International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 01, 2020
ISSN: 1475-7192

A Survey of Existing Approaches for Computer Aided Diagnosis of Cancer

N.V.S.S.K. Suman and K. Senthil Kumar

Abstract--- *Cancer is a deadly disease in which abnormal cells divide uncontrollably and destroy body tissue. It has various diagnosis techniques which are limited to only to some types of cancer. These diagnosis techniques have limited efficiency and do not guarantee the patient's life. We plan diagnosis of this deadly disease at the initial stage itself. We take the patient's medical records and gene data and look for cancer causing mutations(driver) and separate the other mutations (passengers).Thus we classify the gene and detect the mutations so that any patient with harmful mutations are potentially risk patients and thus are given treatment.*

Keywords--- *Digital Pathology, Computer Aided Diagnosis (CAD), Artificial Intelligence (AI), Tumor, Staging and Grading, Histopathology, Gene Expression, Fuzzy, Neural Networks (FNN), Support Vector Machines (SVM), Adaptive Neuro- Fuzzy Interface System (ANFIS).*

I. INTRODUCTION

Cancer is a deadly diseases which involves cells growing abnormally and can spread to different parts of the body. These contrast with benign tumors, which don't spread to other parts of the body. Signs and symptoms could be a lump, abnormal bleeding, prolonged cough, unexplained weight loss and a change in bowel movements while these symptoms may indicate cancer, they may have other causes too. More than 100 types of cancers affect humans.

Current technologies

There are various techniques which have been explored for performing cancer diagnosis starting from the Support Vector Machines, Convolutional Neural Networks and Artificial Neural Networks. These techniques are being discussed below

Cancer Classification Using Expressions of genes using SVM

This techniques finds the smallest collection of genes that assures highly accurate classification various types of cancers from data with the help of supervised machine learning algorithms. The need to find the gene subsets is :

1) It decreases the burden of complex computation and "noise" arising from irrelevant genes. The example in this paper and that is to find out the minimum gene subsets helps in the extraction of simple rules which yields the accurate result without the help of any classifiers.

2) Only a small amount of genes can be considered for test instead of hundreds of genes which can drastically reduce the cost for cancer testing.

3) It requires further investigation upon these genes and their biological relationship for furthur development. The simple and effective method specified in this paper is a two step process. Firstly few genes are chosen based on a

N.V.S.S.K. Suman, SRM Institute of Science and Technology.
K. Senthil Kumar, Assistant Professor, SRM IST.

DOI: 10.37200/IJPR/V24I1/PR200296
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

ranking scheme. This is followed by the classification of all the possible combinations of the important genes to be tested using a good classifier. For three "small" data sets with two, three, and four cancer (sub) types, the approach has high accuracy w. But for a "large" and "complex" data set with 14 cancer types, the problem as a whole is divided into a group of binary classification problems and this 2-step approach is applied to each of the binary classification problems. In general, the method reduces the number of genes significantly in requirement for highly reliable diagnosis.

Method

Step 1: Gene Importance

Ranking a) T-test gene I has the following T-test score and is defined as : Eq: 1

$$TS_i = max\left\{\left|\frac{\overline{x}_{ik} - \overline{x}_i}{m_k s_i}\right|, k = 1, 2, \ldots K\right\},$$

$$\overline{x}_{ik} = \sum_{j \in C_k} \overline{x}_{ij}/n_k,$$

$$\overline{x}_i = \sum_{j=1}^{n} x_{ij}/n,$$

$$s_i^2 = \frac{1}{n-K}\sum_k \sum_{j \in C_k} (x_{ij} - \overline{x}_{ik})^2,$$

$$m_k = \sqrt{\frac{1}{n_k} - \frac{1}{n}}.$$

Step 2: Using SVM and finding the Minimum Gene

Results (data sets used) are obtained with the help of the following datasets:

a) Lymphoma dataset. b) Liver cancer dataset. c) SRBCT dataset. d) GCM data.

The results obtained were convincible and had a good efficiency. The SVM is a good classifier to separate the cancer causing genes from the other genes.

b) Selection of a Gene and Classification of Cancer using a FNN.

Classification of Cancer based on tiny gene expressions is a very effective method. In this paper, a t-test is used to filter the genes from a few hundreds of genes. The data set is taken and FNN classification technique is applied. This Fuzzy Neural Network combines the self- generating gene optimizes the parameters and applies simplification based on some predetermined rules.

Three datasets are chosen and Fuzzy Neural Network classification technique is applied. The data sets are:

- SRBCT Dataset (4 sub types)

DOI: 10.37200/IJPR/V24I1/PR200296
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

- Lymphomia Dataset (3 sub types)
- Liver Cancer Dataset (2 classes).

The results conclude that FNN has a very good accuracy compared to other techniques provided that the genes are few in number. The FNN concludes and describes the need to classify the genes and separate the important genes that cause cancer. It helps research people to focus on fewer genes for further examination.

The paper is structured as follows. , firstly ranking of gene with the help of t-test, followed by the review of the structure and the algorithm of the FNN**.** Finally**,** the FNN is used classify the three data sets**,** i.e ; the lymphoma data set, the SRBCT data set**,** and the liver cancer data set. All these datasets are microarray datasets.

Method

1) Gene importance ranking.

2) The FNN Structure and flow chart analysis.

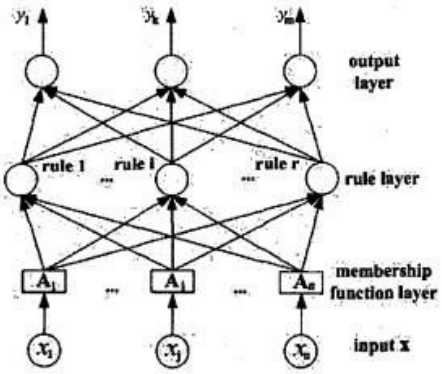

Fig. 1: FNN Structure

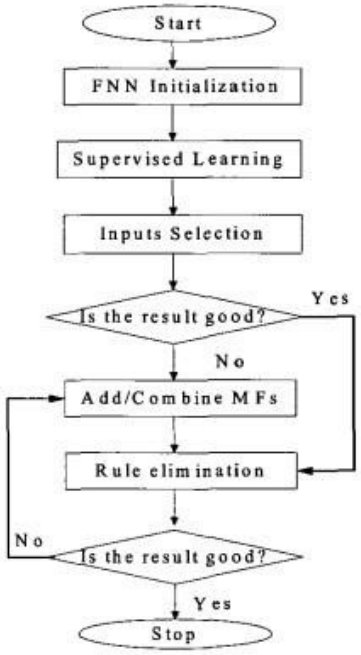

Fig. 2: FNN Flow chart

DOI: 10.37200/IJPR/V24I1/PR200296
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

FNN Training

Training rule for the FNN

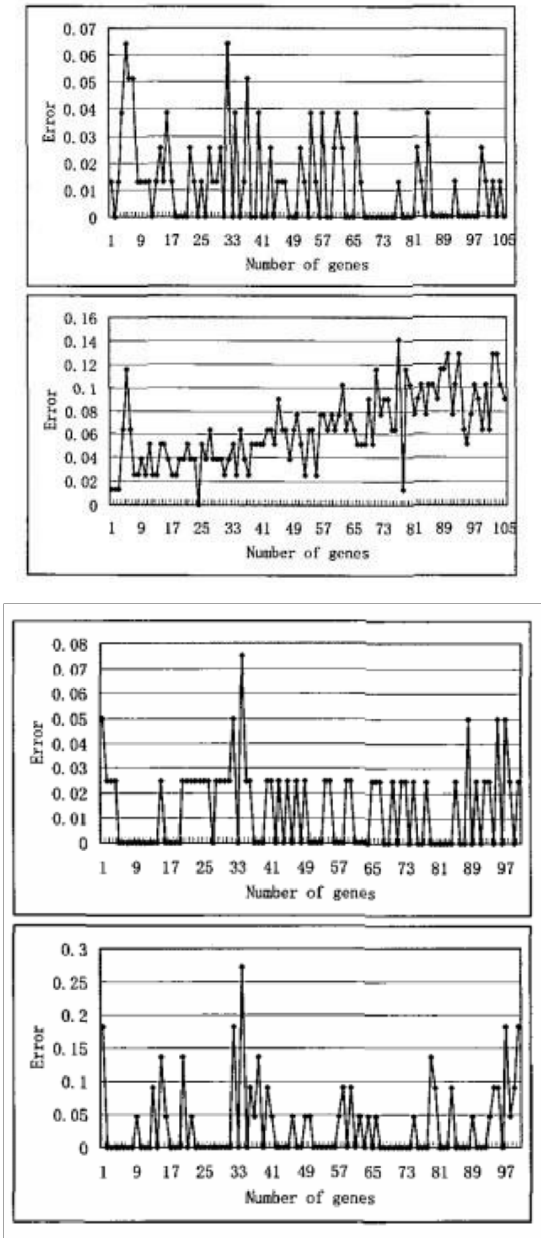

Fig. 3: Lymphoma dataset

$$y_l^{\,i}(k+1) = y_l^{\,i}(k) - \eta \frac{\partial \varepsilon_l}{\partial y_l^{\,i}}$$

The learning rule

$$\omega_l^{\,i}(k+1) = \omega_l^{\,i}(k) - \eta \frac{\partial \varepsilon_l}{\partial \omega_l^{\,i}}$$

$$A_q^{\,i}(k+1) = A_q^{\,i}(k) - \eta \frac{\partial \varepsilon_l}{\partial A_q^{\,i}}$$

DOI: 10.37200/IJPR/V24I1/PR200296
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

Error function

$$\varepsilon_l = \frac{1}{2} \times (y_l - y_{dl})^2$$

Results

1) Lymphoma dataset The training result is in the upper section and the testing results is in the lower section.

2) SRBCT dataset

The training result is in the upper section and the testing results is in the lower section.

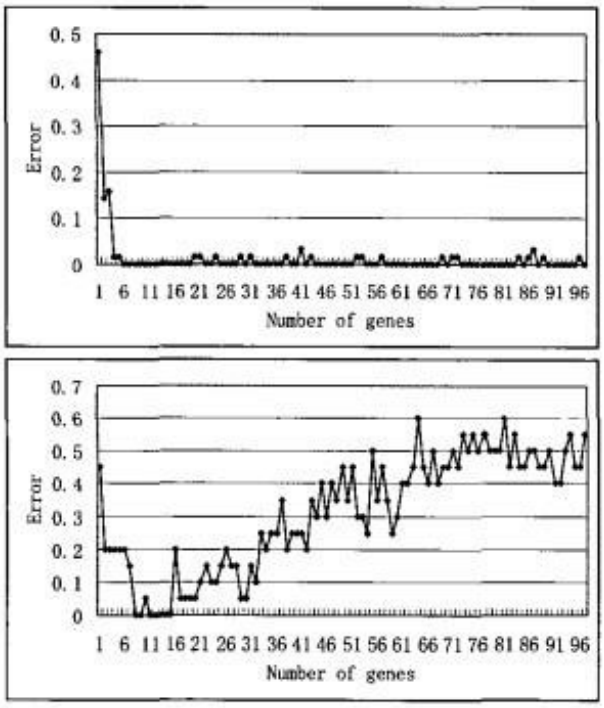

Fig. 4: SRBCT dataset

3) Liver cancer dataset

The training result is in the upper section and the testing results is in the lower section.

The paper provides conclusive evidence that FNN has better performance compared to traditional methods like nearest shrunken centroids. FNN also shows that fewer number of important genes yields a better result. FNN uses only 5 genes in lymphomia data set.

Evolutionary algorithm reported by Deutsch produces 100% accuracy with 12 genes whereas FNN requires only 8 genes to obtain the same accuracy.

The number of genes were reduced to 24 in the liver cancer dataset without much impact on the accuracy.

The FNN helps biological researchers in the because it uses less number of genes and produces results with great accuracy.

DOI: 10.37200/IJPR/V24I1/PR200296
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

Thus the biological researchers can focus on small no of genes and find relationship among them.

Computer aided diagnosis of lung cancer

Most important algorithms used in the computer aided diagnosis of cancer are considered and mentioned in this paper. The ROC characteristics are shown for these algorithms. These algorithms are compared to get a good understanding of their pros and cons.

Classification algorithms

1. (1-NN) A-One Nearest Neighbour.

2. (SVM) Support Vector Machines. 3. Random Forecasting 4. (QDA) Quadratic Linear Analysis.

5. Model based chromatic segmentation

6. Level-set image processing and analysis 7. Fractal based classification

8. Gaussian mixture model algorithm

9. Naïve Bayesian classifier Comparison of the algorithms (Lung cancer):

Table 1

Classification/pattern recognition method	Potential application in digital pathology
Geometric and Appearance Histogram Features	Lung cancer detection on histological slides
Fuzzy Soft Set Based Classification	Genetic expression detection in HE image
Hybrid Shape and Appearance Features	Lung cancer detection
Combination of PCA with SMOTE Resampling	Lung cancer dataset mining and classification
Hyper-Heuristic Algorithm (HHA)	Lung nodule metastasis assessment
Naive Bayes Classifier and C 4.5	Cancer survivability estimation based on tissue imaging data
Multilayer perceptron and machine learning platform	Clinical data forecasting for lung cancer patient (Histopathological data)

Conclusions drawn from the computer aided diagnosis

1) Much better integration among pattern recognition algorithms which drastically improves the computer aided diagnosis by setting up a good platform in digital pathology especially in diagnosis of lung cancer

2) Betterment in the testing of microarray data and thus yield better choices in diagnosis of cancer.

3) The Computer aided diagnosis can be improved using quantatiive image analysis and histological data annotation.

4) Outlier trimming and supervised learning in histopathological slides can reduce the difficulties in medical imaging objective.

DOI: 10.37200/IJPR/V24I1/PR200296
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

5) The CAD can be improved with the help of these supervised learning algorithms.

Breast Cancer Diagnosis using ANFIS

The paper shows the diagnosis of WBCD (Wisconsin Breast Cancer) using ANFIS (Adaptive Neuro Fuzzy Interface System)

The diagnosis of Breast Cancer is a very important real-world medical problem. The complexity of the problem is very high due to the large interdependency among various features. However, it is possible to solve the problem with the present AI technique's.

ANFIS is a mixture of Fuzzy interface system and neural nets. Therefore, it is possible to deal with complex and ambiguous data and learn from the past data. ANFIS is used in this paper as a diagnostic system.

The main concern is the performance of the model. The model should have good computational power as well as correctness of the output form the interface system.

A few of the methods using recommended inputs are proposed such as: genetic algorithm, decision tree and correlation coefficient computation with ANFIS to decrease the computational complexity and also enhance the performance by removing less-relevant input features.

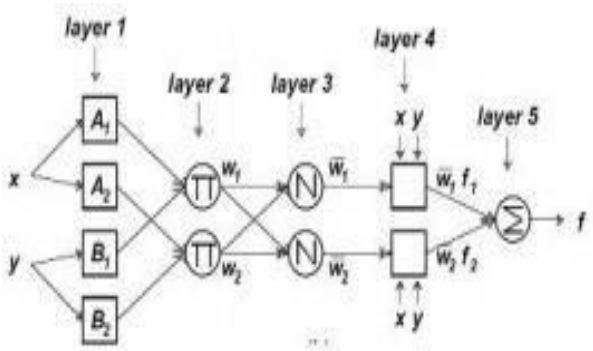

Fig. 6: ANFIS Structure

The paper focuses on automatic breast cancer diagnosis system using ANFIS. It shows the performance against various inputs. Other AI algorithms such as FNN, Neural Nets are also a good option for the diagnosis.

The input reduction methods are a great option in decreasing the large computation and storage. The experiment comes up with a better computational performance and better accuracy.

The benefits of using ANFIS along with input reduction are not just limited to this problem. It can be used for other problems in other context as well which have large data to be analysed.

II. CONCLUSION

This paper has presented a survey of the techniques that exist for performing the diagnosis of cancer using gene data. We have seen a wide variety of techniques in detail starting with SVMs up to the ANFIS. There are a lot of scopes for future improvement of the system. Most of the systems are limited to only few techniques and does the diagnosis using them. We can come up with future systems where we can have a system which is capable of detecting various cancers types with the help of various AI techniques.

DOI: 10.37200/IJPR/V24I1/PR200296
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

REFERENCES

[1] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science*, vol. 270, pp. 467-470, 1995.

[2] J.M. Khan et al., "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine*, vol. 7, pp. 673- 679, 2001.

[3] J. Deutsch, "Evolutionary Algorithms for Finding Optimal Gene Sets in Microarray Prediction," *Bioinformatics*, vol. 19, pp. 4552, 2003.

[4] R. Tibshirani, T. Hastie, B. Narashiman, and G. Chu, "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proc. Nat'l Academy of Sciences USA*, vol. 99, pp. 6567-6572, 2002.

[5] M. P. Brown, W.N. Grundy, D. Lin, N. Cristianini, *C.*W. Sumet, T.*S.*F- . M. Jr. Ares. D. Haussler. "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. N dA c d* Sei *US.4,*Vol. 97,pp 262-267.2000.

[6] Reta, C., et al. Segmentation of Bone Marrow Cell Images for Morphological Classification of Acute Leukemia in FLAIRS Conference. 2010. *USA: Abbrev of Publisher, year, ch.* 4, sec. 21, pp. 142 –147.

[7] Demir, C. and B. Yener, Automated cancer diagnosis based on histopathological images: a systematic survey. *Rensselaer Polytechnic Institute, Tech. Rep*, 2005, p. 232-237

[8] Jyh-Shing Roger Jang. "ANFIS: Adaptive Network Based Fuzzy Inference System". *IEEE Trans. on System, Man and Cybernetics,*vol. 23, no. 3. 1993

[9] O.L. Mangasarian, W.N. Street, and W.H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Mathematical Programming Technical Report 9410, University of Wisconsin*, 1994.

[10] W. H. Land, Jr., T. Masters, and J. Y. Lo."Application of a New Evolutionary Programming/Adaptive Boosting Hybrid to Breast Cancer Diagnosis." *IEEE Congress on Evolutionary Computation Proceedings*, 2000.

DOI: 10.37200/IJPR/V24I1/PR200296
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020