Review on Mesothelioma Diagnosis

Prof. Vaishali Surjuse, Anish Khobragade, Ajeet Sah, Shubham Soneji

Department of Computer Science and Engineering, KDK College
Rd Opposite Telephone Exchange, Nandanvan Nagpur-440009, Maharashtra ,India

Abstract-* Asbestos is a carcinogenic substance, and threatens human health. Malignant Mesothelioma disease is one of the most dangerous kind of cancer caused by asbestos mineral. The most common symptom of the disease, progressive shortness of breath and constant pain. Early treatment and diagnosis are necessary. Otherwise, the disease can lead people to die in a short period of time. In this paper, different types of artificial intelligence methods are compared for effective Malignant Mesothelioma's diseases classification. Support Vector Machine, Neural Network and Decision Tree methods are selected in terms of regular machine learning concept. Additionally, Bagging and Adaboost re-sampling within ensemble learning terminology is also adapted. Totally 324 Malignant Mesothelioma data which consists of 34 features is used in this study. K-fold cross-validation technique is performed to compute the performance of the algorithms with different K values. 100% classification accuracies are obtained from three tested methods; Support Vector Machine, Decision Tree and Bagging. Additionally, the process time of methods are measured in case of using method in lots of data. In this sense, methods are evaluated based on accuracy and time complexity. The results of this paper are also compared with previous studies using same Malignant Mesothelioma's dataset.

***Keywords—* Malignant Mesothelioma, Support Vector Machine, Decision Tree, Neural Network, Ensemble Learning**

1. INTRODUCTION

Malignant Mesothelioma (MM) is one of the cancer type. It appears on the thin layer of tissue and rapidly affects to various internal organs [1]. Lining parts of lungs and the chest wall is the most infected parts and organs in cases [2] [3]. Different symptoms such as difficulties in breathing, affliction in chest wall, cough, bloated abdomen, exhausted morality, extremely loss in weight etc. can be seen. Disease advances rapidly while the symptoms appear slowly [4].

1) The asbestos mineral plays important role on mesothelioma disease. According to medical report, 80% of disease is caused by the mineral [3]. More exposure to mineral increase the risk of developing the disease. In this sense, people living in industrialized countries encounters more than small towns. More specifically, disease is mostly seen in miners and produces who deals with the asbestos mineral. Normally, incubation stage of the disease is around 40 years for [3]. The late awareness of Malignant Mesothelioma disease has made it impossible to diagnosis.

2) The diagnosis are performed by observation of the X-ray images of chest and the scan findings of computed tomography. In both techniques, doctors mainly examine the fluid produced by the cancer in results or the tissue obtained by biopsy [4].

3) Addition to regular techniques, computerized methods are also utilized in few studies. Currently, computer based diagnosis systems, which named as Computer Assisted Systems (CAS) become more popular due to high accurate, consistent and efficient results [5]. CAS mainly employs the artificial intelligence methods such as Support Vector Machine (SVM), Decision Tree (DT), Neural Networks (NN) etc. on the stored numerical data. Similar to various medical application, MM disease diagnose is, basically, also a significant classification problem. Methods might conclude different results according to arranged data [6]. In this sense, in order to define the useful method for the corresponding data, several artificial algorithms need to be tested. *5)* In the study, the classification of the data for the Malignant Mesothelioma disease is performed and test results is compared. This study also provides a decision support system, which contributes to the doctors in their diagnosis decisions. Paper is organized as follow; current studies over MM disease diagnose are presented in Section 2. Methods used in testing are briefly explained in Section 3 with data information. Results and explanations are given in Section 4. Paper is concluded with future works and final decisions as last chapter.

2. LITERATURE REVIEW

Visual investigation technique on the diagnosis of medical images is a time-consuming and subjective procedure. Experiences of doctors play effective roles on decision step. In this sense, using the image processing algorithms and artificial intelligent methods prevent diagnoses from different decisions of doctors such as in computed tomography analyses. Computer based technique presented in [7] easily identifies the pleural contours and detects pleural thickenings with two steps. Firstly, they detect the thorax and then remove the air and trachea. In both steps, they implemented 3D morphological operations. According to paper, image retrieval system over MM diagnose is a promising method to detect the disease.

Another study published by Chen et al. [8] explains the implementation of the random walk-based segmentation [9] method. They used mesothelioma computed tomography image datasets and aimed to establish an automatic segmentation. They observed the progression of the disease by volumetric assessments to decide the treatments. Similar to this approach, Onama et al. used 3D version of random walk-based segmentation method on PET images [10]. They aimed to increase success rates for the detection of Lung Tumor.

Er et al. used numerical dataset instead of images. They adapted probabilistic neural networks (PNN) for using in the diagnosis of MM disease. They compared the results to multilayer and learning vector quantization neural networks. They reported in [11] that PNN is evaluated as best classifier with 96.30% accuracy.

A different approach to MM disease diagnose is presented in [12] by K. Chaisaowong et al. They observed the contours of the pleura form in healthy and patient cases. According to comparison of tracing, they detected the thickenings. In this meaning, they formed a tissue-specific segmentation by implementation of the 3D Gibbs-Markov random field (GMRF) [13]. It is adopted to distinguish thickenings from thoracic tissue. Then, morphometric analyses and volumetric assessments are performed to 3D modeling. According to results of the paper, authors assure that the automated approach can help physicians to diagnose pleural mesothelioma in its early stage.

3. METHODOLOGY

Currently, several machine learning algorithms are already utilized for mesothelioma dataset. However, classification results might be increased with other methods. Hence, in this study, different machine learning methods tested on mesothelioma dataset. Methods are selected due to not applied on dataset before. Hereby, in case of more accurate results, method can be used for advanced diagnosis. Five fundamental classification methods are tested in this study. Methods are categorized into two titles: a.) Machinelearning and b.) Ensemble-learning methods. The brief descriptions of the used methods and parameter arrangement are separately explained in following subsections.

Machine Learning Methods

A great deal of machine learning algorithms and their variation with differently selected parameters are stated in literature by means of classification. Majority of them are highly modified for biomedical datasets. Accurate results provides more informative and meaningful diagnosis. In that meaning, three fundamental methods of machine learning is adapted for mesothelioma dataset.

Support Vector Machine (SVM)

SVM is one of the prominent classification algorithms that can be used large-scale datasets and provides more accurate results. It can be achieved by even small size trainsets with the help of well-fitted cost function in kernel space [15].

SVM uses the core idea of kernel based learning. It aims to separate data in high dimensional feature space with a kernel function. SVM creates a decision surface between the samples of different classes over optimal hyperplane. SVM provides binary classification of two-class datasets. "One against one" or "one against all" are the most popular strategies in literature. Each strategy has own advantages and disadvantages mentioned in [16]. In our study, “one against one” strategy is used owing to 2 classes’ presence in datasets.

In order to define well-fitted settings of SVM for mesothelioma dataset, different kernels, penalty and kernel parameters are tested at the initial part of study. Table 2 indicates the all parameter test results.

a) Decision Tree (DT)

Decision Tree is known as rule based machine-learning method [17]. Principally, it works based on tree terminology. The path from root to leaf presents classification rules. The roots represent the most informative features and the leaves indicate the labels. Information gain (IG) is the rule defining criteria. The most widely used algorithms are entropy, twoing, and Gini to calculate the IG.

Decision Tree is easy to implement. Additionally, interpretation of the classification is much easier than other methods. It is useful for some regression problems. However, DT results low performance on large scale datasets with few training samples compare to SVM [18]. Pruning process is another obstacle point to avoid overfitting. According the results of preliminary studies on parameter settings, DT model is modified with pruning functionality and Gini Diversity Index for IG.

a) Neural Networks - Multi Layer Perceptron (MLP)

Multi-Layer Perceptron (MLP) is the advanced version of NN [19]. Minimum two layers connected with two functions should be utilized. Different parameters and functions are tested at initial studies. According to results, MLP network is arranged as the weight and bias are fixed with 0.8 and 1, respectively.

Ensemble Learning Methods

Ensemble learning is emerged from the principles of machine-learning concepts. The key point behind the ensemble is the proper combination of several machine learning algorithms. Not only one learner as in regular methods, multi learners gather in decision step for ensemble methods, therefore it gives more success. Machine-learning classifiers such as DT, KNN etc. is named as base learner.

Mainly two ensemble models having the same base learner (Decision Tree - DT) combinations but different sample selection strategies are evaluated in this study. Majority voting is used to define final decision of base learners.

a) Bagging with DTs

Bagging, in other words bootstrap aggregation, is a way for improving the classification by the aid of well-formed train samples. It is also cited as re-sampling process in literature [20]. The idea of bagging is to distort the dataset by resampling, and to train weak learners using re-sampled trainsets. The distortion of the samples is made by a voting process of weight parameters. The weights of the samples are fixed equally; therefore, trainsets are randomly selected. Consequently, different samples are used in trainset iteratively. It provides more diversity in the samples' distribution. The average of the each decision of base learners determines the final decision. More information can be found in [20].

a) Adaptive Boosting (Adaboost) with DTs

Boosting is another technique in re-sampling process similar to bootstrap. The difference is that bootstrap ignores the weight values of the samples and re-samples randomly, however boosting technique defines different weights for each samples after first iteration. Then, the probabilities of misclassified samples are boosted for the second step, and subsequent classifiers are trained. Likewise, other steps are sustained with different weight parameters. Readers are referred to an essential guide [21] for boosting theorem.

Adaptive boosting is mainly outperforms other regular boosting techniques and more robust for over-fitting problem. However, it is still easily affected by noise and outliers owing to iteratively arranging process for weights.

Dataset

Dataset is obtained from UCI dataset repository [22]. It includes the patient's records obtained from Dicle University, Faculty of Medicine. 324 MM patient data were recorded and tested by aforementioned AI methods. These data were also investigated by Orhan Er et al. in terms of PNN as mentioned in Section 2 [12].

In the dataset, 324 samples individually have 34 features with multivariate variables. There is no “unidentified” or “missing value” presence in dataset. Details of data and features can be found in [12]. Decision labels provided by doctors as sick and healthy (2 classes).

4. RESULT & DISCUSSION

Classification of mesothelioma dataset is performed by three regular machine learning and two ensemble learning methods. DT, SVM and NN methods are selected within the regular machine learning concept. On the other hand, Bagging and Adaboost with same weak learners (DT) is performed as ensemble idea. Accuracy and computational time are considered as the evolution metrics. Computational time is recorded to estimate efficiency of method for big data problems due to so many patients suffering from MM disease. In case of future studies with more patient record, time complexity become more important factor according to including 34 features besides plenty of patients.

Only 10 Fold Cross validation tests are measured in terms of computational time. Less computational time and high accuracy rate are preferred to indicate the best algorithm. Over all results are presented in Table 1

	DT	SVM (Linear)	MLP	Bagging	Adaboost
10-Fold	100	100	96,87	100	70,54
5-Fold	100	100	95,82	100	65,35
2-Fold	100	100	94,44	100	68,82
Time	0,019	0,095	13,89	17,52	0,25

Table 1 : overall results of methods

According to Table 1, simple DT and SVM as regular machine learning idea and DT with Bagging in terms of ensemble method outperform over other methods with common 100% accuracy rates. Differently formed train sets (2, 5 and 10 Fold) has no effect in general. However, another ensemble method, Adaboost using same form of DT as base learner but different sample selection strategy as weighed re-sampling, stay far behind over all methods. In this sense, randomly selection of train samples is more effective strategy in the detection of mesothelioma. Selection of sample with weight parameter is useless due to lots of features (34 features) using in classification. However, Bagging needs more computational time because of irregular sample selection process. In that meaning, Bagging is not preferred method when compare to DT and SVM because of the same accuracy rates.

One of the prominent Kernel based method, SVM, is tested with different kernels and parameters. Obtained the highest results of each kernel with different parameters are individually registered in Table 2. Linear kernel gives the best result with 100% in all *K* values. RBF (radial basis function) outcome is depended on training size. It resulted 100% accuracy rate with more training samples, but success is decreased when train set reduced. Besides the inconsistent results of RBF, it includes exponentially operations, thus, needs more time to classify big data. To avoid that time consuming process, Linear SVM might be utilized in practice owing to simplicity of algorithm and less time complexity. Polynomial, quadratic and MLP (MultiLayer Perceptron) kernels generally concluded with 97%, 88% and 90% respectively. These kernels are also directly related with training sample size. Addition to low accuracies, computational time analyses of kernels are not too far ahead from linear kernel. Therefore, SVM should be utilized with Linear Kernel to classify mesothelioma dataset. Results emphasize that it might give better results with big data over other methods.

As a final method, MLP in Neural Network terminology is adapted. Normally, MLP gives higher accuracies on nonlinear classification problems, but deals with all samples in dataset. In that meaning, algorithm success might be decreased easily by outliers and needs more computational time as it is emphasized in Table 1. Dataset has 34 features over 324 observation which means 34 dimension data. In that case, MLP is resulted with 97% accuracy rate owing to complexity of data set. On the other hand, SVM focus on the samples near support vectors. Therefore, SVM surpasses MLP due to less complexity and using pre-arranged data.

Kernels	Polynomial	Quadratic	MLP	Linear	RBF
10-Fold	97,72	88,98	90,93	100	100
5-Fold	97,18	88,75	89,21	100	99,84
2-Fold	92,40	84,01	86,11	100	99,07
Time	0,186	0,385	0,089	0,095	0,286

5. CONCLUSION

In this study, different machine and ensemble learning methods are tested on the detection of mesothelioma disease. In that meaning, a prevalent dataset provided by Orhan Er et al. [8] is utilized to measure the methods.

Orhan Er et al. published a study about the classification of their dataset with PNN before. They reported 96% success with 3Fold cross validation. In this study, we also perform a MLP network having 0.8 weight and 1 bias parameters and obtained same results. This indicates the testing methodology is similar and analogous. In that meaning, other obtained results express consistent output.

DT and SVM as regular machine learning, and Bagging as ensemble learning are highly compatible algorithms for mesothelioma dataset considering to Table 1. Methods successfully provide 100% accuracy rate in classification. However, linear kernel SVM and DT are simpler algorithm and require less computational time compare to Bagging. In this sense, Bagging is not preferably. Rule Based algorithm, DT, fails on big data problem according to report [14]. Therefore, it is also useless in practice owing to numerous patient suffering from Mesothelioma disease. In order to generalize the results, more record is necessary. In that condition, DT might give misleading diagnose. As a result, Linear SVM might be better to utilize in practice due to abovementioned results and reasons.

As future works, abovementioned methods will be tested on more obtained data in classification. Then, more generic diagnose system can be improved.

6.REFERENCES

[1] Malignant Mesothelioma, Retrieved 3 May 2016, http://www.cancer.gov/types/mesothelioma

[2] General Information About Malignant Mesothelioma, Retrieved 3 May 2016, http://www.cancer.gov/types/mesothelioma/patient/mesothelioma-treatment-pdq

[3] B.M. Robinson, “Malignant pleural mesothelioma: an epidemiological perspective”, Annals of cardiothoracic surgery vol. 1 (4), 2012.

[4] S. Kondola, D. Manners, A.K. Nowak, "Malignant pleural mesothelioma: an update on diagnosis and treatment options", Therapeutic Advances in Respiratory Disease, 2016.

[5] Delp, S. L., Loan, J. P., Robinson, C. B., Wong, A. Y., & Stulberg, S. D. (1997). U.S. Patent No. 5,682,886. Washington, DC: U.S. Patent and Trademark Office.

[6] H. Kadoz, S. Ozsen, A. Arslan, and S. Gunes, “Medical application of information gain based artificial immune recognition system (AIRS): diagnosis of thyroid disease”, Expert Syst Appl vol. 36 (2), 2008.

[7] J. Lerdsinmongkol, K. Chaisaowong, S. Roongruangsorakarn, T. Kraus, and T. Aach, “Efficient Application of 3D Morphological Operations in the Framework of a Computer-Assisted Diagnosis System”, 9th International Conference on Signal Processing, pp. 857-860, 2008.

[8] M. Chen, E. Helm, N. Joshi, S.M. Brady, "Random walk-based automated segmentation for the prognosis of malignant pleural mesothelioma", IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1978-1981, 2011.

[9] Grady, L. (2006). Random walks for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 28(11), 1768-1783.

[10] Onoma, D. P., Ruan, S., Gardin, I., Monnehan, G. A., Modzelewski, R., & Vera, P. (2012, May). 3D random walk based segmentation for lung tumor delineation in PET imaging. In 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI) (pp. 1260-1263). IEEE.

[11] O. Er, A.C. Tanrikulu, A. Abakay, and F. Temurtas, "An approach based on probabilistic neural network for diagnosis of Mesothelioma's disease", Computers & Electrical Engineering, vol. 38(1), pp. 75-81, 2012.

[12] K. Chaisaowong, C. Akkawutvanich, C. Wilkmann, and T. Kraus, "A fully automatic probabilistic 3D approach for the detection and assessment of pleural thickenings from CT data", Computational Intelligence in Medical Imaging (CIMI), IEEE Fourth International Workshop on, pp. 14-21, 2013.

[13] Schroder, M., Rehrauer, H., Seidel, K., & Datcu, M. (1998). Spatial information retrieval from remote-sensing images. II. Gibbs-Markov random fields. IEEE Transactions on geoscience and remote sensing, 36(5), 1446-1455.

[14] E. Lotfi, A. Keshavarz, "Gene expression microarray classification using PCA–BEL", Computers in Biology and Medicine, vol. 54, pp. 180–187, 2014. [15] B. Scholkopf, A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT press, 2001.

[16] J. Milgram, M. Cheriet, R. Sabourin, one against one or one against all: Which one is better for handwriting recognition with svms in: Tenth International Workshop on Frontiers in Handwriting Recognition, Suvisoft.

[17] Quinlan, J. Ross. "Induction of decision trees." Machine learning 1.1 (1986): 81-106.

[18] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont, CA, 1984.

[19] Hagan, Martin T., Howard B. Demuth, and Mark H. Beale. Neural network design. Boston: Pws Pub., 1996.

[20] L. Breiman, Random forests, Machine learning 45 (2001) 5– 32.

[21] R. Rojas, Adaboost and the super bowl of classifiers a tutorial introduction to adaptive boosting (2009).

[22] UCI Machine Learning Repository, Mesothelioma Disease Data Set, Retrieved 3 May 2016, http://archive.ics.uci.edu/ml/machine-learning-databases/00351/