00	0.00	0.00
Cardiomegaly	0.00	0.03	0.02	0.01	0.00	0.94	0.00	0.00
Effusion	0.01	0.00	0.00	0.95	0.02	0.00	0.02	0.00
Infiltration	0.00	0.05	0.93	0.00	0.00	0.00	0.00	0.02
Mass	0.02	0.01	0.00	0.00	0.00	0.00	0.02	0.95
Nodule	0.00	0.00	0.00	0.00	0.01	0.02	0.96	0.91
Pneumothorax	0.96	0.03	0.01	0.00	0.00	0.00	0.00	0.00
Pneumonia	0.00	0.92	0.02	0.03	0.03	0.00	0.00	0.00

Table 7 Comparison of different researcher methods

Related study	Dataset	
	ELT-COPD dataset	NIH clinical dataset
Li et al. [40]	0.78	0.757
LeNet [41]	0.829	0.819
Huang et al. [24]	0.86	0.845
Selvanambi et al. [25]	0.879	0.856
ALuzubi et al. [26]	0.89	0.87
Alex net [18]	0.923	0.90
Pretrained VGG16 [17]	0.937	0.923
Pretrained Dense net [42]	0.956	0.918
Proposed approach	0.989	0.98

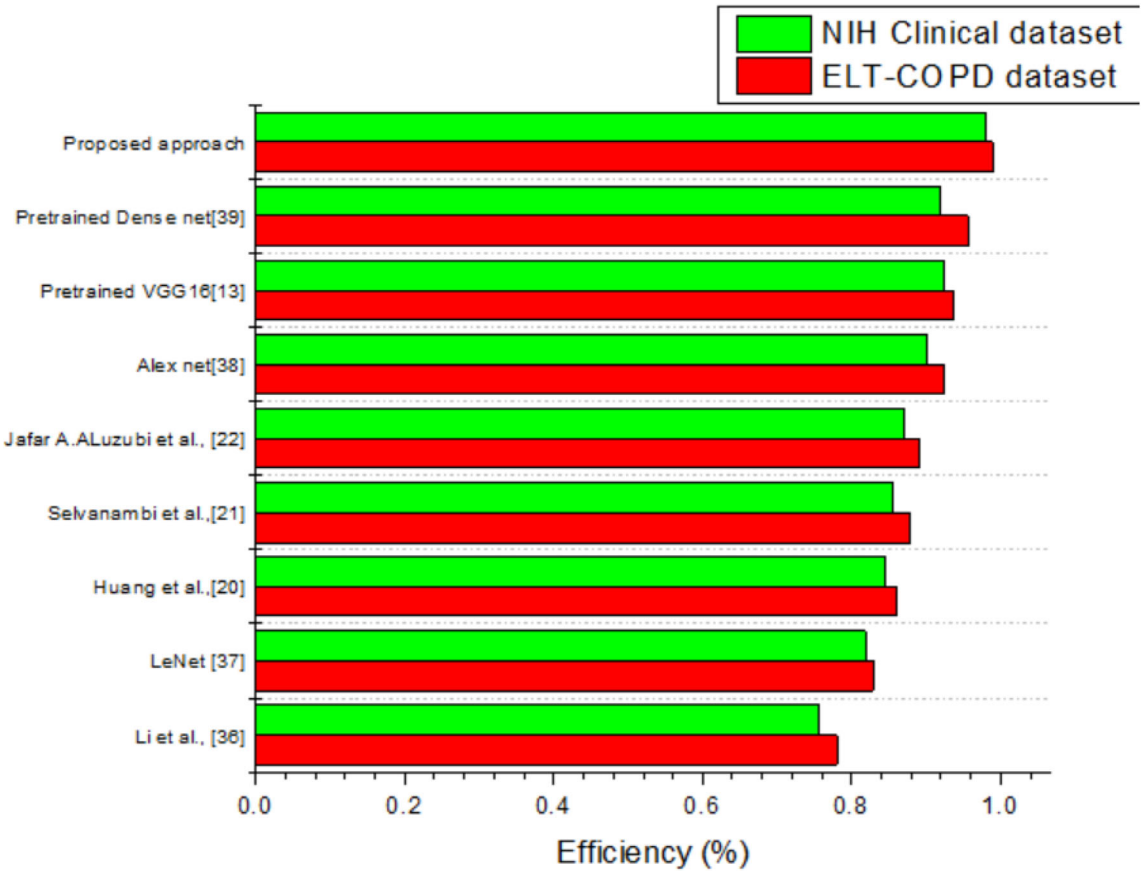

Fig. 7 Efficiency comparison of different researcher methods

$$max_{S_k}\left[\frac{r_{cf1}+r_{cf2}+\ldots\ldots+r_{cfk}}{\sqrt{k+2(rf_1f_2+\ldots+rf_if_j+\ldots+rf_kf_{k-1}}}\right]$$

are estimated and the maximum ranked features are selected as the best features. Due to the effective feature selection process, entire lung health condition identification system [44] performance is improved which is proved in Table 4, Figs. 5 and 6. Further, the lung health condition identification system process is enhanced via the effective convolution neural network-based learning process because it automatically trains the features [45] also reduces the error by reconstructing the output $J = Loss_{recon} + Loss_{cls}$. Because of the training process, the obtained results are improved which is shown in Tables 5 and 6. At last, the overall system performance is improved by decomposing the input into the small task [35, 46, 47] and the output error is adjusted by the harmony memory updating process.

The obtained results are depicted in Table 7 and Fig. 7. Thus the introduced system attains the discussed contribution by overcoming the defined problems in the conventional lung disease identification process.

6 Conclusion

Thus the manuscript analyzes the urban people's health condition due to air pollution. The air pollution creates several health problems, especially it affects the lung function. So, in this work, lung health condition is analyzed by applying the meta-heuristic optimization algorithm is used. Initially, the lung information is collected from two datasets such as ELT-COPD and NIH clinical datasets. Before processing the data, it converted into processing format 0 to 1 or 1 to − 1. Then various features are extracted which are analyzed and optimized features are selected depending on the feature correlation. The convolution layers are applied to train these features. At last, the modular neural network independently processes the input and produces the output by propagating the error value. Finally, the efficiency of the system is evaluated using the F1 score value, in which the system ensures the 98.9% - ELT-COPD and 98% -NIH clinical dataset. More ever, the excellence of the system is evaluated in feature selection and without a feature selection process. In the future, the large volume of lung images is taken to analyze further critical disease using optimized learning networks.

Acknowledgements This work was supported in part the Deanship of Scientific Research at King Saud University for funding this work through research group No. (RG-1439-53). This work was supported in part by Zayed University, office of research under Grant No. R17089.

References

1. von Schneidemesser, E., Monks, P.S., Plass-Duelmer, C.: Global comparison of VOC and CO observations in urban areas. Atmos. Environ. **44**(39), 5053–5064 (2010)
2. Fouad, H., Hassanein, A.S., Soliman, A.M., Al-Feel, H.: Analyzing patient health information based on IoT sensor with AI for improving patient assistance in the future direction. Measurement **159**, 107757 (2020)
3. Weber, N., Haase, D., Franck, U.: Assessing modelled outdoor traffic-induced noise and air pollution around urban structures using the concept of landscape metrics. Landsc. Urban Plan. **125**, 105–116 (2014). https://doi.org/10.1016/j.landurbplan.2014.02.018
4. Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., et al.: A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. **42**(33), 7561–7578 (2008). https://doi.org/10.1016/j.atmosenv.2008.05.057
5. Shakeel, P.M., Baskar, S., Dhulipala, V.S., Mishra, S., Jaber, M.M.: Maintaining security and privacy in health care system using learning based Deep-Q-Networks. J. Med. Syst. **42**(10), 186 (2018). https://doi.org/10.1007/s10916-018-1045-z
6. Lim, S.S., Vos, T., Flaxman, A.D., Danaei, G., Shibuya, K., Adair-Rohani, H., et al.: A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet **380**(9859), 2224–2260 (2012). https://doi.org/10.1016/S0140-6736(12)61766-8
7. Schindler, M., Caruso, G.: Urban compactness and the trade-off between air pollution emission and exposure: lessons from a spatially explicit theoretical model. Comput. Environ. Urban. **45**, 13–23 (2014). https://doi.org/10.1016/j.compenvurbsys.2014.01.004
8. Jarjour, S., Jerrett, M., Westerdahl, D., de Nazelle, A., Hanning, C., Daly, L., et al.: Cyclist route choice, traffic-related air pollution, and lung function: a scripted exposure study. Environ. Health. **12**(1), 1–12 (2013). https://doi.org/10.1186/1476-069X-12-14
9. Shakeel, P.M., Burhanuddin, M.A., Desa, M.I.: Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks. Measurement (2019). https://doi.org/10.1016/j.measurement.2019.05.027
10. Al Kheraif, A., Wahba, A., Fouad, H.: Detection of dental diseases from radiographic 2d dental image using hybrid graph-cut technique and convolutional neural network. Measurement **146**, 333–342 (2019)
11. Alsiddiky, A., Awwad, W., Bakarman, K., Fouad, H., Mahmoud, N.: Magnetic resonance imaging evaluation of vertebral tumor prediction using hierarchical hidden Markov random field model on Internet of Medical Things (IOMT) platform. Measurement **159**, 107772 (2020)
12. Suzuki, A., Sakanashi, H., Kido, S., Shouno, H.: Feature representation analysis of deep convolutional neural network using two-stage feature transfer-an application for diffuse lung disease classification. arXiv:1810.06282 (2018).
13. Manogaran, G., Shakeel, P.M., Fouad, H., Nam, Y., Baskar, S., Chilamkurti, N., Sundarasekar, R.: Wearable IoT smart-log patch: An edge computing-based Bayesian deep learning network system for multi access physical monitoring system. Sensors **19**(13), 3030 (2019)
14. Mahmoud, N.M., Fouad, H., Alsadon, O., Soliman, A.M.: Detecting dental problem related brain disease using intelligent bacterial optimized associative deep neural network. Cluster Comput. (2020). https://doi.org/10.1007/s10586-020-03104-3
15. Fouad, H., Hassanein, A., Soliman, A., Al-Feel, H.: Internet of medical things (IoMT) assisted vertebral tumor prediction using heuristic hock transformation based gautschi model–a numerical approach. IEEE Access **8**, 17299–17309 (2020)
16. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: full training or fine tuning. IEEE Trans Med Imag **35**(5), 1299–1312 (2016)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014).
18. Krizhevsky, A., Sutskever, I., Hinton, GE.: Imagenet classification with deep convolutional neural networks. Proc. Adv. Neural Inf. Process Syst. 1097–1105 (2012).
19. . Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 1–9 (2015).
20. Al Kheraif, A.A., Alshahrani, O.A., Al Esawy, M.S.S., Fouad, H.: Evolutionary and Ruzzo-Tompa optimized regulatory feedback neural network based evaluating tooth decay and acid erosion from 5 years old children. Measurement **141**, 345–355 (2019)

21. Liaghat, S., Mansoori, E.G.: Filter-based unsupervised feature selection using Hilbert-Schmidt independence criterion. Int. J. Mach. Learn. Cyber. **10**, 2313–2328 (2019). https://doi.org/10.1007/s13042-018-0869-7
22. https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-37768/
23. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R. M.:Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)
24. Huang, S., Lee, F., Miao, R., et al.: A deep convolutional neural network architecture for interstitial lung disease pattern classification. Med. Biol. Eng. Comput. (2020). https://doi.org/10.1007/s11517-019-02111-w
25. Selvanambi, R., Natarajan, J., Karuppiah, M., et al.: Lung cancer prediction using higher-order recurrent neural network based on glowworm swarm optimization. Neural Comput. Appl. (2018). https://doi.org/10.1007/s00521-018-3824-3
26. ALzubi, J.A., Bharathikannan, B., Tanwar, S., Manikandan, R., Khanna, A., Thaventhiran, C.: Boosted neural network ensemble classification for lung cancer disease diagnosis. Appl. Soft Comput. **80**, 579–591 (2019)
27. Hankey, S., Marshall, J.D.: Urban form, air pollution, and health. Curr. Environ. Health Rep **4**, 491–503 (2017). https://doi.org/10.1007/s40572-017-0167-7
28. Zhang, J., Day, D.: Urban air pollution and health in developing countries. In: Nadadur, S., Hollingsworth, J. (eds.) Air Pollution and Health Effects Molecular and Integrative Toxicology. Springer, London (2015)
29. Jiao, K., Xu, M., Liu, M.: Health status and air pollution related socioeconomic concerns in urban China. Int. J. Equity Health **17**, 18 (2018). https://doi.org/10.1186/s12939-018-0719-y
30. Mohamad Mohsin, M.F., Hamdan, A.R., Abu, B.A.: The effect of normalization for real value negative selection algorithm. In: Noah, S.A., et al. (eds.) Soft Computing Applications and Intelligent Systems. M-CAIT 2013. Communications in Computer and Information Science. vol. 378, Springer, Berlin, Heidelberg (2013)
31. Liu, C., Ma, Q., Xu, J.: Multi-label feature selection method combining unbiased Hilbert-Schmidt independence criterion with controlled genetic algorithm. In: Cheng, L., Leung, A., Ozawa, S. (eds.) Neural Information Processing. ICONIP 2018. Lecture Notes in Computer Science. vol. 11304, Springer, Cham (2018)
32. Xu, J.: Effective and efficient multi-label feature selection approaches via modifying Hilbert-Schmidt independence criterion. In: Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D. (eds.) Neural Information Processing. ICONIP 2016. Lecture Notes in Computer Science. vol. 9949, Springer, Cham (2016)
33. Savić, M., Kurbalija, V., Ivanović, M., Bosnić, Z.: A feature selection method based on feature correlation networks. In: Ouhammou, Y., Ivanovic, M., Abelló, A., Bellatreche, L. (eds.) Model and Data Engineering. MEDI 2017. Lecture Notes in Computer Science. vol. 10563, Springer, Cham (2017)
34. Chen, S., Ding, C.H.Q., Zhou, Z., et al.: Feature selection based on correlation deflation. Neural Comput. Appl. **31**, 6383–6392 (2019). https://doi.org/10.1007/s00521-018-3467-4
35. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imag. **35**(5), 1285–1298 (2016)
36. Agnes, S.A., Anitha, J., Pandian, S.I.A., et al.: Classification of mammogram images using multiscale all convolutional neural network (MA-CNN). J. Med. Syst. **44**, 30 (2020). https://doi.org/10.1007/s10916-019-1494-z
37. Shukla, A., Tiwari, R., Kala, R.: Modular neural networks. In: Towards Hybrid and Adaptive Computing. Studies in Computational Intelligence, vol 307. Springer, Berlin, Heidelberg, pp. 307–335 (2010). https://doi.org/10.1007/978-3-642-14344-1_14
38. Du, K. L., Swamy, M. N. S.: Harmony search. In: Search and Optimization by Metaheuristics, pp. 227–235. Birkhäuser, Cham (2016). https://doi.org/10.1007/978-3-319-41192-7_14
39. Lipton, Z.C., Elkan, C., Naryanaswamy, B.: Optimal thresholding of classifiers to maximize F1 measure. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2014. Lecture Notes in Computer Science. vol. 8725, Springer, Berlin, Heidelberg (2014)
40. Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, DD., Chen, M.: Medical image classification with convolutional neural network. Proc. 13th Int. Conf. Control Automat. Robot Vis. IEEE, 844–848 (2014)
41. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
42. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, KQ.: Densely connected convolutional networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 4700–4708 (2017)
43. Baskar, S., Shakeel, P.M., Kumar, R., Burhanuddin, M.A., Sampath, R.: A dynamic and interoperable communication framework for controlling the operations of wearable sensors in smart healthcare applications. Comput. Commun. **149**, 17–26 (2020)
44. Prasad, A., Gray, C.B., Ross, A., Kano, M.: Metrics in urban health: current developments and future prospects. Annu. Rev. Public Health. **37**, 113–133 (2016). https://doi.org/10.1146/annurev-publhealth-032315-021749
45. Tarando, S.R., Fetita, C., Faccinetto, A., Brillet, P.Y.: Increasing cad system efficacy for lung texture analysis using a convolutional network. In: Medical imaging 2016: Computer-aided diagnosis, 9785, 97850Q (2016)
46. Rey Gozalo, G., BarrigónMorillas, J.M., Trujillo Carmona, J., Montes González, D., Atanasio Moraga, P., Gómez Escobar, V., et al.: Study on the relation between urban planning and noise level. Appl. Acoust. **111**, 143–147 (2016). https://doi.org/10.1016/j.apacoust.2016.04.018
47. Zhuang, F., Cheng, X., Luo, P., Pan, S.J., He, Q.: Supervised representation learning: transfer learning with deep autoencoders. Proc. 24th Int. Conf. Artif. Intell. (2015)

Ahed Abugabah is an Assistant Professor at the College of Technological Innovation at Zayed University. He worked in higher education in Australia where he received his degrees in information systems. Dr. Ahed also worked in the airline industry in the Aircraft Engineering and Supply Chain management. Before Joining Zayed University in 2016. Dr. Abugabah was involved in administration as an Associate Dean and a University Council Member at the American University, UAE. His research interests

include Information Systems, Enterprise Applications, and Development, Healthcare Information Systems and RFID in Healthcare.

Ahmad Ali AlZubi is a full Professor at King Saud University (KSU). He obtained his Ph.D. from the National Technical University of Ukraine NTUU in Computer Networks Engineering in 1999. His current research interest include Computer Networks, Cloud computing, Big Data and Data Extracting. AlZubi is a certified member of the Board of Assessors of Quality Management System at King Saud University (BOA-QMS). He also served for 3 years as a consultant and a member of the Saudi National Team for Measuring EGovernment in Saudi Arabia.

Feras Al-Obeidat is an Associate Professor at the College of Technological Innovation at Zayed University. Dr Al-Obeidat received both his Master and PhD in Computer Science from the University of New Brunswick, Canada. Dr Al-Obeidat's primary field of research is Artificial Intelligence, Data mining and Machine Learning.

Abdulaziz Alarifi is currently a associate professor at Computer Science Department, Community College, King Saud University (KSU) Saudi Arabia. He received his Ph.D. degrees in Information Security from University of Wollongong, Australia. He is currently the head of Research Unit at Community College, KSU. His main research interests include Information Security, Information Technology Management, Cloud Computing, Big Data, Information Privacy, Risk Assessment and Management, Internet of Things (IOT), E- Governance and Mobile Applications

Ayed Alwadain is an Assistant Professor at the Computer Science Department at the Community College, King Saud University, Saudi Arabia. He received his PhD in 2014, from Queensland University of Technology in Australia. In his research, he focuses on Enterprise Architecture, Service Management and Engineering, Business Process Management, Requirement Engineering and Big Data. Ayed has published his work at many international conferences and journals such as Data & Knowledge Engineering and the International Journal of Intelligent Information Technologies.