GENE EXPRESSION MINING FOR PREDICTING SURVIVABILITY OF PATIENTS IN EARLY STAGES OF LUNG CANCER

Shoon Lei Win, Zaw Zaw Htike, Faridah Yusof, Ibrahim A. Noorbatcha

Faculty of Engineering, IIUM, Kuala Lumpur, Malaysia

ABSTRACT

After numerous breakthroughs in medicine, microbiology, and pathology in the past century, lung cancer still remains as a leading cause of cancer-related death even in the developed countries. Lung cancer accounts roughly for 30% of all cancer-related deaths in the world. Diagnosis and treatments are still based on traditional histopathology. It is of paramount importance to predict the survivability of patients in early stages of lung cancer so that specific treatments can be sought. Nonetheless, histopathology has been shown by previous studies to be inadequate in predicting lung cancer development and clinical outcome. The microarray technology allows researchers to examine the expression of thousands of genes simultaneously. This paper describes a state-of-the-art machine learning based approach called averaged one-dependence estimators with subsumption resolution to tackle the problem of predicting whether a patient in early stages of lung cancer will survive by mining DNA microarray gene expression data. To lower the computational complexity, we employ an entropy-based gene selection approach to select relevant genes that are directly responsible for lung cancer survivability prognosis. The proposed system has achieved an average accuracy of 92.31% in predicting lung cancer survivability over 2 independent datasets. The experimental results provide confirmation that gene expression mining can be used to predict survivability of patients in early stages of lung cancer.

KEYWORDS

Lung cancer, survivability prediction; microarray gene expression

1. INTRODUCTION

Lung cancer continues to contribute as one of the most common cause of cancer-related mortality even in the developed countries [1]. Lung cancer accounts roughly for 30% of all cancer deaths in the world. The number of deaths from lung cancer is greater than the total number of deaths from the next three most notorious cancers (breast, colon, and prostate) combined [2]. Despite recent advances in cancer diagnosis and treatment, the overall 10-year survival rate still remains at a mediocre level of 8–10% [3]. Table 1 lists the lung cancer survival rates. Lung cancer prognosis essentially depends on the subtype and stage of the lung cancer. The 5-year survival rate is the percentage of people who are still alive 5 years or longer after being diagnosed with lung cancer [4]. Although survival rates are statistically known, it is not currently possible to identity high-risk patients [5-7]. Diagnosis of lung cancer has remained essentially unchanged for 30 years [8], and continues to be based on conducting histopathological examination of tissue samples as shown in Figure 1 [9]. Histopathology has been shown by previous studies to be inadequate in both prognosis and treatment [10]. Therefore, it is extremely difficult for physicians to know which patients diagnosed with lung cancer will survive. The probability that a lung cancer will recur and the probable timing depend on the type of the primary cancer. Gene expression profiling using the DNA microarray technology could bring about the ability to predict the survivability of lung cancer patients by analyzing the gene expression levels of the cancer cells.

DOI: 10.5121/ijbb.2014.4201

Table 1. Lung cancer survival rates [4, 11].

Stage	Tumor, node, metastasis (TNM) staging	5-year survival rate
IA	T1, N0, M0	More than 70%
IB	T2, N0, M0	60%
IIA	T1, N1, M0	50%
IIB	T2, N1, M0	30%
	T3, N0-N1, M0	40%
IIIA	T1-T3, N2, M0	10%-30%
IIIB	Any T4, any N3, M0	Less than 10%
IV	Any M1	Less than 5%

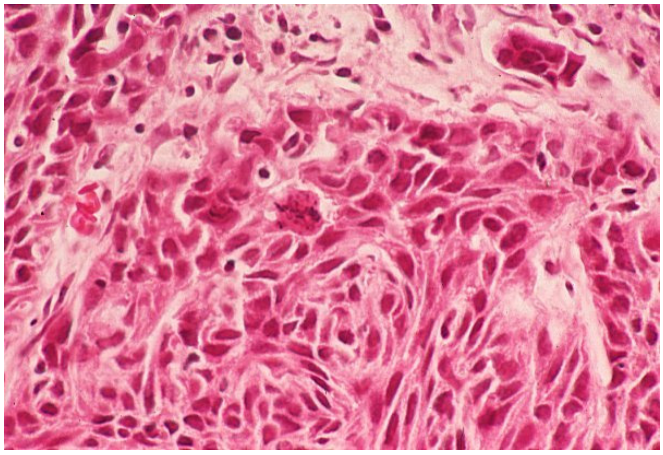

Figure 1. Lung cancer tissue for histopathogical examination [12].

In this paper, we tackle the problem of predicting lung cancer survivability by mining DNA microarray gene expression data. During the past few decades, applications of pattern recognition and machine learning techniques have emerged in many domains [13-22]. Pattern recognition and machine learning techniques have also recently become popular in the arena of microarray gene expression analysis. There have been some attempts to predict cancer survivability using machine learning techniques. For instance, Beer et al. [3] employed univariate Cox regression to predict survivability of patients with lung adenocarcinoma using microarray gene expression analysis. Ensemble techniques have also become popular in gene expression mining. Shedden et al. [23] utilized an ensemble of eight classifiers, each of which produced either categorical or continuous risk scores. Final scoring and classification was carried out using penalized Cox regression. Urgard et al. [24] applied Bayesian regression analysis and Kaplan-Meier survival analysis to determine metagenes associated with lung cancer survival. Their experiments showed that gene-expression patterns and metagene profiles could be applied to predict the probability of survival outcomes of lung cancer patients. Ford et al. [25] proposed a General Regression Neural Network (GRNN) Oracle ensemble that combined several Partial least squares (PLS) models trained to predict lung cancer outcome from 12 different gene networks. They concluded that it was possible to correctly predict lung cancer outcome by combining the results based on their proposed individual gene network models. Similarly, Norris et al.[26] applied the very same GRNN oracle to predict cancer outcome. They confirmed that GRNN led to high prediction accuracy. This paper describes an approach based on a state-of-the-art machine learning technique called averaged one-dependence estimators with subsumption resolution to tackle the problem of predicting lung cancer outcome from microarray gene expression data.

2. CANCER SURVIVABILITY PREDICTION

The objective of lung cancer survivability prediction is to predict whether a particular a patient in an early stage of lung cancer will survive given the gene expression data from tissue samples. We propose a three-layered framework that consists of gene selection, discretization, and prediction as shown in Figure 2. The complexity of a machine learning classifier depends upon the dimensionality of the input data [27]. There is also a phenomenon known as the 'curse of dimensionality' that arises with high dimensional input data [28]. In the case of genetic data classification, not all the genes in a genetic sequence might be responsible for predicting cancer survivability. Therefore, we propose to employ a gene selection process to select relevant prognostic genes in an unsupervised manner and an entropy-based discretization process to discretize gene expression levels. Section 2.1 describes the process of gene selection and Section 2.2 describes the process of discretization. After dimensionality reduction, we propose to perform cancer survivability prediction using the averaged one-dependence estimators with subsumption resolution (AODEsr). Section 2.3 describes the process of prediction.

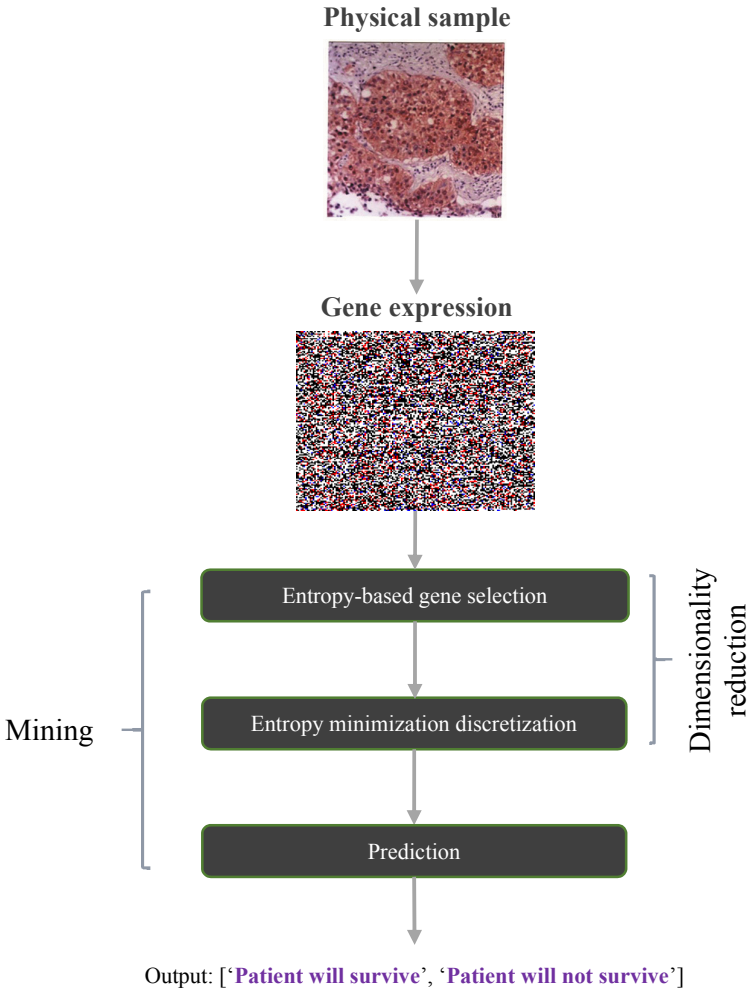

Figure 2. High-level flow diagram of lung cancer survivability prediction framework (part of the image obtained from [29]).

2.1. Entropy-based gene selection

The complexity of any machine learning classifier depends upon the dimensionality of the input data [27]. Generally, the lower the complexity of a classifier, the more robust it is. Moreover, classifiers with low complexity have less variance, which means that they vary less depending on the particulars of a sample, including noise, outliers, etc. [27]. In the case of cancer lung survivability prediction, not all the genes in a genetic sequence might be prognostic biomarkers for survival ability. Therefore, we need to have a gene selection method that chooses a subset of relevant prognostic genes, while pruning the rest of the genes in the microarray gene expression data [30, 31]. In essence, we are interested in finding the best subset of the set of genes that can sufficiently predict cancer survivability. Ideally, we have to choose the best subset that contains the least number of genes that most contribute to the prediction accuracy, while discarding the rest of the genes. There are 2^n possible subsets that can arise from an n-gene long genetic sequence. In essence, we have to choose the best subset out of 2^n possible subsets. Because performing an exhaustive sequential search over all possible subsets is computationally expensive, we need to employ heuristics to find a reasonably good subset that can sufficiently predict cancer survivability.

We employ a prognostic gene selection process based on an information-theoretic concept of entropy. Given a set of genes X and $p(x_i)$ which represents the probability of the i^{th} gene, then the entropy of genes, which measures the amount of 'uncertainty', is defined by:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\ log\, p(x_i) \tag{1}$$

Entropy is a non-negative number. $H(X)$ is 0 when X is absolutely certain to be predicted. The conditional entropy of class label Y given the genes is defined by:

$$H(Y \mid X) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,\, y_j)\, ln\, \frac{p(y_j)}{p(x_i,\, y_j)} \tag{2}$$

The information gain (IG) of the genes from the class label Y is defined to be:

$$IG(Y \mid X) = H(Y) - H(Y \mid X) \tag{3}$$

The gain ratio (GR) between the genes and the class label Y is defined to be:

$$GR(Y \mid X) = \frac{IG(Y \mid X)}{H(Y)} \tag{4}$$

The GR of a gene is a number between 0 and 1 which approximately represents the 'prognostic capacity' of the gene. A GR of 0 roughly indicates that the corresponding individual gene has no significance in cancer survivability prediction while a GR of 1 roughly indicates that the gene is significant in cancer survivability prediction. During the training phase, the GR for each gene is calculated according to (4). All the genes are then sorted by their GRs. Genes whose GRs are higher than a certain threshold value are selected as discriminating genes while the rest are discarded. Training needs to be carried out only once.

2.2. Entropy minimization discretization

Microarray gene expression heat map is essentially a matrix of gene expression levels. Each gene expression level is a continuous number. It has been demonstrated in a number of studies that many classification algorithms seem to work more effectively on discrete data or even more strictly, on binary data [32]. Therefore, discretization is a desired step. Discretization is a process in which continuous gene expression levels are transformed into discrete representation which is comparable to linguistic expressions such as 'very low', low', 'high', and 'very high'. There are numerous discretization techniques in the literature [33]. However, we have adopted EMD (Entropy Minimization Discretization) [34] because of its reputation in discretization of high-dimensional data. The training instances are first sorted in an ascending order. The EMD algorithm then evaluates the midpoint between each successive pair of the sorted values of an attribute as a potential cut point [35]. While evaluating each candidate cut point, the data are discretized into two intervals and the resulting class information entropy is calculated. A binary discretization is determined by selecting the cut point for which the entropy is minimal amongst all candidates [32]. The binary discretization is applied recursively, always selecting the best cut point. A minimum description length criterion (MDL) is applied to decide when to stop discretization [34]. The results of the discretization process are carried forward to the prediction stage.

2.3. Classification

Naive Bayes (NB), which is fundamentally built on the strong independence assumption, is a very popular classifier in machine learning due to its simplicity, efficiency and efficacy [36-39]. There have been numerous applications of NB and variants thereof. The conventional NB algorithm uses the following formula for classification [40]:

$$Output = \underset{y}{\operatorname{argmax}} \left(P(y \mid x_1, \cdots, x_n)\right) \tag{5}$$

NB performs fairly accurate classification. The only limitation to its classification accuracy is the accuracy of the process of estimation of the base conditional probabilities. One clear drawback is its strong independence assumption which assumes that attributes are independent of each other in a dataset. In the field of genetic sequence classification, NB assumes that genes are independent of each other in a genetic sequence despite the fact that there are apparent dependencies among individual genes. Because of this fundamental limitation of NB, researchers have proposed various techniques such as one-dependence estimators (ODEs) [41] and super parent one-dependence estimators (SPODEs) [42] to ease the attribute independence assumption. In fact, these approaches alleviate the independence assumption at the expense of computational complexity and a new set of assumptions. Webb [36] proposed a semi-naive approach called averaged one-dependence estimators (AODEs) in order to weaken the attribute independence assumption by averaging all of a constrained class of classifiers without introduction of new assumptions. The AODE has been shown to outperform other Bayesian classifiers with substantially improved computational efficiency [36]. The AODE essentially achieves very high classification accuracy by averaging several semi-naive Bayes models that have slightly weaker independence assumptions than a pure NB. The AODE algorithm is effective, efficient and offers highly accurate classification. The AODE algorithm uses the following formula for classification [40]:

$$Output = \underset{y}{\operatorname{argmax}} \left(\sum_{i:\, 1 \le i \le n \,\wedge F(x_i) \ge m} P(y, x_i) \prod_{j=1}^{n} P\left(x_j \mid y,\, x_i\right) \right) \tag{6}$$

Semi-naive Bayesian classifiers attempt to preserve the numerous strengths of NB while reducing error by relaxing the attribute independence assumption [40]. Backwards sequential elimination (BSE) is a wrapper technique for attribute elimination that has proved to be effective at this task. Zheng et al. [40] proposed a new approach called *lazy estimation* (LE), which eliminated highly related attribute values at classification time without the computational overheads that are intrinsic in classic wrapper techniques. Their experimental results show that LE significantly reduces bias and error without excessive computational overheads. In the context of the AODE algorithm, LE has a significant advantage over BSE in both computational efficiency and error. This novel derivative of the AODE is called the averaged one-dependence estimators with subsumption resolution (AODEsr). In essence, the AODEsr enhances the AODE with a subsumption resolution by detecting specializations among attribute values at classification time and by eliminating the generalization attribute value [40]. Because the AODEsr has a very weak independence assumption, it performs well in classification. Therefore, we employ an AODEsr classifier to perform lung cancer survivability prediction.

3. EXPERIMENTS

The proposed lung cancer survivability framework was implement in C# 5.0 programming language using IKVM. Figure 3 illustrates a screenshot of the implemented lung cancer survivability prediction system.

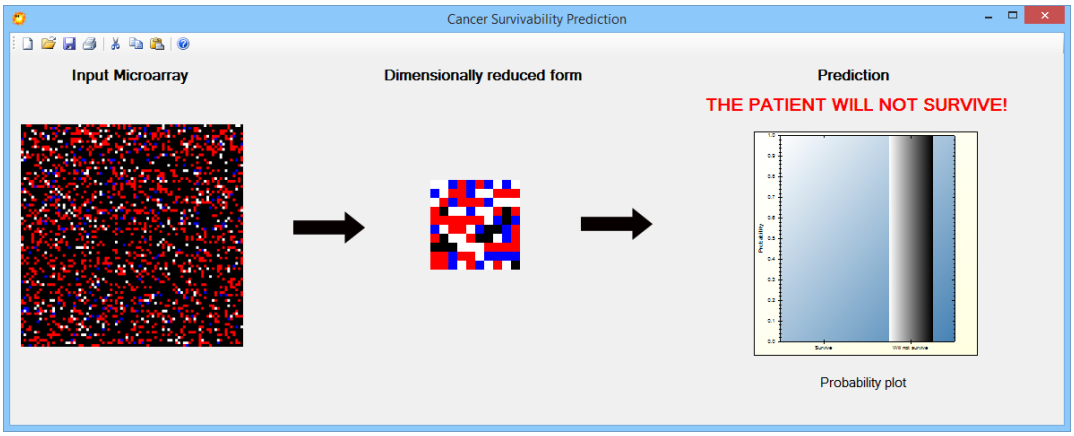

Figure 3. Screenshot of implemented lung cancer survivability prediction system.

We tested our proposed system using two independent datasets from the literature as listed in Table 2. For each dataset, we carried out leave-one-out cross validations (LOOCV) where an *N*-sized dataset was partitioned into *N* equal-sized sub-datasets. Out of the *N* sub-datasets, a single sub-dataset was retained as the validation data for testing the model, and the remaining *N* - 1 sub-datasets were used as training data. The whole cross-validation process was then repeated *N* - 1 more times such that each of the *N* sub-datasets got used exactly once as the validation data. The results were then averaged over all the *N* trials. We used a critical value of 1, frequency limit of 250, an M-estimate weight value of 0.03 for the AODEsr model for all the trails.

Table 2. Two independent lung cancer datasets used in our experiments.

Dataset	#Genes	#Samples
Beer et al. [3]	7129	96
Wigle et al. [2]	2880	39

For each dataset, we performed one LOOCV experiment for varying number of selected genes ranging from 5 to 75. The genes for each trail were selected using the entropy-based technique outlined in Section 2.2. Table 3 lists our experimental results in a tabular format. The most surprising finding was that the system achieved a 100% accuracy in predicting lung cancer survivability for any number of genes for Beer et al.'s dataset. The results indicate that there are only a few genes that act as prognostic biomarkers for lung cancer survivability. In other words, using a few genes will be enough to predict whether a particular lung cancer patient will survive. The system achieved much lower accuracy for Wigle et al.'s dataset. This might be because of outliers and very low number of samples in Wigle et al.'s dataset. Theoretically, accuracy should monotonically increase with an increase in the number of selected genes because AODEsr, like any other Bayesian classifiers, is not sensitive to irrelevant features. Adding extra genes should not theoretically downgrade accuracy. However, as shown in Table 3, for Wigle et al.'s dataset, the system achieved an 87.18% accuracy with 20 genes and an 84.62% accuracy for 50 genes, exhibiting a departure from monotonicity. The disruption in monotonicity might be because of the intrinsic imperfection in the gene selection procedure. The steady-state average LOOCV accuracy of our lung cancer survivability predictor is 92.31%. It is worth iterating the fact that we used the AODEsr classifier with the same set of parameters (critical value of 1, frequency limit of 250, an M-estimate weight value of 0.03) throughout all the experiments to prevent bias. The accuracy rate of the proposed cancer survivability prediction system using the AODEsr classifier with the entropy-based selection process seems to be significantly higher than the base line accuracy of binary classification, which is 50%.

Table 3. LOOCV accuracy of the system on 2 datasets with varying number of selected genes.

Number of genes	LOOCV accuracy	
	Beer et al. [3] Dataset	Wigle et al. [2] Dataset
5	100	82.05
10	100	84.62
15	100	84.62
20	100	87.18
50	100	84.62
75	**100**	**84.62**

4. CONCLUSION

Lung cancer is a major leading cause of cancer-related deaths in the world. Current treatments for lung cancer are still based on traditional histopathology. It is of paramount importance to predict the survivability of patients suffering from lung cancer so that specific treatments can be sought. Nonetheless, histopathology has been shown by previous studies to be inadequate in predicting lung cancer development and clinical outcome. We have presented a machine learning based approach to predict lung cancer survivability from microarray gene expression data. We employ a state-of-the-art machine learning approach called the averaged-on dependence estimator with subsumption resolution (AODEsr) to tackle the problem of predicting lung cancer survivability. Given a set of gene expression data, the system predicts whether a patient suffering from lung cancer will survive. To lower the computational complexity and to increase the generalization capability of the system, we employ an entropy-based gene selection algorithm to select relevant prognostic genes for lung cancer survival. We have carried out experiments on 2 independent datasets acquired from the literature. This proposed system has achieved an accuracy of 92.31% in predicting lung cancer survivability. The experimental results demonstrate the efficacy of our framework.

ACKNOWLEDGEMENTS

This research was supported by Research Acculturation Collaborative Effort (RACE) Grant from the Ministry of Education, Malaysia.

REFERENCES

[1] R. T. Greenlee, et al., "Cancer statistics, 2001," *CA: a cancer journal for clinicians,* vol. 51, pp. 15-36, 2001.

[2] D. A. Wigle, et al., "Molecular profiling of non-small cell lung cancer and correlation with disease-free survival," *Cancer Research,* vol. 62, pp. 3005-3008, 2002.

[3] D. G. Beer, et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature medicine,* vol. 8, pp. 816-824, 2002.

[4] "Lung Cancer Survival Rates." Retrieved from: http://www.webmd.com/lung-cancer/lung-cancer-survival-rates Retrieved on: 8 November 2013.

[5] D. Williams, et al., "Survival of patients surgically treated for stage I lung cancer," *The Journal of thoracic and cardiovascular surgery,* vol. 82, pp. 70-76, 1981.

[6] P. C. Pairolero, et al., "Postsurgical stage I bronchogenic carcinoma: morbid implications of recurrent disease," *The Annals of thoracic surgery,* vol. 38, pp. 331-338, 1984.

[7] T. Naruke, et al., "Prognosis and survival in resected lung carcinoma based on the new international staging system," *The Journal of thoracic and cardiovascular surgery,* vol. 96, pp. 440-447, 1988.

[8] W. A. Fry, et al., "Ten-year survey of lung cancer treatment and survival in hospitals in the United States," *Cancer,* vol. 86, pp. 1867-1876, 1999.

[9] C. F. Mountain, "Revisions in the international system for staging lung cancer," *Chest Journal,* vol. 111, pp. 1710-1717, 1997.

[10] Y. Sekido, et al., "Progress in understanding the molecular pathogenesis of human lung cancer," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer,* vol. 1378, pp. F21-F59, 1998.

[11] V. T. DeVita and R. Govindan, *Devita, Hellman, and Rosenberg's Cancer, Principles & Practice of Oncology Review*: Lippincott Williams & Wilkins, 2005.

[12] "Basic Histology -- Some Abnormally Large Nuclei."

[13] Z. Z. Htike, "Multi-horizon ternary time series forecasting," in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2013*, 2013, pp. 337-342.

[14] Z. Z. Htike, "Can the future really be predicted?," in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2013*, 2013, pp. 360-365.

[15] E.-E. M. Azhari, et al., "Brain Tumor Detection And Localization In Magnetic Resonance Imaging," *International Journal of Information Technology Convergence and services,* vol. 4, 2014.

[16] N. A. Mohamad, et al., "Bacteria Identification from Microscopic Morphology Using Naïve Bayes," *International Journal of Computer Science, Engineering and Information Technology,* vol. 4, 2014.

[17] E.-E. M. Azhari, et al., "Tumor Detection in Medical Imaging: A Survey," *International journal of Advanced Information Technology,* vol. 4, 2014.

[18] S. N. A. Hassan, et al., "Vision Based Entomology – How to Effectively Exploit Color and Shape Features," *Computer Science & Engineering: An International Journal,* vol. 4, 2014.

[19] N. A. Mohamad, et al., "Bacteria Identification from Microscopic Morphology: A Survey," *International Journal on Soft Computing, Artificial Intelligence and Applications,* vol. 3, 2014.

[20] S. N. A. Hassan, et al., "Vision Based Entomology: A Survey," *International Journal of Computer science and engineering Survey,* vol. 5, 2014.

[21] S. L. Win, et al., "Cancer Recurrence Prediction Using Machine Learning," *International Journal of Computational Science and Information Technology,* vol. 6, 2014.

[22] S. L. Win, et al., "Cancer Classification from DNA Microarray Gene Expression Data Using Averaged One-Dependence Estimators," *International Journal on Cybernetics & Informatics,* vol. 3, 2014.

[23] K. Shedden, et al., "Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study," *Nature medicine,* vol. 14, pp. 822-827, 2008.

[24] E. Urgard, et al., "Metagenes Associated with survival in non-small cell Lung cancer," *Cancer informatics,* vol. 10, p. 175, 2011.
[25] W. Ford, et al., "Classifying Lung Cancer Recurrence Time Using Novel Ensemble Method with Gene Network based Input Models," *Procedia Computer Science,* vol. 12, pp. 444-449, // 2012.
[26] J. Norris, et al., "A Novel Application for Combining CASs and Datasets to Produce Increased Accuracy in Modeling and Predicting Cancer Recurrence," *Procedia Computer Science,* vol. 20, pp. 354-359, 2013.
[27] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed.: The MIT Press, 2010.
[28] C. M. Bishop, *Pattern Recognition and Machine Learning*: Springer, 2007.
[29] "Exalpha." Retrieved from: https://www.exalpha.com/image-library/P51 Retrieved on: Retrieved on: 8 November 2013.
[30] Z. Z. Htike and S. L. Win, "Recognition of Promoters in DNA Sequences Using Weightily Averaged One-dependence Estimators," *Procedia Computer Science,* vol. 23, pp. 60-67, 2013.
[31] Z. Z. Htike and S. L. Win, "Classification of Eukaryotic Splice-junction Genetic Sequences Using Averaged One-dependence Estimators with Subsumption Resolution," *Procedia Computer Science,* vol. 23, pp. 36-43, 2013.
[32] H. Liu and R. Setiono, "Feature selection via discretization," *IEEE Transactions on knowledge and Data Engineering,* vol. 9, pp. 642-645, 1997.
[33] V. Bolón-Canedo, et al., "A combination of discretization and filter methods for improving classification performance in KDD Cup 99 dataset," presented at the Proceedings of the 2009 international joint conference on Neural Networks, Atlanta, Georgia, USA, 2009.
[34] U. M. Fayyad and K. B. Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," in *13th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, 1993, pp. pp. 1022-1029.
[35] S. Kotsiantis and D. Kanellopoulos, "Discretization techniques: A recent survey," *GESTS International Transactions on Computer Science and Engineering,* vol. 32, pp. 47-58, 2006.
[36] G. I. Webb, et al., "Not So Naive Bayes: Aggregating One-Dependence Estimators," *Machine Learning,* vol. 58, pp. 5-24, 2005.
[37] D. Hand and K. Yu, "Idiot's Bayes---Not So Stupid After All?," *International Statistical Review,* vol. 69, pp. 385-398, 2001.
[38] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Mach. Learn.,* vol. 29, pp. 103-130, 1997.
[39] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI-01 workshop on "Empirical Methods in AI"*.
[40] F. Zheng and G. I. Webb, "Efficient lazy elimination for averaged one-dependence estimators," presented at the Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, 2006.
[41] M. Sahami, "Learning Limited Dependence Bayesian Classifiers," in *Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 335-338.
[42] Y. Yang, et al., "Ensemble Selection for SuperParent-One-Dependence Estimators," in *AI 2005: Advances in Artificial Intelligence*. vol. 3809, S. Zhang and R. Jarvis, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 102-112.