International Journal of Bio-Science and Bio-Technology
Vol.8, No.1 (2016), pp.143-152
http://dx.doi.org/10.14257/ijbsbt.2016.8.1.13

Ameliorated MLP based Approach for Identification of Lung Diseases

Ramandeep Kaur[#1] and Prince Verma[*2]

[#]*M.Tech CSE Department,* [#]*CT Group of Engg. Mgt.&Tech, Jalandhar, India*
[*]*Assistant proff.,M.Tech, CSE Department,* [#]*CT Group of Engg., Mgmt., &Tech, Jalandhar, India*
Sokhiraman89@gmail.com

Abstract

Data Mining is the process of extract information from a data set and transform it into an understandable structure. Classification is a data mining task, used to classify data among different predefined clusters. The paper proposes a new classifier utilizing MLP approach by maintaining the clusters based on similar feature. The improved MLP approach can handle noisy data, reduce complexity and increases the performance. This technique has been applied for automating medical diagnosis. This paper analyzes the lung CT-scan images for identifying and classifying them among the various lung diseases (i.e., bronchitis, emphysema, pleural effusion or normal) by emphasizing on the problematic area.

Keywords*: Data Mining, Classification, Multilayer Perceptron*

1. Introduction

Classification is a data mining process that assigns items in a group to target classes. The purpose of classification it to try to discover forecast the target class for every single case in the data. The algorithm tries to find out relationships between the attributes that will make it possible to predict the outcome. Classification is an essential data mining technique with wide applications. It is used to categorize each and every item in a set of data into one of predefined set of classes or groups. There are different types of classifier to classify the data like MLP, Naïve Bayes, and SVM *etc.*, The MLP classification algorithm is used to classify because MLP do not make any assumption relating to the underlying probability density functions and has ability to classify untrained patterns. Anita chaudhary, *et al.*, [1] in this paper discuss computed tomography (CT) in many cases are better than X-ray. And drawback arises as a result of period constraint in detecting the current of pulmonary cancer regarding on the varied diagnosis method used. Hence, a pulmonary detection system using image method can be utilized to classify the present of lung cancer in a CT- image. MATLAB are used through each procedures created. And also discuss more accurate result using segmentation and enhancement technique. Roy, *et al.*, [2] the adaptation of network weights using Particle Swarm Optimization (PSO) was proposed as to enhance the performance of ANN and classification is conducted on IRIS dataset. By merging the PSO and BP it provides optimized outcome. PSO algorithm is a global algorithm, that has a substantial capability to discover global optimistic outcome and BP algorithm has a substantial capability to discover local optimistic outcome, but global optimistic outcome is weaker. By mixing the PSO with the BP, in starting phase of searching for the optimum, the PSO is used to accelerate the training speed. Lin-Yu Tseng, *et al.*, [3] in this Paper identifying the lung diseases, computed tomography (CT) scan is wide applied in diagnose. The lung segmentation is the pre-processing step in the systems. In manually segmentation the

ISSN: 2233-7849 IJBSBT

lungs is tedious, difficult and take plenty of time for the large CT databases. And also define a Novel segmentation technique that will find the threshold value for all CT Slice in a patient.

In the past few years, medical CT Images are widely-used in medical diagnosis. Computed Tomography images might be recognized for several body tissues based on their distinct gray levels. Computed tomography, more commonly called a CT scan. This diagnostic medical test done like classic X-rays produce multiple images inside of the body. The cross-sectional of CT scan images produced variety planes of body and this output can be generate in 3-D view. CT imaging had evolved far enough in terms of speed and resolution to make it a valuable tool in the imaging of the lungs these images can be viewed on a PC or printed on film. The lungs are an imperative a part of the whole body as thousands of times each day it performs responsibility for supplying oxygen to blood whereas exhaling Carbon dioxide. Many people suffer from lung disease which occurs due to smoking, infection and biological reasons.

This paper considers four types of images *i.e.*, bronchitis, emphysema, pleural effusion, and normal. The bronchitis is a condition in which the bronchial tubes become inflamed. People who have bronchitis often have a cough that brings up mucus. Mucus is a slimy substance made by the lining of the bronchial tubes. Bronchitis also may cause wheezing (a whistling or squeaky sound when you breathe), chest pain or discomfort, a low fever, and shortness of breath. The Emphysema is a form of chronic (long-term) lung disease. People with emphysema have difficulty breathing from a limitation in blowing air out. There are multiple causes of emphysema, but smoking is by far the most common. The pleura are the membrane that lines your thoracic (chest) cavity and covers your lungs. It is a large sheet of tissue that wraps around the outside of your lungs and lines the inside of your chest cavity. There is a small space between the layers of your pleura, which contains a small amount of fluid that serves as a lubricant for the two layers of the pleural.

Classification is done with new classifier i.e. improved MLP approach that can handle noisy data and increases the accuracy. The Proposed Approach would be executed in four phases, *i.e.*, in the very first phase, the preprocessing is done by a filter, then features are extracted from preprocessed image by MAD Technique, the features are selected by applying genetic algorithm to select the top ranked features. In the finalized phase, the classifiers Multi-Layer Perceptron Neural Networks (MLP-NN) and Improved MLP-NN are used to classification of the lung diseases.

2. Material and Method

The classification of CT scan images for different Lung diseases like Bronchitis, Pleural effusion and Emphysema are considered in this work. The age of patients whose lung images are considered, ranges from 20 to 60. The images are taken from radiopaedia.org website. Four datasets are considered. The first data set contains 20 images, 5 images of each disease, second data set contains 40 images, 10 images of each disease, third data set contains 60 images, 15 images of each disease and last dataset contains 80 images, 20 images of each disease. Classification helps people to classify the items, classify data. It also helps people to contrast and compare items. The classification algorithm helps to describe the disease. It classify according to the feature of images and texture of images. With the help of new algorithm easily classification is done and also reduces time. In lungs CT scan image identification of disease is difficult but these algorithms can easily identify images within few second.

3. Proposed Approach

3.1. Preprocessing

Preprocessing is a diverse and important set of image preparation programs to improve the image in ways that increase the chances for success of the further processes. It takes an input image to produce enhanced output image. In the Proposed Approach preprocessing is performed by two filters *i.e*. Morphological Smoothening Filter and Median Filter. This filter is help to enhance the image. The median filter is utilized to remove the noise and salt while the morphological smoothening filter enhances the image by erosion and dilation. The erosion is process of removing pixel from edge of image and dilation is the reverse process.

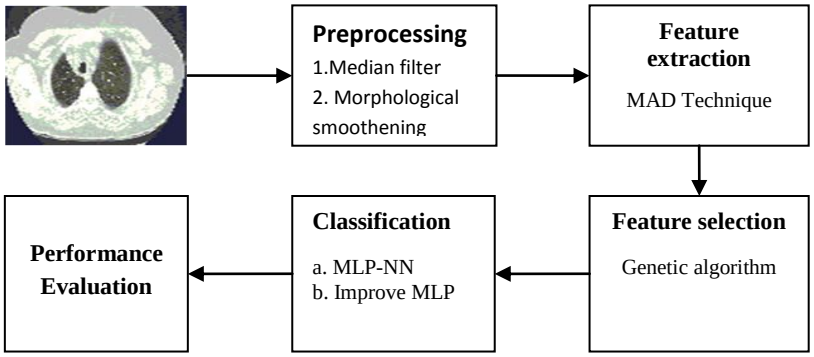

Figure 1. Block Diagram of Proposed Approach

3.2. Feature Extraction

Feature extraction is process of extracting a unique feature by including an existing feature. These features are non redundant and provides the succeed learning as well as act as a generalization approach. It requires reducing the amount of resources necessary to describe a huge set of facts. In our Proposed Approach Feature extraction is performed by MAD as it is a strong measure of the variability of a univariate sample of quantitative information. MAD can use the median for the discrepancy scores *i.e.,* healthier quality compared to the regular deviation as a method of dispersion and is far less at risk of the result of outliers compared to the regular deviation.

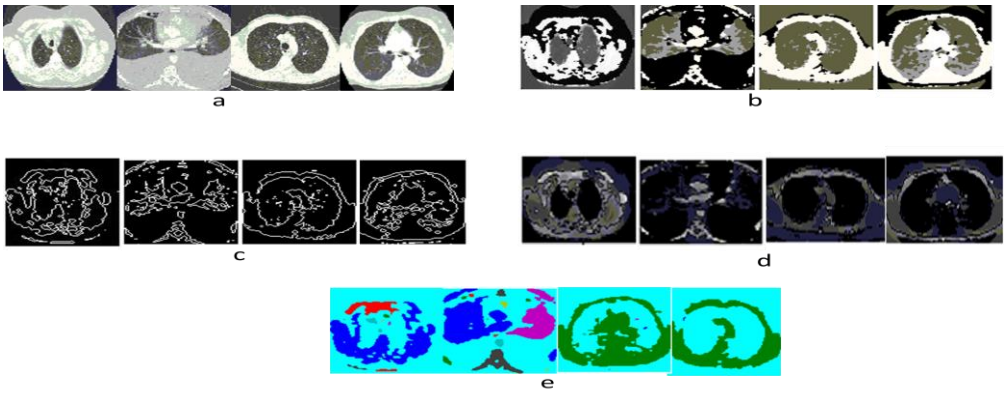

Figure 2. (a) Input image (b) Images After Application of Medium Filter, (c) Images after Application of Morphological Filter, (d) Images after Application of Mad Technique, (e) Images after Application of Genetic Algorithm

3.3 Feature Selection

Feature selection is the fully automated process that performs the selection of features in the data, out of which majority data is related to the current predictive modeling problem. This process is completely different from dimensionality reduction as it eliminates all of the features whose variance doesn't fulfill some threshold. Its selection methods are commonly used in extensions having various features and relatively number of free samples. Feature selection is also known as variable selection or attributes selection. Genetic Algorithms (GAs) are flexible heuristic finder algorithm based on the evolutionary strategies of natural selection and genetics. In this Algorithm firstly set of several answers (represented by chromosomes) named population. The population contributes to a solution that is used by another population for better solution. According to the fitness of population, the new population will probably be significantly better than the older one.

3.4 Classification

Classification is a data mining operations that assigns items in a set to target classes or categories. The purpose of classification is to get the forecast of the target class for every case in the data. The algorithm will try to find out relationships between the attributes that will ensure it is possible to forecast the outcome. In this paper classification is done using MLP-NN and the proposed Improved MLP algorithm. These algorithms are used for classification and compared for the detailed analysis in this paper further.

3.4.1 MLP-NN Algorithm:-Multilayer perception (MLP) is an information processing technique based on biological nervous systems process information, such as in brain. MLP-NN design that maps sets of input information onto a set of desirable results. A MLP consists of a number of layers of nodes in a directed graph, with every single layer totally linked with the next one. The each unit carries out the weighted number of their inputs and then passes into activation level with help of transfer function to generate their output. The approach uses the form of feed forward topology.

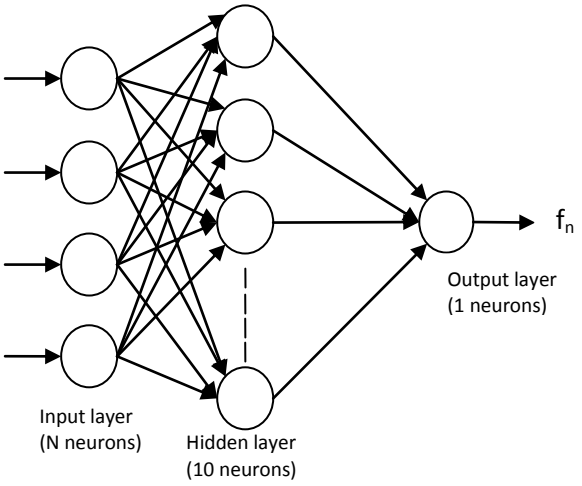

Figure 3. Block Diagram of MLP-NN

3.4.2 Proposed Improved MLP:-The MLP makes initial clusters on the randomized approach which needs to eliminate vague information. So, the improved MLP will firstly create the clusters based on the same feature using nearest neighbor. The nearest neighbor is find according to distance formula. When clustering is done then initialize all weights *i.e.*, $\vec{x}=(x_1, x_2 \ldots\ldots x_n)$ according to input. After initialize weights then calculate the sum

of all weights with there activation functions. The activation function is $a_i \leftarrow x_i$. Then find activation of all hidden layer. The processing is done all neuron of the network step by step. With the cluster initialization based on similar feature the complexity of MLP can be reduced. When vague information is removed time complexity is reduced, automatically. Improved Multilayer Perception (IMLP) has a Hidden layer between the input layer and output layer. Improve MLP is implementing by following step:-

Step1: Make a cluster based on similar feature which is selected by GA
Step2: Initialize input of all weights with small random numbers.
Step3: Calculate the weight sum of the inputs.
Step4: Calculate activation function of all hidden layer.
Step5: Output of all layers.

4. Implementation & Results

Our proposed method is implemented on lung diseases CT dataset. There are five dataset. The first data set consider 40 images, 10 images of each disease. The second dataset consist of 60 images, 15 of each disease. The third dataset consist of 80 images, 20 images of each disease. The fourth dataset consist of 100 images. The 25 images are normal lung image and each diseases of bronchitis, emphysema, pleural effusion images each comprising of 25 images respectively. The improve MLP is used for classify diseases. In improve MLP optimized value of threshold variable and training time. To evaluate the results of the MLP-NN and Improve MLP for the lung diseases image parameters used are *i.e.*, accuracy, F-measure, precision, recall and correctly classified diseases.

A. Result of MLP an Improved MLP with 40 images dataset

The output of two classification algorithms with 40 images dataset the classification accuracy is show in Figure 4. The classification accuracy is tabulated form in Table 1.

Table 1. Classification Accuracy with 40 Images Dataset

Algorithm	Accuracy (%)
MLP	73.7
Improve MLP	78.5

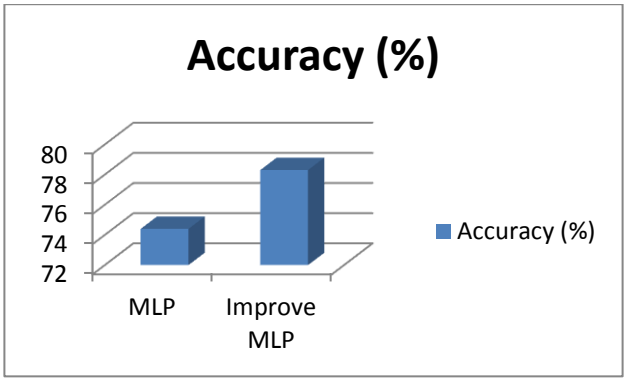

Figure 4. Classification Accuracy (in Percentage) with 40 Images Dataset

The performance measures such as precision, recall, F-Measure are presented in the Table 2 and graphical representation in the Figure 5.

Table 2. Performance Measure of Two Classifiers with 40 Images Dataset

Algorithm	Precision	Recall	F-measure
MLP	.88	.83	.81
Improve MLP	.90	.91	.91

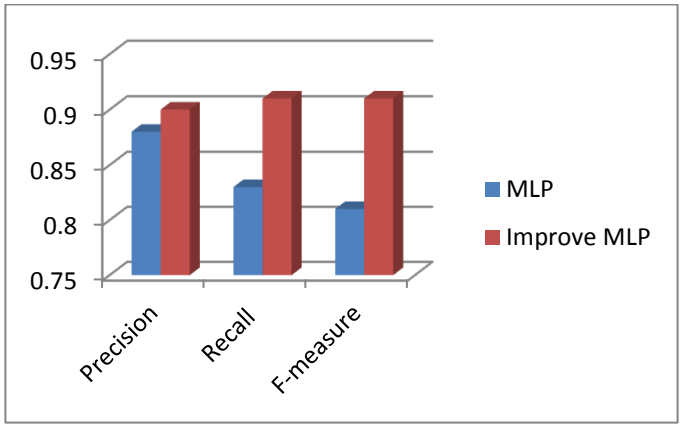

Figure 5. Performance Measurements with 40 Images Dataset

Table 3 presents the classification of datasets by their respective diseases for a given set of data.

Table 3. Classification Details Disease Wise 40 Images Dataset

Algorithm	Correctly classified as Emphysema	Correctly classified as Bronchi stasis	Correctly classified as pleural effusion	Correctly classified as normal
MLP	8	8	7	6
Improve MLP	9	8	8	7

B. Result of MLP and Improved MLP with 60 Images Dataset

The output of two classification algorithms with 60 images dataset the classification accuracy is show in Figure 6. The classification accuracy is tabulated form in Table 4.

The performance measures of 60 images dataset such as precision, recall, F-Measure are presented in the Table 5 and graphical representation in the Figure 7.

Table 4. Classification Accuracy with 60 Images Dataset

Algorithm	Accuracy (%)
MLP	74.4
Improve MLP	78.3

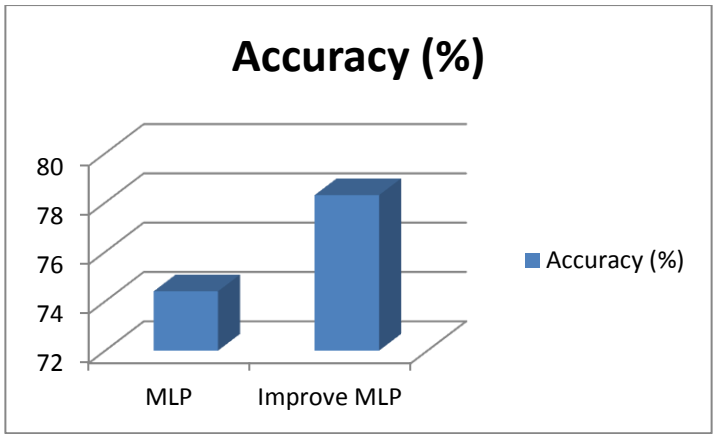

Figure 6. Classification Accuracy (in Percentage) with 60 Images Dataset

Table 5. Performance Measure of two Classifiers with 60 Images Dataset

Algorithm	Precision	Recall	F-measure
MLP	.91	.83	.83
Improve MLP	.92	.91	.91

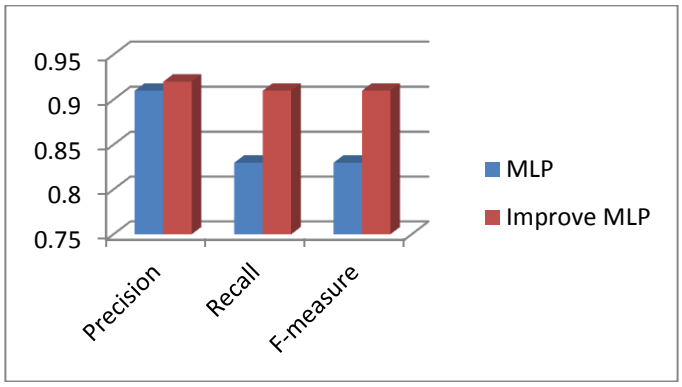

Figure 7. Performance Measurements with 60 Images Dataset

Table 6 presents the classification of 60 images datasets by their respective diseases for a given set of data.

Table 6. Classification Details Disease Wise 60 Images Dataset

Algorithm	Correctly classified as Emphysema	Correctly classified as Bronchi stasis	Correctly classified as pleural effusion	Correctly classified as normal
MLP	8	8	7	8
Improve MLP	10	11	10	9

Correctly classification of all diseases using MLP and Improve MLP on variant data set *i.e.*, 40 images, 60 images, 80 images and 100 images data set.

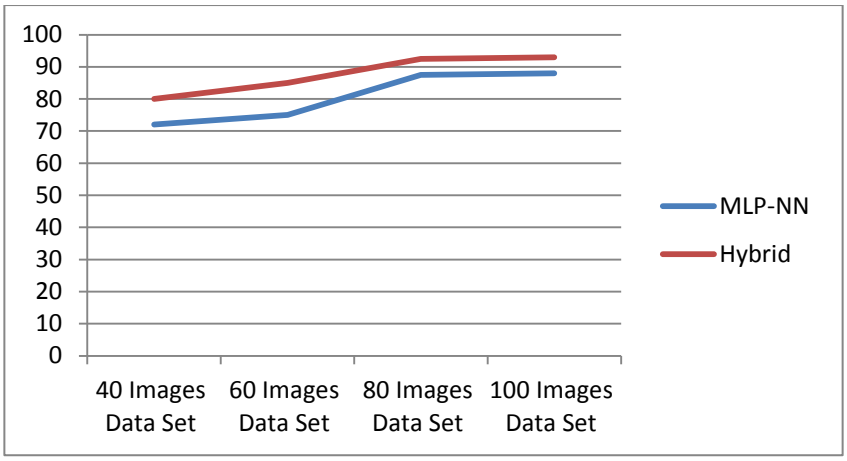

Figure 8. Correctly Classification of All Disease (in %age)

The correctly classification of all diseases is 72% in MLP and 80% in Improve MLP in 40 images dataset, 75% in MLP and 85% in Improve MLP in 60 images dataset, 87.5% in MLP and 92.5% in Improve MLP in 80 images dataset and 88 % in MLP and 93% in Improve MLP in 100 images dataset.

5. Conclusion

In this work feature extraction is performed by MAD and feature selection is performed by genetic algorithm that chooses the best ranked features. For classification, MLP and New proposed Improve MLP re used. The proposed approach effectively works for the detection of lung diseases with high precision, Recall, F-measure and accuracy as shown in the results. With very small number of images(*i.e.*, 20 Images in database), both approaches performs better in identification of correct disease *i.e.*, traditional MLP gives 72% and Improved MLP gives 80%. With further increase of the number of images in dataset, the accurate classification of diseases increases, as with increased number of images there is better training for the features extracted. It can be also being seen that, the proposed approach performs better training for the features extracted than traditional MLP approach. The work can further be extended by including more feature extraction or/and feature selection methods for classifying more lung disease categories like asthma, chronic cough, lung cancer, pleurisy and influenza.

References

[1] A. Chaudhary and S. S. Singh, "Lung cancer detection on CT images by using image processing", In Computing Sciences (ICCS), 2012 International Conference on IEEE, **(2012)**, pp. 142-146.
[2] A. Roy, D. Dutta and K. Chaudhary, "Training artificial neural network using particle swarm optimization algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 3, **(2013)**, pp. 430-434.
[3] L.-Y. Tseng and L.-C. Huang, "An adaptive thresholding method for automatic lung segmentation in CT images", In AFRICON, 2009. AFRICON'09, IEEE, **(2009)**, pp. 1-5.
[4] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, **(2000)**.
[5] D. L. Abd AL-Nabi and S. S. Ahmed "Survey on Classification Algorithms for Data Mining, (Comparison and Evaluation)", Computer Engineering and Intelligent Systems, vol. 4, no. 8, **(2013)**.
[6] V. Vaithiyanathan, K. Rajeswari, K. Tajane and R. Pitale, "comparison of different classification technique using different datasets", International Journal of Advances in Engineering & Technology, ©IJAET, **(2013)** May.
[7] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Elsevier, ISBN1558609016, **(2006)**.
[8] M. Kantardzic, "Data Mining - Concepts, Models, Methods, and Algorithms", IEEE Press, Wiley-Interscience, ISBN 0-471-22852-4, **(2003)**.
[9] R. Shojaii, J. Alirezaie and P. Babyn, "Automatic lung segmentation in CT images using watershed transform", In Image Processing, ICIP, IEEE International Conference on, vol. 2, **(2005)**, pp. II-1270.

[10] K. Devaki and V. MuraliBhaskaran, "Study of computed tomography images of the lungs: A survey", In Recent Trends in Information Technology (ICRTIT), International Conference on IEEE, **(2011)**, pp. 837-842.
[11] N. Mesanovic, S. Mujagic, H. Huseinagic and S. Kamenjakovic, "Application of lung segmentation algorithm to disease quantification from CT images", In System Engineering and Technology (ICSET), International Conference on IEEE, **(2012)**, pp. 1-7.
[12] Y. Qian and W. Guirong, "Lung nodule segmentation using EM algorithm", In Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2014 Sixth International Conference on IEEE, vol. 1, **(2014)**, pp. 20-23.
[13] Q. Gao, S. J. Wang, D. Zhao and J. Liu, "Accurate lung segmentation for X-ray CT images", In Natural Computation, ICNC Third International Conference on, IEEE, vol. 2, **(2007)**, pp. 275-279.
[14] Y. Itai, H. Kim, S. Ishikawa, S. Katsuragawa, T. Ishida, K. Nakamura and A. Yamamoto, "Automatic segmentation of lung areas based on SNAKES and extraction of abnormal areas", In Tools with Artificial Intelligence, ICTAI, 17th IEEE International Conference on, **(2005)**, pp. 5-pp.
[15] X. W. V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand and D. Steinberg, "Top 10 algorithms in data mining", Springer-Verlag London Limited, **(2007)** December 4.
[16] A. Desai And S. Rai, "Analysis of Machine Learning Algorithms using WEKA", International Conference & Workshop on Recent Trends in Technology, (TCET) Proceedings published in International Journal of Computer Applications® (IJCA), **(2012)**.
[17] D. Kumar and Suman, "Performance Analysis of Various Data Mining Algorithms: A Review", International Journal of Computer Applications (0975 – 8887), vol. 32, no. 6, **(2011)** October.
[18] N. J. Chatap and A. K. Shrivastava, "A Survey on Various Classification Techniques for Medical Image Data", International Journal of Computer Applications, vol. 97, no. 15, **(2014)** July.
[19] G. Kesavaraj And S. Sukumaran, "A Comparison Study on Performance Analysis of Data Mining Algorithms in Classification of Local Area News Dataset using WEKA Tool", International Journal Of Engineering Sciences &Research Technology, vol. 2, no. 10, **(2013)** October.
[20] M. S. Chen, J. Han and P. Yu, "Data mining: an overview from a database perspective", IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, **(1996)**, pp. 866-883.
[21] S. J. Lee and K. Siau, "A review of data mining techniques", Industrial Management and Data Systems, University of Nebraska-Lincoln Press, USA, **(2001)**, pp. 41-46.
[22] H. S. Kim, H.-S. Yoon, K. N. Trung and G. S. Lee, "Automatic lung segmentation in Ct images using anisotropic diffusion and morphology operation", In Computer and Information Technology, CIT, 7th IEEE International Conference on IEEE, **(2007)**, pp. 557-561.
[23] S. W. Purnami, A. Embong, J. M. Zain and S. P. Rahayu, "A New Smooth Support Vector Machine and Its Appli- cations in Diabetes Disease Diagnosis", Journal of Computer Science, vol. 5, no. 12, pp. 1006-1011.
[24] R.-M. Ştefan, "A Comparison of Data Classification Methods", Procedia Economics and Finance, vol. 3, **(2012)**, pp. 420-425.
[25] G. I. Salama, M. Abdelhalim and M. A. E. Zeid, Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC), vol. 32, no. 569, **(2012)**, pp. 2.

Authors

Ramandeep Kaur, He received the B. Tech degree in Computer Science from CTIEMT, Jalandhar (Pb), India in 2013 and M.Tech degree in Computer Science in 2015 from CTIEMT, Jalandhar (Pb), India. Currently she is doing research work in Data Mining. Her research focuses on Data Mining and its algorithm, digital image processing on documents and database.

Prince Verma, He received the B.Tech degree in Computer Science from MIMIT, Malout (Pb), India in 2008 and M.Tech degree in Computer Science in 2013 from DAVIET, Jalandhar (Pb), India. Currently he is Assistant Professor in Computer Science Department of CTIEMT, Jalandhar (Pb), India. His research focuses on Data Mining, Algorithm optimization techniques