IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 21, 2018, accepted October 15, 2018, date of publication December 20, 2018,
date of current version January 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886604

Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach

JAYADEEP PATI
Department of Computer Science and Engineering, Indian Institute of Information Technology Ranchi, Jamshedpur 831014, India
e-mail: jayadeeppati@gmail.com

ABSTRACT Cancer as a multifactorial disorder develops due to the complex interaction between gene and environment. A person may be susceptible to cancer due to his individual genetic makeup. Cancer causes maximum death worldwide as per data given by the World Health Organization. Some cases are reported with particular genetic makeup. Hence, the proper understanding of eco-genomics of cancer is necessary to interpret the underlying cause and risk factor for cancer. Combining huge gene expression date available in cyberspace and advanced machine learning technique may bring out the strongest candidate genes, which individually or as a part of the complex system, have more accurate prognostic value to determine someone's susceptibility toward cancer. In this paper, we have analyzed gene expression data for the lung cancer available in the Kent Ridge Bio-Medical Dataset Repository. The microarray gene expression data are analyzed to select and predict the optimal subset of genes, which are the most probable causing agent of lung cancer.

INDEX TERMS Gene expression analysis, machine learning technique, eco-genomics, multilayer perceptron, information gain attribute ranking.

I. INTRODUCTION

Predicting who can be more susceptible to have cancer in future and the response to therapy is a challenging field of research. The launch of human genome project has completely changed the way of medical research. A microarray is an efficient tool in the simultaneous study of gene expression of thousands of genes or their RNA product. It gives a clear of up-regulation and downregulation of different genes in different cell or sample under study. Gene expression of disease tissues can be compared to normal tissues for understanding the disease pathology, more accurate diagnosis and clear prognosis. Gene expression can be helpful in determining the type of cancer and also temporal evaluation of tumors. The gene expression patterns from tumors samples derived from different stages of progression can be compared to identify the early and late stage of the disease. Early prediction of cancer can give effective treatment to cancer. We can also predict the chance that a person may be affected by any cancer in future which will help in taking preventive measures to avoid cancer.

Histopathology is not efficient enough for identification of disease progression and clinical outcome in lung adenocarcinoma [32], [33]. Gene expression profiles based on microarray analysis are efficient method to predict patient survival in early-stage lung adenocarcinomas [34]–[37]. In this paper, we have analyzed gene expression data for the Lung cancer available in the Kent Ridge Bio-Medical Dataset Repository [25]. The microarray gene expression data is analyzed to select and predict the optimal subset of genes which are the most probable causing agent of different type of cancer.

The first step is to select the optimal subset of genes to from a large gene set which requires advanced machine learning method. In this paper, we have used a hierarchical approach to optimal gene subset selection. The next step is to classify the cancers datasets and to know the accuracy of the classifier in predicting the cancerous samples. There are number of papers [1]–[6] which focuses on cancer prediction based on different gene expression technique. But they have not considered environmental factor associated with the genes which cause cancer simultaneously.

In this paper, We have used advanced machine learning classifier for classification tasks. We have also given a comparison of the classification accuracy, precision, and recall

value for the different classifier. We have got 72 genes out of 7129 genes in the Lung cancer data which are most probably associated with cancer.

After getting the optimal subset of genes, we have selected top six genes from each cancer type to validate their association with the particular type of cancer and with environmental factors such as heavy metal, air pollutant, ionizing and non-ionizing radiations, food habit, smoking habit, using research article published in scientific journals.

Combing huge gene expression date available in cyber space and advanced machine learning technique may bring out the strongest candidate genes, which individually or as a part of complex system, have more accurate prognostic value to determine someones susceptibility towards cancer. When we add an ecological context, it helps us in better understanding the causes of cancer.

II. MACHINE LEARNING MODELS FOR ANALYSIS

In this section, we present a three advanced machine learning classifier for classifying the cancerous sample.

A. ATTRIBUTE SELECTION

Attribute selection is the process of reducing the number of attributes for use in the classifiers. In applications like gene expression analysis, this becomes all the more important owing to the huge number of genes in Human genome. Building a model on so many attributes can incur huge performance overheads and will suffer from the issue of overfitting. Attribute selection involves a combination of search and attribute estimation which in turn is evaluated against certain classifiers. The techniques employed may be supervised or unsupervised with different performance and accuracy values. The attributes in our application refer to gene expression values that we aim to reduce for an effective prediction, techniques for which are listed below. In our paper, we have used Information Gain Attribute Ranking [4], [17] as a model for selecting a set of most probable genes from a set of vast gene set based on gene expression scores.

1) INFORMATION GAIN ATTRIBUTE RANKING

Information gain [26], [27] makes use of entropy model to rank the attributes. Let us consider an attribute A.

Let's take C as a class,

We can find the entropy of the class as given by these equations. Equation 1 gives entropy before observation and Equation2 gives entropy after observation.

$$H(C) = -\sum_{c \epsilon C} p(c) log_2 p(c) \tag{1}$$

$$H(C|A) = -\sum_{\alpha \epsilon a} p(c|a) log_2 p(c|a) \tag{2}$$

The information gain by the attribute is the amount by which the entropy of the class decreases after addition of the attribute.

B. CLASSIFICATION

After the set of genes have been reduced from the original set, we apply supervised learning on the gene expression data to evaluate the accuracy of prediction. In this paper, we have compared two advanced classifiers on the basis of Accuracy, Precision and Recall.

1) MULTILAYER-PERCEPTRON

Multilayer Perceptron [28] is a feedforward Artificial Neural Network which models the input to output relation through a series of hidden layers which can be supposed to account for individual or a subset of attributes that actively take part in the classification. Each layer has some nodes, and any two consecutive layers form a completely connected bipartite graph with each edge weight representing the probability of transition from one layer to another. It is a supervised learning technique which makes use of error back propagation to correct the edge weights with each new instance. The major parameters that affect the accuracy of an MLP include:

1) Transition weights from one layer to another.
2) The number of hidden layers selected. Higher number leads to deeper learning.
3) Activation function selected, which maps the output from the MLP to response.

MLPs have been found to be extremely useful and accurate and have wide applications in multiple domains.

2) RANDOM SUBSPACE

Random subspace [29] classification technique makes use of multiple instances of same or different classifiers each working on the subset of attributes or a subspace. The output of Random Subspace is based on the class results produced by these individual classifiers. This type of classification is also referred to as attribute bagging.

1) Let us take N number of training objects with D number of features
2) Let us assume L to be the number of Individual Classifiers in the ensemble.
3) Let us choose $d_i (d_i < D)$ as the number of input variable for I for each classifier I. There exist one value of d_i for all the individual classifiers.
4) Let us create training set by choosing di features from D without replacement For each individual classifier I. The features are chosen without replacement.
5) The outputs of L individuals classifiers are combined by majority voting or the posterior probabilities to classify a new object.

3) SMO (SEQUENTIAL MINIMAL OPTIMIZATION)

SMO (Sequential Minimal Optimization) [30] is an improvement of the basic support vector machine and is one of the most popular training tool for SVM. The aim of SVM is to find a hyperplane that gives the best accuracy in categorizing the inputs into one of the two classes. The input instances are projected into higher dimensional space and a hyperplane is

Rank	Attribute Score	Attribute Number in original expression Data	Attribute Name
1	0.482	17	J02874_at
2	0.482	24	L34657_at
3	0.482	51	U60115_at
4	0.482	47	U39447_at
5	0.482	69	X64559_at
6	0.482	70	Z11793_at
7	0.432	55	U89336_cds3_at
8	0.432	29	M83186_at
9	0.432	72	Z18951_at
10	0.432	43	U24488_s_at
11	0.432	5	D13628_at
12	0.432	18	J03890_rna1_at
13	0.401	61	X05130_s_at
14	0.4	44	U29171_at
15	0.377	28	M61906_at
16	0.377	40	U13219_at
17	0.377	53	U76764_s_at
18	0.356	4	D13626_at
19	0.356	23	L27479_at
20	0.356	3	AF001294_at
21	0.351	21	L10955_cds1_s_at
22	0.339	26	M19722_at
23	0.339	7	D26070_at
24	0.339	62	X07695_at
25	0.339	50	U49020_cds2_s_at
26	0.323	19	K03195_at
27	0.323	42	U24152_at
28	0.323	32	M90516_at
29	0.323	36	S74017_at
30	0.323	58	U97105_at
31	0.322	2	AF000959_at
32	0.322	11	D50683_at
33	0.322	49	U45973_at
34	0.309	35	M94250_at
35	0.309	34	M93221_at
36	0.305	59	X00129_at
37	0.3	37	S85655_at
38	0.299	8	D26599_at
39	0.292	67	X61118_rna1_at
40	0.286	63	X54326_at
41	0.285	20	L06139_at
42	0.285	33	M93036_at
43	0.285	41	U19247_rna1_s_at
44	0.28	39	U03105_at
45	0.274	14	HG1612-HT1612_at
46	0.274	1	AB003102_at

FIGURE 1. Genes selected on the basis of info gain ranking results.

47	0.274	38	U02493_at
48	0.274	12	D88422_at
49	0.264	27	M21305_at
50	0.264	56	U93237_rna2_at
51	0.264	48	U43077_at
52	0.264	68	X64044_at
53	0.254	65	X57766_at
54	0.254	22	L13773_at
55	0.254	52	U73379_at
56	0.249	60	X01060_at
57	0.249	54	U83461_at
58	0.236	9	D49394_at
59	0.225	10	D49410_at
60	0.222	13	HG1153-HT1153_at
61	0.214	30	M84332_at
62	0.208	66	X58288_at
63	0.201	25	M14200_rna1_at
64	0.197	46	U37707_at
65	0.197	45	U30872_at
66	0.195	16	HG4683-HT5108_s_at
67	0.189	15	HG2868-HT3012_s_at
68	0.173	64	X57152_rna1_s_at
69	0.173	6	D21063_at
70	0.171	57	U93867_at
71	0.149	31	M88461_s_at
72	0.14	71	Z14000_at

FIGURE 1. ***(Continued.)*** **Genes selected on the basis of info gain ranking results.**

found that is at maximum distance from the closest inputs to the hyperplane from both the classes. It suffers from a Quadratic Programming problem which is addressed in SMO, explanation of which is beyond scope of this report.

III. DATA COLLECTION

The gene expression data for 86 primary lung adenocarcinomas samples and 10 non-neoplastic lung samples are collected [25]. All the samples contain 7129 genes. Gene expression data for the Lung cancer available in the Kent Ridge Bio-Medical Dataset Repository [25]. The entire dataset is divided into 70 % as training set and 30% as a test set. The result is validated on both training and test dataset. The model which gives more accurate result on test dataset is selected as the most suitable model.

IV. IMPLEMENTATION OF MACHINE LEARNING MODELS

After collecting the gene expression data for 86 primary lung adenocarcinomas samples and 10 non-neoplastic lung samples, the next step is to get an optimal number of genes from a huge gene set.

A. INFO GAIN RANKING BASED ATTRIBUTE SELECTION

We applied the Info gain ranking method for selecting a set of most probable genes from a set of 7129 genes The Info Gain Ranking Method is implemented in Java Environment. The Set of attributes we got after applying the Info Gain Ranking Method is presented in Figure 1. We have also sorted the attributes based on score based on the information gain. After applying the Info Gain Ranking method, the number of genes reduced to 72 from a huge number of 7129 genes. Here the attribute name corresponds to particular gene in the sample. Among them we have given biological relevance for top six ranked genes.

B. CLASSIFICATION

After getting a set of most probable genes, the next step is to apply advanced classification model for classifying

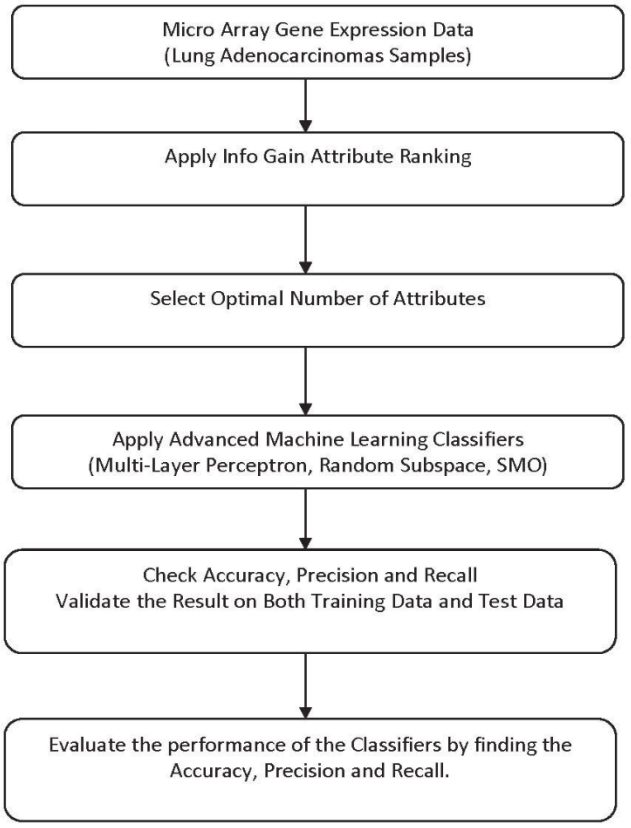

FIGURE 2. Flow diagram for computational modeling of gene expression data for lung cancer.

cancerous samples. In this paper, we have applied Multilayer Perceptron and Random Subspace on the set of selected attributes for classifying the cancerous samples. We have also validated our results on both training and test data. The flow diagram of computational modeling of the gene expression data is given in figure 2. Finally, the classification models are compared on the basis of Accuracy, Precision and Recall. The results of the attribute selection were stored for future reference to establish a biological correlation between expressed gene and the disease. This information may be used to carry out DNA sequence analysis for the gene to find out mutations which may result in genetic defects.

V. EVALUATION AND INTERPRETATION

The models are evaluated on the basis of Accuracy, Precision and Recall [23]. The definition of all these terms is given below.

- Accuracy is the commonly used performance measure. $Accuracy = (TP + TN)/(TP + TN + FP + FN)$
- Precision is simply the ratio of correct positive observations. $Precision = TP/(TP + FP)$
- The recall is also called sensitivity or true positive rate. $Recall = TP/(TP + FN)$

TABLE 1. Parameter description.

Parameter	Defination
True Positives (TP)	A number of positive examples, labeled as such.
False Positives (FP)	A number of negative examples labeled as positive.
True Negatives (TN)	A number of negative examples, labeled as such.
False Negatives (FN)	A number of positive examples labeled as negative.

TABLE 2. Classification result (lung cancer Michigan 70/30 percent split).

Classifier Name	Accuracy	Precision	Recall
Multi-Layer Perceptron	86.6667	0.8714	0.8315
Random Sub Space	68.3333	0.6458	0.6025
SMO	91.6667	0.9125	0.9029

TABLE 3. Lung cancer most probable 6 genes (as from computational result).

Gene (Computational Name)	Gene (Biological Name)
$J02874_{at}$	Fatty acid binding protein 4, adipocyte
$L34657_{at}$	Platelet/endothelial cell adhesion molecule
$U60115_{at}$	Four and a half LIM domains1
$U39447_{at}$	Amine oxidase, copper containing 3
$X64559_{at}$	C-type lectin domain family 3, member B
$Z11793_{at}$	Selenoprotein P, plasma, 1

The definition of TP, FP, TN and FN is given in Table 2. In Table 3, we presented the Accuracy, Precision, and Recall after classifying the cancer samples using the Multi-Layer Perceptron and Random Subspace model. From the table, we observe the SMO is the most accurate model for classifying the cancerous samples. The precision and recall value as given by SMO is also much better than Multi-Layer Perceptron and Random Subspace. Hence SMO can be used as a most suitable model for predicting the cancerous samples.

VI. BIOLOGICAL RELEVANCE

From the gene expression data, we have identified the most probable genes associated with cancerous samples. From the computational point of view, these are the genes which are associated the cancerous samples. The next step to validate the genes from a biological point of view and also to find the environmental factor associated with the particular type of cancer. From a set of 73 genes, we have selected the best 6 genes having higher Info Gain Score. The genes with its biological names are presented in Table 4. We have searched in various research papers to find out the role of these genes in lung cancer patients. We have also searched for the associated environmental factors and their effect on a particular type of gene. We found significant contribution of these genes in lung cancer samples.

- Significant contribution of FABP4 gene is found in lung adenocarcinoma, lung cancer, nonsmall cell lung cancer, normal airway epithelia, non-small cell lung carcinoma [8]. We also found that, under the conditions of fasting and cold stress Fatty acid binding protein 4 and 5 has an important part to play in thrombogenesis [9]. So this thermogenesis can be an environmental factor affecting the expression of FABP4.
- Platelet endothelial cell adhesion molecule-1 1 is found in non-small-cell lung cancer patients [10], [11].

In another paper it is found that Platelet Endothelial Cell Adhesion Molecule-1 and Pigment Epithelium-Derived Factor has role in Non-Small-Cell- Lung Cancer. We also found a significant effect of photoactivation, laser light exposure, electrical stimulation, or topical application of caustic chemicals (e.g., ferric chloride) in controlling gene expression of cell adhesion molecule-1 [12].

- There is evidence of inhibitory role of FHL1 in lung cancer [13], [14]. A significant contribution of an environmental factor on the gene expression of FHL1 also exists [15].
- In one paper, it is found that downregulation of Monoamine Oxidase-A in causes cancer in multipler organs [16]. We also found the significant role of environmental exposure on the gene expression of Monoamine Oxidase-A [17].
- There also exists evidence of Expression of CLEC3B in cancer [18]. Another paper presents a significant effect of High occupational exposure on the gene expression of CLEC3B [19].
- We also found Significance of selenium levels in non-small cell lung cancer patients [20], [21]. We also found evidence of
- Occupational Exposure on the Selenoprotein P in causing Lung cancer [22].

VII. CONCLUSION

Cancer has been a potential threat to human eco system as it affects in persons identity, roles and normal functioning in the environment. The solution is early prediction of cancer and also to find out cancer probabilities of an individual in future which will help to avoid cancer. Gene expression of cancerous cells can be compared to normal cells for understanding the pathology of cancer. In this paper we used advanced machine learning technique to analyses the gene expression of cancerous samples to find out the genes which have maximum probability of causing cancer. Combing huge gene expression date available in cyber space and advanced machine learning technique may bring out the strongest candidate genes, which individually or as a part of complex system, have more accurate prognostic value to determine someones susceptibility towards cancer. When we add an ecological context, it helps us in better understanding the causes of cancer. This has changed the cancer research magnificently. Our approach can be considered as an application to Eco-genomics [7].

REFERENCES

[1] L. Shoon *et al.*, "Cancer recognition from DNA microarray gene expression data using averaged one-dependence estimators," *Int. J. Cybern. Inform.*, vol. 3, no. 2, pp. 1–10, 2014.

[2] G. Russo, C. Zegar, and A. Giordano, "Advantages and limitations of microarray technology in human cancer," *Oncogene*, vol. 22, no. 42, pp. 6497–6507, 2003.

[3] X. Wang and O. Gotoh, "Microarray-based cancer prediction using soft computing approach," *Cancer Inform.*, vol. 7, Jan. 2009.

[4] A. Bashetha and G. U. Srikanth, "Effective cancer detection using soft computing technique," *IOSR J. Comput. Eng.*, vol. 17, no. 1, pp. 1–5, 2015.

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

[6] M. C. Ungerer, L. C. Johnson, and M. A. Herman, "Ecological genomics: Understanding gene and genome function in the natural environment," *Heredity*, vol. 100, no. 2, pp. 178–183, 2008.

[7] S. A. Pavey, L. Bernatchez, N. Aubin-Horth, and C. R. Landry, "What is needed for next-generation ecological and evolutionary genomics?" *Trends Ecol. Evol.*, vol. 27, no. 12, pp. 673–678, 2012.

[8] *FABP4 Fatty Acid Binding Protein 4 [Homo Sapiens (Human)]*. [Online]. Available: http//www.ncbi.nlm.nih.gov/gene/2167

[9] M. R. A. A. Syamsunarno *et al*, "Fatty acid binding protein 4 and 5 play a crucial role in thermogenesis under the conditions of fasting and cold stress," *PLoS ONE*, vol. 9, no. 3, p. e90825, 2014.

[10] B.-H. Kuang *et al.*, "The prognostic value of platelet endothelial cell adhesion molecule-1 in non-small-cell lung cancer patients," *Med. Oncol.*, vol. 30, no. 2, p. 536, 2013.

[11] A. Emmert *et al.*, "Interplay between platelet endothelial cell adhesion molecule-1 and pigment epithelium-derived factor in non-small-cell-lung cancer," *J. Thoracic Surg.*, vol. 6, no. 4, pp. 47–56, 2016.

[12] B. Piedboeuf, M. Gamache, J. Frenette, S. Horowitz, H. S. Baldwin, and P. Petrov, "Increased endothelial cell expression of platelet–endothelial cell adhesion molecule-1 during hyperoxic lung injury," *Amer. J. Respiratory Cell Mol. Biol.*, vol. 19, no. 4, pp. 543–553, 1998.

[13] *FHL1 Four and a Half LIM Domains 1 [Homo Sapiens (Human)]*. [Online]. Available: http//www.ncbi.nlm.nih.gov/gene?Db=geneCmd=DetailsSearchTerm=2273

[14] C. Niu *et al.*, "Downregulation and growth inhibitory role of FHL1 in lung cancer," *Int. J. Cancer*, vol. 130, no. 11, pp. 2549–2556, 2012.

[15] A. Raskin *et al.*, "A novel mechanism involving four and a half lim domain protein-1 and extracellular-signal-regulated kinase-2 regulates titin phosphorylation and mechanics," *J. Biol. Chem.*, vol. 287, no. 35, pp. 29273–29284, 2012.

[16] L. A. Rybaczyk, M. J. Bashaw, D. R. Pathak, and K. Huang, "An indicator of cancer: Downregulation of monoamine oxidase—A in multiple organs and species," *BMC Genomics*, vol. 9, no. 1, p. 134, 2008.

[17] H. Hu, W. Wang, H. Tang, and P. Xu, "Characterization of pseudooxynicotine amine oxidase of *pseudomonas putida* S16 that is crucial for nicotine degradation," *Sci. Rep.*, vol. 5, Dec. 2015, Art. no. 17770.

[18] *Expression of CLEC3B in Cancer—Summary—The Human Protein Atlas*. [Online]. Available: http://www.proteinatlas.org/ENSG00000163815-CLEC3B/cancer

[19] A. Arita *et al.*, "Gene expression profiles in peripheral blood mononuclear cells of Chinese nickel refinery workers with high exposures to nickel and control subjects," *Cancer Epidemiol. Prevention Biomarkers*, vol. 22, no. 2, pp. 261–269, 2013.

[20] M. Epplein, R. F. Burk, Q. Cai, M. K. Hargreaves, and W. J. Blot, "A prospective study of plasma Selenoprotein P and lung cancer risk among low-income adults," *Cancer Epidemiol. Prevention Biomarkers*, vol. 23, no. 7, pp. 1238–1244, 2014.

[21] G. Elassal, H. Samy, M. Said, and S. Elbatrawy, "Significance of selenium levels in non-small cell lung cancer patients: A comparative study," *Egyptian J. Chest Diseases Tuberculosis*, vol. 63, no. 4, pp. 1019–1023, 2014.

[22] J. Gromadzinska, W. Wasowicz, K. Rydzynski, and N. Szeszenia-Dabrowska, "Oxidative-stress markers in blood of lung cancer patients occupationally exposed to carcinogens," *Biol. Trace Element Res.*, vol. 91, no. 3, pp. 203–215, 2003.

[23] S. Uwagbole, W. Buchanan, and L. Fan, "Applied Web traffic analysis for numerical encoding of SQL injection attack features," in *Proc. 15th Eur. Conf. Cyber Warfare Secur. (ECCWS)*, 2016.

[24] A. Bashetha and G. U. Srikanth, "Effective cancer detection using soft computing technique," *IOSR J. Comput. Eng.*, vol. 17, no. 1, pp. 1–5, 2015.

[25] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *J. Comput. Biol.*, vol. 6, nos. 3–4, pp. 281–297, 1999.

[26] Li. (2002). Kent ridge bio-medical data set repository. Institute Infocomm Research. [Online]. Available: http://sdmc.lit.org.sg/GEDatasets/Datasets.html

[27] B. Azhagusundari and A. S. Thanamani, "Feature selection based on information gain," *Int. J. Innov. Technol. Exploring Eng.*, vol. 2, no. 2, pp. 18–21, 2013.

[28] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 6, pp. 1437–1447, Nov./Dec. 2003.

[29] A. Mateos, J. Dopazo, R. Jansen, Y. Tu, M. Gerstein, and G. Stolovitzky, "Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons," *Genome Res.*, vol. 12, no. 11, pp. 1703–1715, 2002.
[30] A. Bertoni, R. Folgieri, and G. Valentini, "Bio-molecular cancer prediction with random subspace ensembles of support vector machines," *Neurocomputing*, vol. 63, pp. 535–539, Jan. 2005.
[31] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods*. 1999, pp. 185–208.
[32] M. E. Garber *et al.*, "Diversity of gene expression in adenocarcinoma of the lung," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 24, pp. 13784–13789, 2001.
[33] J. Lu *et al.*, "MicroRNA expression profiles classify human cancers," *Nature*, vol. 435, no. 7043, pp. 834–838, 2005.
[34] S. Ramaswamy *et al.*, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 26, pp. 15149–15154, 2001.
[35] K. Shedden *et al.*, "Gene expression–based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study," *Nature Med.*, vol. 14, no. 8, pp. 822–827, 2008.
[36] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 43, pp. 15545–15550, 2005.
[37] K. A. Olaussen *et al.*, "DNA repair by ERCC1 in non–small-cell lung cancer and cisplatin-based adjuvant chemotherapy," *New England J. Med.*, vol. 355, no. 10, pp. 983–991, 2006.

JAYADEEP PATI received the B.Tech. degree from the Institute of Technical Education and Research, Bhubaneswar, India, in 2010, the M.Tech. degree from the National Institute of Technology, Rourkela, India, in 2012, and the Ph.D. degree from IIT (BHU), Varanasi. He is currently an Assistant Professor with the Indian Institute of Information Technology, Ranchi. His research interests include machine learning and software engineering.