CANCER RECURRENCE PREDICTION USING MACHINE LEARNING

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

Faculty of Engineering, IIUM, Kuala Lumpur, Malaysia

ABSTRACT

Cancer is one of the deadliest diseases in the world and is responsible for around 13% of all deaths world-wide. Cancer incidence rate is growing at an alarming rate in the world. Despite the fact that cancer is preventable and curable in early stages, the vast majority of patients are diagnosed with cancer very late. Furthermore, cancer commonly comes back after years of treatment. Therefore, it is of paramount importance to predict cancer recurrence so that specific treatments can be sought. Nonetheless, conventional methods of predicting cancer recurrence rely solely on histopathology and the results are not very reliable. The microarray gene expression technology is a promising technology that could predict cancer recurrence by analyzing the gene expression of sample cells. 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, from DNA microarray gene expression data, whether a particular cancer will recur within a specific timeframe, which is usually 5 years. To lower the computational complexity, we employ an entropy-based gene selection approach to select relevant prognostic genes that are directly responsible for recurrence prediction. This proposed system has achieved an average accuracy of 98.9% in predicting cancer recurrence over 3 datasets. The experimental results demonstrate the efficacy of our framework.

KEYWORDS

Cancer recurrence prediction; microarray gene expression

1. INTRODUCTION

Today cancer kills more people than AIDS, tuberculosis, and malaria combined [1]. According to the World Health Organization (WHO), cancer is a leading cause of death and responsible for around 13% of all deaths world-wide [2]. Cancer incidence rate is growing at an alarming rate. Despite the fact that cancer is preventable and curable in early stages, the vast majority of patients are diagnosed with cancer very late. Furthermore, it is not uncommon for cancer to come back after years of treatment. Cancer recurs because a tiny portions of cancer cells may remain undetected in the body after treatment. Over time, these cells may proliferate and grow large enough to be identified by conventional tests. Depending on the type of cancer, recurrence can occur weeks, months, or even many years after the primary cancer was treated. It is extremely difficult for physicians to know which cancer patients will experience recurrence. The likelihood that a cancer will recur and the likely timing and location of a recurrence depend on the type of the primary cancer. Some cancers have a predictable and distinguishable pattern of recurrence which can be picked up by pattern recognition and machine learning techniques. Therefore, a computerized cancer recurrence prediction system is required to prevent people from dying as a consequence of this unfortunate disease. Technically, cancer is a family of diseases that involve uncontrolled cell growth wherein cells divide and grow exponentially, generating malignant tumors and spreading to other parts of the body. The destructive power of the cancer is that it may not only spread to the neighboring tissues, but also to the whole body through the lymphatic

DOI : 10.5121/ijcsity.2014.2202

system or bloodstream. There are a few hundreds of known cancers found in humans [3]. Because there are an astronomical number of causes of cancer, researchers are still trying to understand the basis of cancer which still remain only partially understood. However, one thing that is apparent is that in order for a healthy cell to transmute into a cancer cell, the genes which regulate cell growth and differentiation must be modified [4]. It is known that cancers are caused by a chain of mutations in the genetic sequence. The development of a cancer cell is caused by a series of mutations which makes the cell proliferate more than its immediate neighbors by a process which transforms a normal healthy cell into a micro-invasive cell at the genetic level.

The nucleus of a human cell contains 46 chromosomes, each of which comprises a single linear molecule of deoxyribonucleic acid (DNA), which is intimately complexed with proteins in the form of chromatin [5]. DNA is the building block of life, which contains encoded genetic instructions for living organisms. A DNA is transcribed to become a precursor mRNA, which is then spliced to become an mRNA, which is in turn translated to become a protein. Because all the cells (except some) in a human body contain an identical set of genes, the expression level of each gene must differ from cell to cell. If we can somehow measure the expression levels of individual genes in a cell, we can use machine learning techniques to predict whether a cell is cancerous and what type of cancer it is. Fortunately, the DNA microarray technology allows researchers to measure expression levels of genes in a cell. A DNA microarray, also known as DNA chip, gene chip, gene array or biochip, is a densely packed array of identified DNA sequences attached to a solid surface, such as glass, plastic or silicon chip [6]. On a microarray chip, DNA fragments are attached to a substrate and then probed with a known gene or fragment. DNA sequences representing tens of thousands of genes are spotted or in situ synthesized on a very small slide. Microarray chips are scanned using an microarray scanner [7] and digitized on a computer. The scanner generates a 2D heat map, also known as, microarray image or microarray data. Therefore, DNA microarrays can be used to determine which genes are “turned on” (expressed) and which genes and “turn off” in a particular cell. They determine not only whether individual genes are expressed, but also the level at which these individual genes are expressed.

In this paper, we tackle the problem of recognizing cancer from DNA microarray gene expression data. During the past few decades, applications of pattern recognition and machine learning techniques have emerged in many domains [8-17]. 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 recurrence using machine learning techniques. Peterson et al. [18] applied old-fashioned artificial neural networks (ANNs) with back propagation to predict prostate cancer recurrence of patients after undergoing radical prostatectomy. After gene screening and optimization, they claimed to have achieved 0.99 to 1.0 diagnostic sensitivity and specificity. Ensemble techniques have recently become popular in cancer recurrence prediction. Ford et al. [19] proposed a General Regression Neural Network (GRNN) Oracle ensemble by combing several Partial least squares (PLS) models that were individually trained to predict lung cancer recurrence from 12 different gene networks. They concluded that it was possible to correctly classify recurrence by combining the results based on their proposed gene network models. Similarly, Norris et al.[20] applied the very same GRNN oracle to predict cancer recurrence. They confirmed that GRNN led to high prediction accuracy. Campbell et al. [21] applied the same GRNN oracle model to predict colon cancer recurrence. Lizuka et al. [22] applied Fisher linear classifier to predict recurrence of hepatocellular carcinoma after curative resection. Their system obtained an accuracy of 93% in predicting early intrahepatic recurrence. 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 recognizing cancer recurrence.

2. Cancer recurrence prediction

The aim of cancer recurrence prediction is to predict, given a set of gene expression data, whether or not a particular cancer will recur within a particular time frame. We proposed a three-layered framework that consists of entropy-based gene selection, entropy minimization discretization and prediction as shown in Figure 1. The complexity of any machine learning classifier depends upon the dimensionality of the input data [23]. There is also a phenomenon known as the 'curse of dimensionality' that arises with high dimensional input data [24]. In the case of genetic data classification, not all the genes in a genetic sequence might be responsible for predicting cancer recurrence. 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 the 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 recurrence prediction using the averaged one-dependence estimators with subsumption resolution (AODEsr). Section 2.3 describes the process of prediction.

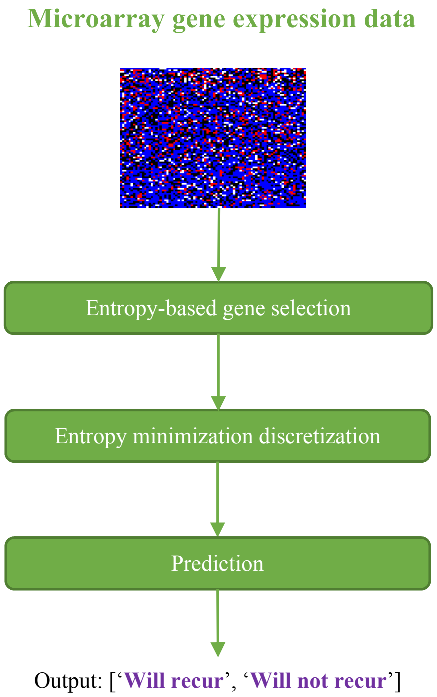

Microarray gene expression data
Entropy-based gene selection
Entropy minimization discretization
Prediction
Output: ['Will recur', 'Will not recur']

Figure 1. High-level flow diagram of cancer recurrence prediction framework.

2.1. Entropy-based gene selection

The complexity of any machine learning classifier depends upon the dimensionality of the input data [23]. 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 [23]. In the case of cancer recurrence prediction, not all the genes in a genetic sequence might be responsible for predicting cancer recurrence. 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 [25, 26]. In essence, we are interested in finding the best subset of the set of genes that can sufficiently predict cancer recurrence. 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 recurrence. There are generally two common techniques: forward selection and backward selection [23]. In forward selection, we start with an empty subset and add a gene (that increases the prediction accuracy the most) in each iteration until any further addition of a gene does not increase the prediction accuracy. In backward selection, we start with the full set of genes and remove a gene (that increases the prediction accuracy the most) in each iteration until any further removal of a gene does not increase the prediction accuracy. There are also other types of heuristics such as scatter search [27] and variable neighborhood search [28]. However, search-based gene selection techniques do not necessarily produce the best subset of the genes.

We employ a 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 [29]. 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 [30]. However, we have adopted EMD (Entropy Minimization Discretization) [31] 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 [32]. 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 [29]. 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 [31]. 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 [33-36]. There have been numerous applications of NB and variants thereof. The conventional NB algorithm uses the following formula for classification [37]:

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

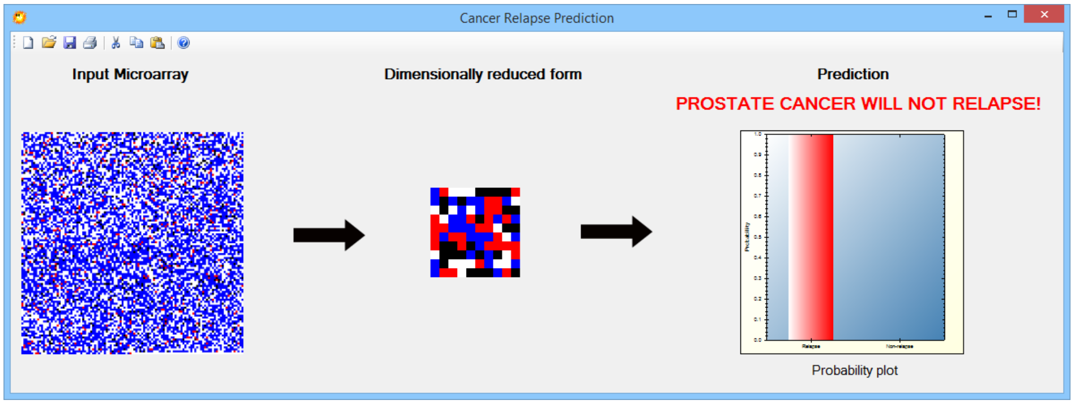
NB performs fairly accurate classification. The only limitation to its classification accuracy is the accuracy of the process of estimation of the base conditional probabilities. One clear drawback is its strong independence assumption which assumes that attributes are independent of each other in a 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) [38] and super parent one-dependence estimators (SPODEs) [39] 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 [33] 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 [33]. 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 [37]:

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

Semi-naive Bayesian classifiers attempt to preserve the numerous strengths of NB while reducing error by relaxing the attribute independence assumption [37]. Backwards sequential elimination (BSE) is a wrapper technique for attribute elimination that has proved to be effective at this task. Zheng et al. [37] 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 [37]. Because the AODEsr has a very weak independence assumption, it performs well in classification. Therefore, we employ an AODEsr classifier to predict cancer recurrence.

3. Experiments

The proposed framework was implement in C# 5.0 programming language using IKVM. Figure 2 illustrates a screenshot of the implemented cancer recurrence prediction system.



PROSTATE CANCER WILL NOT RELAPSE!

Figure 2. Screenshot of implemented cancer recurrence prediction system.

We tested our proposed system using 3 cancer recurrence datasets as listed in Table 1. Each dataset contains samples with more than 7000 genes. 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 1. Three cancer recurrence datasets used in our experiments.

Dataset	#Genes	#Samples
Breast cancer recurrence [40]	24481	78+19
Prostate cancer recurrence [41]	12600	21
CNS cancer recurrence [42]	7129	60

For each dataset, we performed one LOOCV experiment for varying number of selected genes ranging from 1 to 150. The genes for each trail were selected using the entropy-based technique outlined in Section 2.2. Figure 3 illustrates the results of our LOOCV experiments for each of the 11 datasets. The vertical axis represents the accuracy of the cancer recurrence predictor in percentage while the horizontal axis represents the number of selected genes. Table 2 lists the same set of results in a tabular format for a certain number of selected genes. The most surprising finding was that the system achieved a 100% accuracy in predicting prostate cancer recurrence for any number of selected genes. This implies that there is only one gene which acts as a prognostic biomarker for prostate cancer. In other words, using one gene is enough to predict whether prostate cancer will recur. The system achieved a 100% accuracy in predicting breast cancer with the number of genes higher than 88. This implies that the number of prognostic biomarkers are higher for breast cancer than for prostate cancer. The system obtained the lowest accuracy in predicting CNS cancer recurrence.

The results show that prediction accuracy does increase with the number of selected genes, albeit without perfect monotonicity. Results also show that at certain instances accuracy decreases with an increase in the number of genes. This may not be because of the classifier because AODEsr, like any other Bayesian classifiers, is not sensitive to irrelevant features. Therefore, adding an extra gene should not theoretically downgrade accuracy. The disruptions in monotonicity might be because of the intrinsic imperfection in the gene selection procedure.

Because our proposed system is able to predict cancer recurrence accurately even with a very few genes, the results reinforce the clinical belief that there are only a few prognostic biomarkers for cancer recurrence. The maximum LOOCV accuracy of our cancer classifier is 100% for two out of three datasets. The average maximum LOOCV accuracy of our cancer classifier across all the three datasets is 98.9%. It is worth iterating the fact that we used the AODEsr classifier with exactly 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 in order to prevent bias. To the best of your knowledge, the accuracy of the proposed cancer recurrence prediction system using the AODEsr classifier with the entropy-based selection process seems to be significantly higher than those of other cancer recurrence systems reported in the literature.

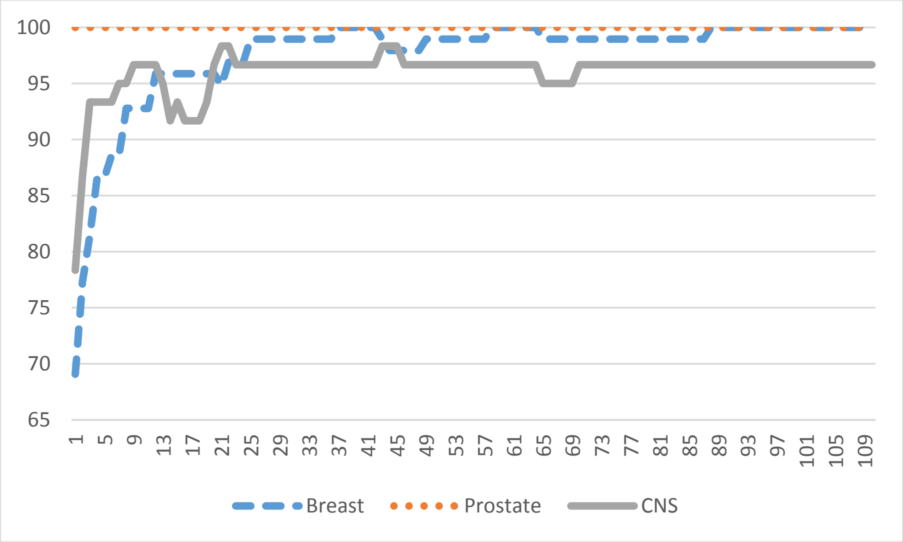


Figure 3. LOOCV accuracy (Y-axis) vs. number of genes (X-axis) [note: the plot maybe hard to read in monochrome print].

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

# of genes	Breast	Prostate	CNS
5	86.6	100	93.3
10	92.8	100	96.7
25	99.0	100	96.7
50	99.0	100	96.7
75	99.0	100	96.7
110	100	100	96.7

3. CONCLUSION

Many people succumb to cancer every day. Although cancer can be treated if detected early, cancer can recur after years of treatment. An automatic cancer recurrence prediction system is highly essential. We have presented a machine learning based approach to predict cancer recurrence from microarray gene expression data. Conventional naïve Bayes classifiers cannot accurately classify gene expression because of their unrealistic assumption that forbids dependencies among individual genes. 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 cancer recurrence. Given a set of gene expression data, the system predicts whether a particular cancer will recur within a particular timeframe. We have carried out experiments on three cancer datasets. This proposed system has achieved an accuracy of 98.9% in predicting cancer recurrence. The accuracy rate of the proposed system was found to be higher than those of other techniques. 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] *Cancer Prevention and Control*. Retrieved from: http://www.cdc.gov/cancer/dcpc/resources/features/worldcancerday/ Retrieved on: 15 November 2013.

[2] *Worldwide cancer statistics*. Retrieved from: http://www.cancerresearchuk.org/cancer-info/cancerstats/world/ Retrieved on: 15 November 2013.

[3] *How many different types of cancer are there?* Retrieved from: http://www.cancerresearchuk.org/cancer-help/about-cancer/cancer-questions/how-many-different-types-of-cancer-are-there Retrieved on: 15 November 2013.

[4] C. M. Croce, "Oncogenes and Cancer," *New England Journal of Medicine,* vol. 358, pp. 502-511, 2008.

[5] N. R. Colledge, et al., *Davidson's Principles and Practice of Medicine*, 21st ed.: Churchill Livingstone, 2010.

[6] M. M. R. Khondoker, "Statistical Methods for Preprocessing Microarray Gene Expression Data," Doctor of Philosophy, University of Edinburgh, 2006.

[7] *ArrayIT*. Retrieved from: http://shop.arrayit.com Retrieved on: 19 November 2013.

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

[9] S. L. Win, et al., "Gene Expression Mining for Predicting Survivability of Patients in Early Stages of Lung Cancer," *International Journal on Bioinformatics & Biosciences,* vol. 4, 2014.

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

[11] 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.

[12] 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.

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

[14] 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.

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

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

[17] 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.

[18] L. E. Peterson, et al., "Artificial neural network analysis of DNA microarray-based prostate cancer recurrence," in *Computational Intelligence in Bioinformatics and Computational Biology, 2005.* pp. 1-8.

[19] 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.

[20] 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.

[21] A. S. Campbell, et al., "Investigating the GRNN Oracle as a Method for Combining Multiple Predictive Models of Colon Cancer Recurrence from Gene Microarrays," *Procedia Computer Science,* vol. 20, pp. 374-378, 2013.

[22] N. Iizuka, et al., "Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection," *The Lancet,* vol. 361, pp. 923-929, 2003.

[23] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed.: The MIT Press, 2010.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*: Springer, 2007.

[25] 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.

[26] 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.

[27] F. García López, et al., "Solving feature subset selection problem by a Parallel Scatter Search," *European Journal of Operational Research,* vol. 169, pp. 477-489, 2006.

[28] M. García-Torres, et al., "Solving Feature Subset Selection Problem by a Hybrid Metaheuristic," presented at the First International Workshop on Hybrid Metaheuristics, 2004.

[29] H. Liu and R. Setiono, "Feature selection via discretization," *IEEE Transactions on knowledge and Data Engineering,* vol. 9, pp. 642-645, 1997.

[30] 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.

[31] 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.

[32] S. Kotsiantis and D. Kanellopoulos, "Discretization techniques: A recent survey," *GESTS International Transactions on Computer Science and Engineering,* vol. 32, pp. 47-58, 2006.

[33] G. I. Webb, et al., "Not So Naive Bayes: Aggregating One-Dependence Estimators," *Machine Learning,* vol. 58, pp. 5-24, 2005/01/01 2005.

[34] D. Hand and K. Yu, "Idiot's Bayes---Not So Stupid After All?," *International Statistical Review,* vol. 69, pp. 385-398, 2001.

[35] 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.

[36] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI-01 workshop on "Empirical Methods in AI"*.

[37] 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.

[38] M. Sahami, "Learning Limited Dependence Bayesian Classifiers," in *Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 335-338.

[39] 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.

[40] L. J. Van't Veer, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *nature,* vol. 415, pp. 530-536, 2002.

[41] D. Singh, et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer cell,* vol. 1, pp. 203-209, 2002.

[42] S. L. Pomeroy, et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature,* vol. 415, pp. 436-442, 2002.