International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 5, September - October 2016 **ISSN 2278-6856**

Prediction of Pneumoconiosis Disease Using Machine Learning Techniques

Soha Safwat Labib[1]

[1] October University for Modern Science and Arts , Faculty of Computer Science, Egypt

Abstract

Pneumoconiosis (Silicosis) is considered to be lung disease and it is a common one in Egypt as its predominance rate ranges from 18.5 % to 45.8% among workers who are at risk of free crystalline silica dust exposure. The main reason behind its occurrence is the inhalation of dust, often in mines. The aim of this paper is to predict the Pneumoconiosis disease among the workers by the use of machine learning techniques as Pneumoconiosis can be prevented but not considered to be a treatable disease. We use principle component analysis algorithm to remove the redundant attributes, and then we use k-nearest neighbor and neural network classifiers in the classification phase. Results show that the weighted k-nearest neighbor classifier outperforms a wide variety of the neural network classifier.

Keywords: Principle Component Analysis, Neural Network, K-Nearest Neighbor, Weighted K-Nearest Neighbor.

1. INTRODUCTION

As most of adults spend nearly about one third of their time at work, workers may be exposing to different hazards, which may have remarkable damaging effect on their health. According to the world population, workers represent 50% and contribute significantly to socioeconomic development. Their health maintained by the occupational health services available to them at their place of work.[1]

World Health Organization (WHO) assessed that people who are suffering from chronic respiratory diseases represent over than 1 billion worldwide, moreover 4 million people die annually. In details, nearly 300 million suffer from asthma, 210 million has chronic obstructive pulmonary disease, including occupational diseases and pulmonary hypertension.[2]

Moreover, respiratory disease (e.g. asthma, chronic obstructive pulmonary disease and silicosis) and malignancies represent about 70% of all occupationally-related deaths.

Occupational diseases are chronic conditions that caused by work or occupational activities. They are the illnesses or conditions that the workers suffered from according to the contact with at the workplace or while they were working at their respective work environment. Globally, there are more than 2.6 billion workers who are exposed to hazardous risks at their work places. Approximately 75% of these workers are in developing countries, where workplace hazards are more severe. If one includes occupational illnesses, an estimated 1.1 million people worldwide die each year. Annually, an estimated 160 million new cases of work- related diseases occur worldwide pneumoconiosis is a group of interstitial lung diseases produced by the inhalation of certain dusts and the lung tissue's reaction to the dust. The main cause of the pneumoconiosis is work-place exposure. The most common pneumoconiosis is asbestosis, silicosis, and coal workers' pneumoconiosis. As their names indicate, they are the result of inhalation of asbestos fibers, silica dust, and coal mine dust.[3]

If iron ore remains in the tissues, it may cause conjunctivitis, choroiditis, and retinitis. Chronic inhalation of iron oxide fumes excessive concentrations or dusts may result in the development of a benign pneumoconiosis. Inhalation of excessive concentrations of iron oxide may enhance the risk of lung cancer development in workers exposed to pulmonary carcinogens.

The term "Pneumoconiosis" is used for the diseases caused by the inhalation of mineral dusts. While many of these broad spectrum substances may be encountered in the general environment, many occur in the work-place for greater amounts as a result of industrial processes; therefore, a range of lung reactions may occur as a result of work-place exposure. Pneumoconiosis are more common in males because of the increased risk of occupational exposure. The age of most of the workers who are suffering from pneumoconiosis are over 40 [4].

The lung disease Pneumoconiosis is caused by the inhalation of various types of industrial dust. This dust is the main cause of lung inflammation and gradually damages the lungs over time. Finally, it may leads to fibrosis which is a condition where the lungs begin to stiffen [5].

1.1 Types of pneumoconiosis

There are several types of pneumoconiosis that affect people living in the United States (U.S). Among them are

coal worker's pneumoconiosis, asbestosis, silicosis and siderosis of the lung, talc pneumoconiosis, and kaolin pneumoconiosis. Each form of the disease can cause serious health apprehensions. Coal worker's pneumoconiosis is caused by inhaling coal dust, graphite, carbon black, or lamp black. People who frequently work with this type of dust, such as coal miners, contract this disease if they do not wear protective equipment. Asbestosis is a disease that often affects construction workers, auto mechanics, and other people who work with asbestos. People who live or work in old buildings that were constructed with asbestos can contract this form of the disease. It may often take about 20 years before symptoms become visible. Silicosis is often diagnosed in people who work with a substance called silica. Miners, sandblasters, quarry workers, silica millers, and those who make glass or ceramic will often suffer from silicosis. Siderosis of the lung is caused by the inhalation of iron particles. There are usually no symptoms present with siderosis of the lung. Talc pneumoconiosis is caused by exposure to talc dust. Kaolin pneumoconiosis results from the inhalation of kaolin. [6]

1.2 Signs and symptoms of pneumoconiosis

Pneumoconiosis can take several years to develop. Large amounts of dust can lead to lung swelling. The main signs and symptoms of lung damage from dust are shortness of breath, wheezing, chronic cough, sore throat, cyanosis, fever, chest pain, bronchitis, loss of appetite and emphysema.[7]

1.3 Diagnosis of pneumoconiosis

National Jewish health, reported that, several tests are commonly used for the diagnosis and monitoring of pneumoconiosis.[8] As a result for that many tests takes place such as detailed medical, occupational and environmental history and physical examination are always required. Moreover, Pulmonary Function. Tests (PFTs) are measured to determine the percentage of lung capacity and function that may have been lost due to scarring; Measurements of Arterial Blood Gases.
(ABGs) are used to determine if the exchange of oxygen and carbon dioxide in the alveoli is impaired, and imaging procedures, such as chest X-rays and CT scans, are used to visualize the nodules and scarring in the lungs.
The clinical diagnosis of pneumoconiosis is usually based on an occupational history, chest radiographic findings and compatible pulmonary function tests. All patients who present with respiratory symptoms should have a work and environmental history recorded. As many occupational exposures to various dusts have their effect years after exposure, a life-time exposure history should be obtained covering all previous and current jobs. It is also important to ask about smoking history. Cigarette smoking is associated with an increased risk of silicosis and coal workers' pneumoconiosis .[9]

1.4 Prognosis of pneumoconiosis

The outlook for this disease depends on the specific type of pneumoconiosis, the length of exposure to mineral dust, the level of exposure and whether the patient is a smoker. In the long term, people with asbestosis and talc pneumoconiosis have an increased risk of lung cancer and malignant mesothelioma (cancer of the membranes lining the lungs and abdominal cavity). The risk of lung cancer is especially high in smokers with asbestosis. Because male workers fill most of the jobs that carry high risks of pneumoconiosis, the majority of deaths from pneumoconiosis occur in men.[10]
The remaining part of the paper is organized as follows. proposed model is discussed in part (2). Part (3) presents the Experimental results. Part (4) presents limitation of the study. Finally, the paper conclusion is in part (5).

2. THE PROPOSED SYSTEM

2.1 Dataset

The study was conducted at El Gedida Iron Mine area at Bahariya Oasis, Giza Governorate. The Bahariya Oasis is located 360 kilometers to the South West of Cairo in the Western Desert of Egypt. Bahariya Oasis mine is the only one working now in Egypt and the exploitation of these mines began in 1973. El Gedida area as shown in figure 1 is located in the Bahariya northern plateau, 15 km south 56° east of Gabel Ghorabi and 11 km north 22° east El–Harra triangulation point. Egyptian iron ore is mined in El Gedida area of El Bahariya Oasis in the Western Desert. The nearly 3 million tons/year produced from this mine deposit is destined for hadid solb Helwan Iron and Steel works near Cairo. Its production provides about three-quarters of Egypt's demands from the iron.
Based on a wide review of recent literature, a structured interviewing questionnaire was developed by the investigator. Five experts from community health nursing department at the Faculty of Nursing, Cairo University were asked to check the tool for its content validity.

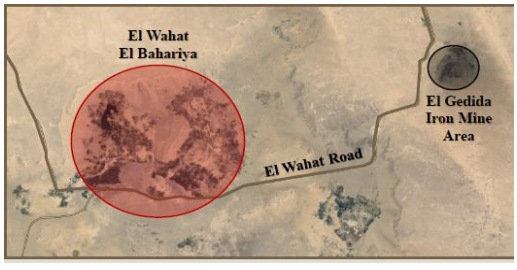

Figure 1: El Gedida Iron Mine area

As shown in table 1 almost half (49.6%) of the workers don't have chronic illnesses, while one fourth (25.4%) of them were diagnosed with silicosis. Regarding workers disabilities, the table shows that, about three quarter (74.6%) of the workers don't have occupational disability with a mean disability percentage of 15.49±6.1. Table 2 show that all attributes of the data set.

TABLE1. Data set description

Items	**Total Sample (n=240)**	
	Number	**%**
Types of chronic illnesses: Hypertension	20	8.3
Diabetes Mellitus	19	7.9
Heart diseases	4	1.6
Asthma	5	2.1
Pneumoconiosis (Silicosis)	61	25.4
More than one illness	12	5.1
No chronic illness	119	49.6
Disability percentage: 5%	1	0.4
10%	21	8.8
15%	22	9.2
20%	9	3.8
25%	6	2.5
More than 25%	2	0.8
No disability percentage	179	74.5
Mean±SD	15.49±6.1	

TABLE 2. All attributes

Attributes Names	**Attributes Names**	**Attributes Names**
Age	Place of compensations:	work accidents
Educational level	pneumoconiosis	Preventive measures
Place of residence	Source of knowledge	dust pollution
Marital status	Symptoms	personal hygiene
Job	early detection	hand washing
Work experience	diagnosis	uniform washing
Work shift	Medical treatment	Cigarettes smoking
Problems at work	iron dust	Effects of cigarettes
Type of Problems	Control measures	Sports/exercise
Types of chronic illnesses	protective measures	of meals/per day
Disability percentage	personal protective	Dietary intake prepared
Types of medical services	Supervisor	Training courses
of follow-up	Health care	Place of health insurance
Work accidents	First-aid	

2.2 The proposed system

As shown in figure 2, we have 48 attributes for each worker; one of them is the class label; there are two classes class 1 that is the presence of the disease and class 0 that is the absence of the disease. There are 240 samples in the data set; 61 samples for class 1 and 179 sample for class 0. There are two phases in the proposed method; the first phase is the preprocessing phase by using principle component analysis algorithm to reduce the redundant attributes; the second phase is the classification phase; in this step we used to classification algorithm; the first algorithm is weighted K-nearest neighbor classifier and the second algorithm is artificial neural network

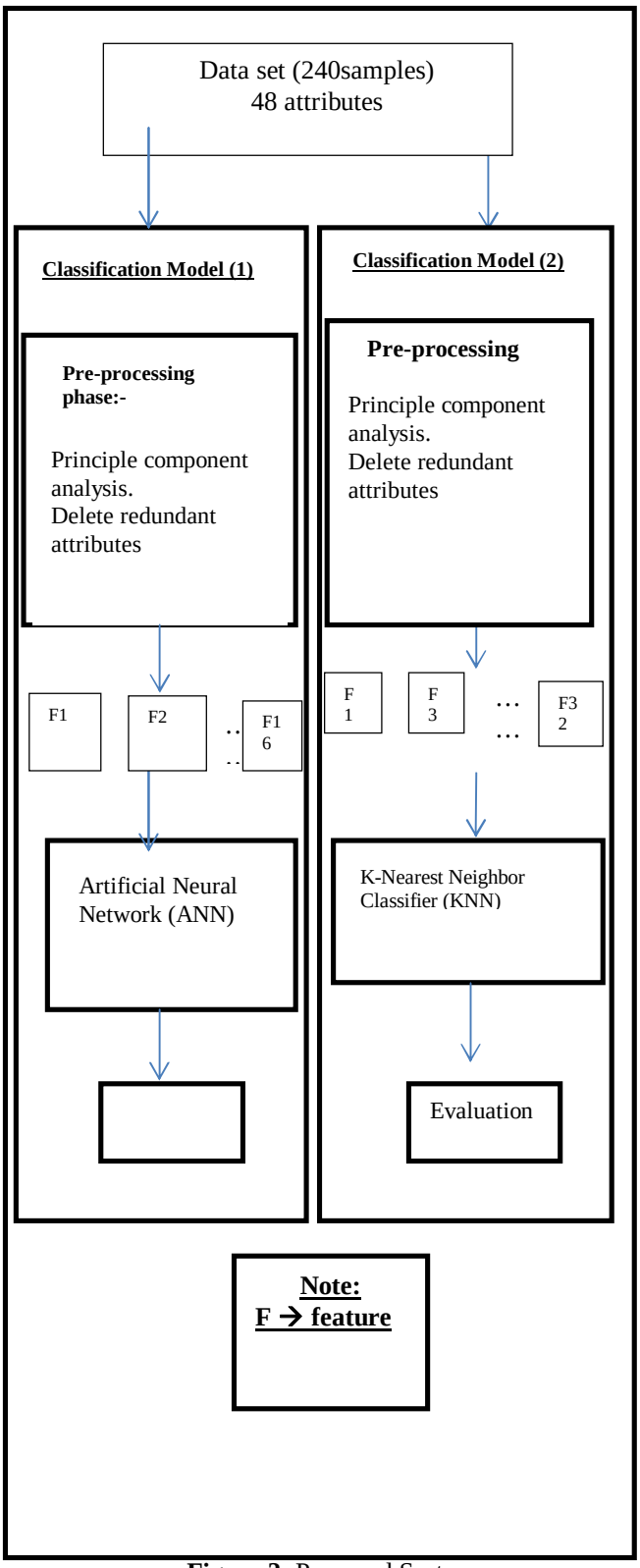

Figure 2: Proposed System

2.2.1 Preprocessing using Principle Component Analysis

Principle component analysis PCA is tool that can be used to reduce the dimensions of a given set of variables to smaller set preserving information in the given large set of variables. Also PCA can use number of correlated variables to transform it to uncorrelated ones which is called principle components.

PCA firstly uses the given dataset but after ignoring the class labels if found. A mean vector is then calculated from the dataset. Covariance matrix is computed using the dataset. Eigen vectors and Eigen values are then computed. By using the Eigen values obtained Eigen vectors are sorted in decreasing order. Finally, number of Eigen vectors is used to transform the samples to the new reduced subspace.[11]

Algorithm 1 " PCA"

Given: N samples $x_1, x_2, \ldots . x_N$ each example $x_n \in R^D$
Goal: Project the data from D dimensions K dimensions (K<D)
Want to capture the maximum possible variance in the projected data
Let $u_1, \ldots, u_D$ *be the principle components, assumed to be orthognal such that:*

$$u_i^T u_j = 0 \text{ if } \neq j \text{ and should be orthonormal such that } u_i^T u_i = 1$$

Each peincipal component is a vector of size $D \times 1$
We will select the frirst K principal components
1: *Compute the mean of the data*

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$$

2: *Compute the Covariance matrix*

$$s = \frac{1}{N}\sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$$

3: *Find the eign values* 4
4: *Take the top k eign vectors according* (*corresponding to the top k eign values*)
5: *Call these vectors as* $u_1, u_2, \ldots u_k$
(*such that* $\lambda_1 \geq \lambda_2 \ldots\ldots \geq \lambda_{k-1} \geq \lambda_k$
6: $Z_n = U^T \times X_n$

2.2.2 Classification

In this paper we classify the data using K-Nearest neighbor algorithm in model 2, in model 1 we use Artificial neural network algorithm.

1.Weighted K-Nearest neighbor algorithm (KNN)

K-nearest neighbor is type of supervised learning techniques where the outcome of new instance query is classified based on majority of K-nearest neighbor category. Classification of a new object based on attributes and training samples is the purpose of this algorithm.

Models to fit are not used by the classifiers as they are only based on memory. We find K number of objects or (training points) closest to the query point by a given query point, Majority vote among the classification of the K objects is used by the classification. Any ties can be broken at random. The prediction value of the new query instance used by K Nearest neighbor algorithm is neighborhood classification.

In pattern recognition, the k-nearest neighbor algorithm (KNN) is a method for classifying objects based on closest training examples in the feature space. The K-Nearest Neighbor algorithm stores the training instances and uses a distance function to determine which k members of the training set are closest to an unknown test instance.

KNN algorithm depends on four main parameters:

- The value of k
- The distance measurements
- The distance weighting measure (We use Euclidean distance)
- The process of votes counting

The final step is used to improve the results of the classification step, we use weighted function that maps the accuracy obtained in the classification step by values ranges from 0 to 1 by giving the best result with the highest value 1 and the minimum classifier with a value 0. Equation 4.1 displays the weighted function that is used in the experiment. After calculating the weighted values, we take the maximum class label.[12]

$$Class = MAX[classifier_{output(j,i)} + ((Results(i) - Minimum)/(Maximum - Minimum)$$

Algorithm2" Weighted K-Nearest Neighbor (KNN) algorithm"

1: Repeat the following steps to all samples
2. Calculate the distance between sample x and all samples in the training data

$$D = \sqrt[2]{\sum (X_i - S_i)^2}$$

3: Sort the distances ascending
4: Pick first K elements
5: $Output_X = majority_{Class}(elements)$

2. Neural Network

We use Artificial Neural Network (ANN) technique in classification step because it is used to solve complexity problems. Artificial network adapts itself by sequential training algorithm and its architecture and connected weights.

The network of neurons in the brains are intended to be represented by an artificial neural network (ANN), which can be described as an interconnected group of nodes. Because of their ability to learn complex patterns, they are widely used in literature.[13, 14]

Algorithm 3 Artificial Neural Network (ANN) algorithm

1: initialize weights (set to small random value).
while stopping condition is false do steps 2-9
2: for each sample in the training set, do steps 3-8
Feed forward :-
3: Each input unit (Xi) receives signal Xi &
broad casts this signal to all units in the layer above (the hidden layer)
4: Each hidden unit (Zj) sums its weighted i/p signals,

$$Z - i_{nj} = V_{aj} + \sum_{i=1}^{n} x_i v_{ij} \text{ , Where } V_{aj} \text{ is a bias}$$

then applies its activation function to compute its output signal

$$Z_j = 1/(1 + e^{-(z-inj)})$$

sends this signal to all units in the layer above
5: Calculate the output:

$$Y - i_{nk} = W_{ok} + \sum_{i=1}^{n} Z_j w_{jk} \text{ , Where } W_{ok} \text{ is a bias}$$

$$Y_k = 1/(1 + e^{-(y-i_{nk})})$$

Back propagation of error:-
6: Computes the error in the output layer

$$\delta_{2k} = Y_k(1 - Y_k) * (T_k - Y_k), T_k \text{ is the target}$$

7: computes its error information in hidden layers

$$\delta_{1j} = Z_j(1 - Z_j) * \sum_{k=1}^{m} \delta_{2k} w_{1k},$$

8: Update weights and bias :-

$$W_{jk}(new) = \eta * \delta_{2k} * Z_j + \alpha * W_{jk}(old)$$
$$V_{ij}(new) = \eta * \delta_{1j} * x_i + \alpha * V_{ij}old$$

9: Test stopping condition.

3. Experimental Results:

This section is divided into 2 parts; A) discussed the results when we used a KNN classifier. In B) we discussed the results when we used artificial neural network.

3.1 Weighted KNN Classifier results

Table 3. Weighted K-Nearest Classifier Results

Method	Accuracy
KNN with k=10	94%
KNN with k=20	90%
KNN with k=15	92%

As shown in table 3 we can see that the Weighted KNN classifier takes only one parameter and the best results were when k=10 that is gives us 94% accuracy.

3.2 Artificial Neural Network Classifier results

Table 4. Proposed System Results

Method	Accuracy
NN, (13,7,5)	83.2%
NN, (15,10,7)	80.7%
NN, (10,7,7)	79.8%

In NN we pick number of layers and nodes in each layer to train the samples and get the weights of the network, we test the data in different cases in case of using three layers, in the first layer we pick 13 nodes, 7 nodes in the second layer and 5 nodes in the output layer we got 83.2% accuracy.

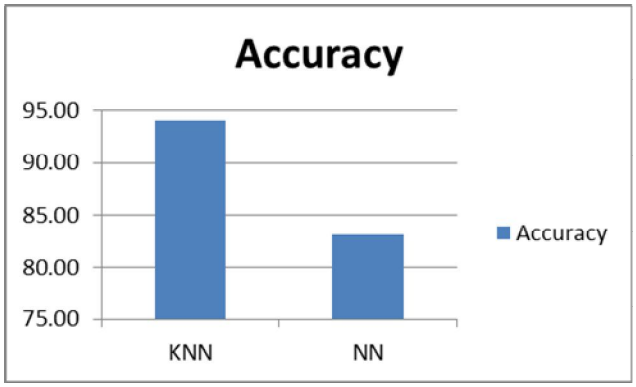

Figure 3: Comparison between NN and KNN

As shown in figure 3 we can see that Weighted KNN classifier give us better accuracy than NN classifier.

4. STUDY LIMITATIONS

One of the limitations of our study was that we have only 240 workers. Approximately 25% with class 0 and 75% with class 1 This limited number of cases may be the reason that NN gives us lower accuracy By the way, data collection of our study is going on and we aim to extend our sample size in the near future.

5. ACKNOWLEDGMENT

We would also like to thank Dr Khadraa Mohamed Mousa from the department of Community Health Nursing, Cairo University, Egypt for providing us with the data set.

6.CONCLUSIONS

In this study we introduce a classification model that used to predict Pneumoconiosis disease. The proposed technique is based on removing redundant attributes using PCA algorithm; finally we classify using two classifiers, neural network and Weighted KNN classifiers. It can be concluded that to improve the performance of the classification and also to have more reliable results for classification we modified the proposed system by using principle component analysis algorithm. Experiments reveal that the proposed weighted KNN classifiers technique results a considerable improvement over the artificial neural network classifier.

References

[1] Munn-Giddings, C. and R. Winter, A handbook for action research in health and social care. 2013: Routledge.
[2] Schluger, N.W. and R. Koppaka, Lung disease in a global context. A call for public health action. Annals of the American Thoracic Society, 2014. 11(3): p. 407-416.
[3] Stewart, B. and C.P. Wild, World cancer report 2014. World, 2015.
[4] Ronge, P.-S., Displacement by development? 2013, uniwien.
[5] Levine, R.A., et al., Mitral valve disease [mdash] morphology and mechanisms. Nature Reviews Cardiology, 2015.
[6] Liu, S.-j., et al., Differential gene expression associated with inflammation in peripheral blood cells of patients with pneumoconiosis. Journal of occupational health, 2016(0).
[7] Chan, C.C., et al., Using WHO's ICF Model on Service Needs of Patients with Pneumoconiosis, in Handbook of Vocational Rehabilitation and Disability Evaluation. 2015, Springer. p. 355-369.
[8] Cox, C.W. and D.A. Lynch, Medical imaging in occupational and environmental lung disease. Current opinion in pulmonary medicine, 2015. 21(2): p. 163-170.
[9] Karkhanis, V.S. and J. Joshi, Pleural effusion: diagnosis, treatment, and management. Open Access Emergency Medicine, 2012. 4(2012): p. 31-52.
[10] Ryerson, C.J. and H.R. Collard, Update on the diagnosis and classification of ILD. Current opinion in pulmonary medicine, 2013. 19(5): p. 453-459.
[11] Shlens, J., A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100, 2014.
[12] Akben, S.B. and A. Alkan, Density Weighted K-Nearest Neighbors Algorithm for Outliers in the Training Set Are So Close to the Test Element. Journal of Electrical Engineering, 2015. 3: p. 150-161.
[13] Ali, F., A. Reda, and E. Mahmoud, EEG Classification based on Machine Learning Techniques. International Journal of Computer Applications, 2015. 128(4): p. 22-27.

[14] Ali, J.B., et al., Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals. Applied Acoustics, 2015. 89: p. 16-27.