A Novel Filter-Wrapper Based Feature Selection Approach for Cancer Data Classification

M.M. Mohamed Mufassirin[1,2]
[1]Department of Mathematical Sciences
South Eastern University of Sri Lanka
Sammanthurai, Sri Lanka
[2]Postgraduate Institute of Science, University of Peradeniya
mufassirin666@gmail.com

Roshan G. Ragel
Department of Computer Engineering
University of Peradeniya
Peradeniya, Sri Lanka
roshanr@pdn.ac.lk

Abstract* - The advancement in DNA microarray dataset technology has become an area of interest among many scholars. Application of this technology can be a great success for cancer data classification. However, DNA microarray data usually contains thousands of irrelevant and redundant gene information which need to be eliminated to improve the accuracy of classification. Thus, in order to select the relevant gene information from cancer data, a novel feature selection technique based on a filter-wrapper approach using machine learning methods is proposed in this study. Wrappers choose all possible subsets of features to evaluate which features are useful by using learning techniques and provide the most informative subset which will increase the accuracy of the classifiers whereas filter methods extract features from the data without any learning involved. However, compared to filters, the computation demand of wrappers are high when applied to cancer data. Hence, in the proposed work, the wrapper is applied after the filter approach with the intention of reducing the computational complexity of wrappers. The datasets were pre-processed initially using a filter called Gain Ratio Filter with the Ranker search method, and then the resultant gene subsets were evaluated using a wrapper called Wrapper Subset Evaluator with the best first forward selection searching strategy using the WEKA machine learning workbench. The selected gene subset by wrapper was then used to classify the cancer microarray using machine learning classifiers namely, Decision Tree (J48), Naïve Bayes, Sequential Minimal Optimization (SMO), Deep Learning and Bayes Net. The proposed approach was tested on five cancer microarray datasets. The accuracy of 89.69%, 95.16% and 97.04% were obtained for Breast, Colon and Lung cancer datasets respectively while Leukaemia and Ovarian cancer datasets scored 100%. According to the findings of this study, the proposed method is capable of accurately classify the dataset based on a few informative genes which is more efficient compared to existing classification models.

Keywords --DNA Microarray, Machine Learning, Feature Selection, Classification

I. Introduction

Machine learning (ML), the ability of machines to learn without being explicitly programmed, has proved to be promising in solving real-world problems. It consists of the making of algorithms that can learn from and make predictions on data fed into it [10]. Machine learning methods can be divided into two core categories namely supervised and unsupervised learning. In supervised learning, a labelled set of training data is used to estimate or map the input data to the desired output. In contrast, under the unsupervised learning methods, no labelled examples are provided, and there is no learning process. As a result, the pattern recognition or discovery is up to the learning model. This practice is considered as a classification problem in supervised learning [7]. Machine learning applications can be seen in various areas like Bioinformatics, Cheminformatics, Computer Networks and many more.

Cancer, a common term in the healthcare sector is the second leading cause of death globally and has caused 8.8 million deaths in 2015 (according to WHO statistics) [24]. Being of several types, cancer now poses a significant threat to human health than any other disease. Whatever the type it was thought to be an incurable disease, in modern days several treatments are being discovered to increase the longevity of a patient's life or to fight certain cancer types. As prevention is better than cure, it is better to know if any person has or might be facing the threat of cancer later in his life. Therefore, in many ways it is the challenge now to try to predict early in the lifespan whether someone will have cancer later in life, whether someone has cancer at present, also to know whether she/he will again suffer from the same cancer even after the cure. Early diagnosis of cancer helps for successful treatment, and can lower the fatality rate [24].

DNA microarray technology has been applied to do prognosis and classification of cancer accurately. Due to the extreme and sparse characteristics of microarray gene expression, the analysis of microarray data is very challenging. Selection of informative gene subset among thousands of genes is also challenging. However, by analysing these microarray gene expression datasets, heterogeneous cancer can be classified and grouped into their appropriate subgroups [5]. Even though many machine learning techniques such as Support Vector Machines, k-Nearest Neighbor (kNN), Decision Trees (DT) and several neural network techniques are used to discover informative knowledge from microarray data, it is difficult to handle a large number of genes thus the accuracy becomes inappropriate. Therefore, feature selection is used in cancer classification which eliminates irrelevant and redundant genes. Identifying a smallest and most informative subset of genes for accurate classification is the goal of feature selection [29-30].

The primary purpose of proposed study is to discuss few selected machine learning methods used in many types of research and discuss the advancements in the field of cancer research using ML methods, also to address a modern method of feature selection approach that can improve the predictive outcomes. Here, certain datasets were chosen, and the results were judged using statistical measures. Compared to existing methods, the proposed approach shows a superior classification accuracy and performance when evaluated using five benchmark cancer datasets based on a few informative genes.

II. Related Work

Physicians and medical experts in cancer can benefit from the mined abstract tumour attributes by better understanding the properties of different types of tumours [2]. The results in [2] show the capability of diagnosis and time saving during the training phase. Different kinds of machine learning and statistical approaches are used to classify tumour cells [5].

According to the better designed and validated studies, machine learning methods have proved to significantly (15-25%) improve the accuracy of predicting cancer vulnerability, mortality, and recurrence [7]. Even though some improvement has been achieved, there are still many challenges lasting and directions for further research, such as emerging improved classification algorithms and integration of classifiers to reduce false positives [1]. Using automated computer tools and in particular machine learning to facilitate medical analysis and diagnosis is a promising area [3]. The idea of using ensemble classifiers is that the outcomes are less reliant on particularities of a single training dataset and because the ensemble model beats the performance and outcome of the best base classifier in it [8]. In one study [4], SSL was proved to be the best among the ANN and SVM, and the differences in performance were statistically significant.

The feature selection methods were broadly used in many researches for the purpose of reducing effects from noise or irrelevant features in order to provide good prediction results [11-12]. The Wrapper methods use learning techniques to evaluate feature subsets which provide better results than filter methods. But wrapper approaches increase the computational cost [9, 51]. In reference [13], the authors have done a comparison study on two techniques of integrating feature selection and ensemble learning, (1) Feature selection for ensemble learning (ENfs) and (2) Ensemble learning for feature selection (FSen). This approach has given a high predictive accuracy than the conventional feature selection methods for supervised machine learning. Moreover, it also gives a better mechanism to efficiently handle stability issue that is usually poor in existing conventional feature selection methods.

The researchers proposed a new wrapper method in [26], called Incremental ANOVA and Functional Networks-Feature Selection (IAFN-FS) for dealing with complex classification problems based in classical algorithms, such as Decision Tree and Naïve Bayes. This approach obtained better results in terms of the accuracy, while it has the shortcoming of picking up a higher number of features for a subset. In reference [27], the authors used a “rotation forest ensemble decision tree technique” and wrapper approach with best first search strategy. The intention of using is to select the optimum and more informative gene subset on the Erythemato-Squamous diseases dataset using forward selection. The refinement ability of selected features/ attributes is evaluated using numbers of machine learning algorithms, and the bagging algorithm was used to assess the diversity of the training data.

Also hybrid methods like Aco-SVM, k-SVM, Hybrid BN can improve the prediction performance significantly. Hybrid Bayesian Networks mentioned in [6] show high accuracy, specificity and the sensitivity of 87.2 %, 0.831, 0.933 respectively which is better in comparison to BN and ANN used for the same breast cancer dataset. Hybrid methods have proved to be very much accurate like k-SVM methodology, a hybrid of ANN and SVM gives the accuracy of 97.38% when trained and tested on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Fast Library for Approximate Nearest Neighbors (FLANN) alone shows 63.4% accuracy whereas PSO-FLANN gives better classification rate of 92.36% [5]. Most machine learning methods like ANN, SVM, Decision tree and Bayes Network give 70-96% accurate results for specific types of cancer [7]. Different machine learning methods and their performance in cancer research obtained from the existing studies are shown in Table II.

III. Materials and Methods

In this study, we have considered mainly four steps for proceeding methods namely DNA dataset pre-processing, feature selection, classification and statistical methods for evaluation. The proposed work is based on pattern discovery where reliable tools (different machine learning methods) are used.

A. Experiment Setup

The dataset pre-processing was done by defining the attributes and filling the missing values with average value of the respective gene. Five cancer microarray datasets were used in this experiment namely Colon cancer, Breast cancer, Lung cancer, Leukaemia and Ovarian cancer. The descriptions of the datasets used for this study are shown in Table I. WEKA (version 3.8.2) machine learning workbench was used in this study for model construction.

TABLE I. Descriptions of Dataset Used for the Experiment

Name of Datasets (Cancer)	No. of genes	No. of Instances	No. of Classes	Description of Classes	Source Available
Colon	2000	62	2	40 Tumor and 22 Normal	[14]
Breast	24481	97	2	Relapse – 46, Non-relapse- 51	[15]
Lung	12600	203	5	1 – 139, 2 – 17, 3 – 6, 4 – 21 & 5 - 20	[16]
Leukaemia	7129	72	2	47 ALL and 25 AML	[17]
Ovarian	15154	253	2	162 – Cancer and 91 – healthy control	[18-19]

TABLE II. DIFFERENT MACHINE LEARNING METHODS AND THEIR PERFORMANCE IN CANCER RESEARCH OBTAINED FROM THE LITERATURE SURVEY [MISSING INFORMATION HAVE BEEN SHOWN WITH "-"]

Methods	Types of cancer					Reference Number	Accuracy (%)	Sensitivity	Specificity	Year of publication
	Colon	Breast	Oral	Basal Cell	Lung					
Decision Tree (DT) (J48/C4.5)		🟧				[35]	93.6	0.958	0.907	2013
						[7]	93.0	-	-	2015
Artificial Neural Network (ANN)				🟥		[6]	88.8	0.885	0.854	2009
		🟧				[36]	95.27	-	-	2016
		⬛				[4]	65.0	0.73	0.58	2013
					🟧	[7]	83.5	-	-	2015
		🟩				[35]	94.7	0.956	0.928	2013
Support Vector Machine (SVM)		🟦				[1]	64.6	1	0.645	2010
						[4]	51.0	0.65	0.52	2013
						[35]	95.7	0.971	0.945	2013
						[7]	69.0	-	-	2015
			🟫			[7]	75.0	-	-	2015
		🟩				[36]	97.13			2016
Bayesian Network			🟧			[7]	100	-	-	2015
				🟦		[6]	70.9	0.885	0.583	2009
Deep Learning				🟩		[34]	92.1	0.887	0.941	2013
		🟦				[3]	63.33	-	-	2013
	⬛					[3]	66.67	-	-	2013
		🟦				[3]	86.67	-	-	2013
	⬛					[3]	66.67	-	-	2013
Semi-supervised Learning		🟧				[4]	71.0	0.76	0.65	2013
		🟦				[7]	80.7	-	-	2015
	🟩					[7]	76.7	-	-	2015

B. *Methodology*

In this work, a novel feature selection approach is proposed as summarised in Fig. 1. The dataset was pre-processed initially using a filter called gain ratio filter (GainRatioAttibuteEval) with Ranker search method and then the resultant gene subset was evaluated using a wrapper called Wrapper Subset Evaluator (WrapperSubsetEval). Filters can be used to reduce dimensionality and overcome overfitting. But, the major problem in the filter approach is the computation of a threshold by which features may be discarded from ranking [32]. One heuristic approach is known as the (n-1) rule in microarray analysis where n denotes the number of instances chooses the top (n-1) genes to start the analysis [20, 22-23]. The analysis is initiated according to (n-1) rule, and gene subset is selected using gain ratio filter. Gain ratio filter evaluates the worth and important of an attribute in classification by measuring the gain ratio with regard to the relevant class. The wrapper subset evaluator is then applied to the resultant gene subset. In wrapper approach different machine learning classifiers (like Naïve Bayes/ Deep Learning) along with best first forwarding selection searching strategy can be used to evaluate all possible gene subsets and ultimately provides the best informative gene subset which does well [21]. The selected gene subset is then used to classify the cancer samples using different classifiers namely, Naïve Bayes, Decision Tree (J48), Sequential Minimal Optimization (SMO), Deep Learning and Bayes Net.

Evaluating a feature subset of genes in machine learning method is done by an internal validation method called k-fold cross-validation [25]. 10-fold cross-validation was used here. The main reason for the selection is that the estimator of k-fold cross-validation has a lower adjustment than a single hold-out set estimator, which will be very imperative if the available data size is limited. In 10-fold cross-validation, the original cancer genes sample was randomly partitioned into ten equal size subsamples. Among that subsamples, a single subsample was taken as the validation data for testing the model, and the remaining 9 (k-1) subsamples were used as training data. The cross-validation procedure was then repeated ten times with each of the ten subsamples where an exactly one subsample was treated as validation data. The average of ten results from the folds was used to produce a single estimation for the dataset. The benefit of this technique over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used exactly once for validation. Here k remains an unfixed parameter [25, 28].

IV. Results and Discussion

There are five classifiers, namely J48 Decision Tree (DT), Naïve Bayes, Sequential Minimal Optimization (SMO), Deep Learning and Bayes Network were used to analyse each cancer dataset. For each machine learning methods/ classifiers as the performance measure, the accuracy, precision and sensitivity were calculated during classification using the WEKA machine learning tool with and without feature selection. The calculated values were used to compare and rank the classifiers. For simplicity, the accuracy of classifiers was used as the primary measure for comparing classifiers. It is assumed that perfect method should give 100% accuracy, 100% specificity, and 100% sensitivity. Deciding the best method depends on many factors like the size of the dataset, missing values in it and the system used to assess the method. However, here we have maintained the same environment for all classifier.

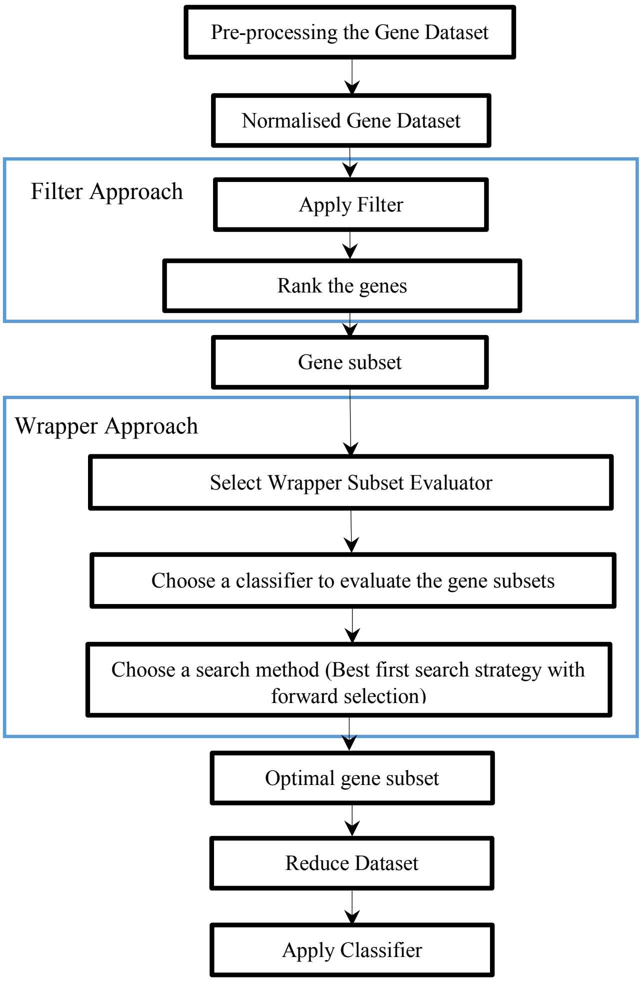

Pre-processing the Gene Dataset
Normalised Gene Dataset
Filter Approach
Apply Filter
Rank the genes
Gene subset
Wrapper Approach
Select Wrapper Subset Evaluator
Choose a classifier to evaluate the gene subsets
Choose a search method (Best first search strategy with forward selection)
Optimal gene subset
Reduce Dataset
Apply Classifier

Fig. 1. Schematic illustration of proposed method

In addition to these measures, we have also considered the time taken to build the model. The comparisons of average run time (in second) to build the model with and without feature selection for different datasets are given in Table VIII. However, it can be noticed that other performance measures are higher than those of the rest of the methods. Time taken to build the model is less significant than other measures in deciding for the best; also time for the same method may vary.

The performances of Colon cancer, Leukaemia, Breast cancer, Lung cancer and Ovarian cancer classification with and without gene selection are given in Table III, Table IV, Table V, Table VI and Table VII respectively.

TABLE III. Improvement of the Accuracy During Feature Selection for Colon Cancer

Classifier/ Method	Accuracy without Feature Selection (%)	Accuracy with Feature Selection (%)	Improvement of Accuracy (%)
J48	82.26	90.92	**8.66**
Naïve Bayes	53.23	87.1	**33.87**
SMO	85.48	88.71	**3.23**
Deep learning	75.81	87.1	**11.29**
Bayes Net	75.81	95.16	**19.35**

TABLE IV. Improvement of the Accuracy During Feature Selection for Leukaemia Cancer

Classifier/ Method	Accuracy without Feature Selection (%)	Accuracy with Feature Selection (%)	Improvement of Accuracy (%)
J48	84.21	97.37	**13.16**
Naïve Bayes	97.74	100	**2.26**
SMO	94.74	100	**5.26**
Deep learning	92.11	100	**7.89**
Bayes Net	94.74	100	**5.26**

TABLE V. Improvement of the Accuracy During Feature Selection for Breast Cancer

Classifier/ Method	Accuracy without Feature Selection (%)	Accuracy with Feature Selection (%)	Improvement of Accuracy (%)
J48	62.89	84.54	**21.65**
Naïve Bayes	54.64	89.69	**35.05**
SMO	68.04	86.6	**18.56**
Deep learning	68.04	79.38	**11.34**

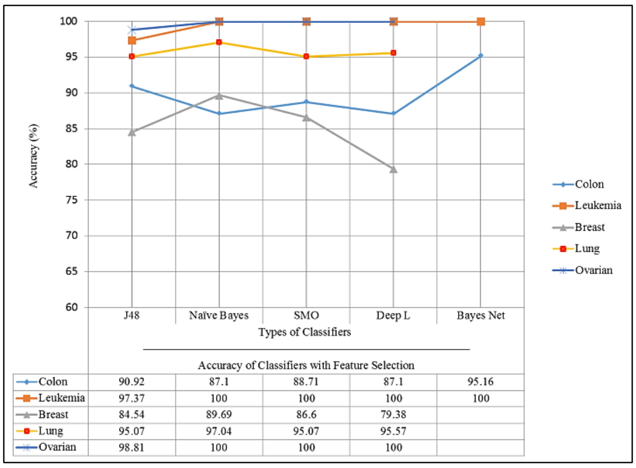
TABLE VI. Improvement of the Accuracy During Feature Selection for Lung Cancer

Classifier/ Method	Accuracy without Feature Selection (%)	Accuracy with Feature Selection (%)	Improvement of Accuracy (%)
J48	93.1	95.07	**1.97**
Naïve Bayes	80.79	97.04	**16.25**
SMO	95.57	95.07	**-0.5**
Deep learning	90.15	95.57	**5.42**

TABLE VII. IMPROVEMENT OF THE ACCURACY DURING FEATURE SELECTION FOR OVARIAN CANCER

Classifier/ Method	Accuracy without Feature Selection (%)	Accuracy with Feature Selection (%)	Improvement of Accuracy (%)
J48	95.65	98.81	**3.16**
Naïve Bayes	92.49	100	**7.51**
SMO	100	100	**0**
Deep learning	98.42	100	**1.58**

In Fig. 2 the highest accuracies of the different machine learning methods with feature selection for different cancers are shown. According to Fig. 2, the proposed approach has given higher performance. 95.16% accuracy for colon cancer, 89.69% accuracy for breast cancer, 97.04% accuracy for lung cancer and 100% accuracy for leukemia and ovarian cancer were obtained during the proposed approach. The increase in classification accuracy implies that the original datasets consist of redundant and irrelevant genes which lead to lack of classification accuracy without feature selection. We have considered the results obtained from two recent studies based on feature selection methods to compare our proposed method. Table IX shows the comparisons of the accuracy of the proposed method with the existing feature selection models for different cancer. The results show that the proposed method has significant improvement in cancer classification.



	J48	Naïve Bayes	SMO	Deep L	Bayes Net
Colon	90.92	87.1	88.71	87.1	95.16
Leukemia	97.37	100	100	100	100
Breast	84.54	89.69	86.6	79.38	
Lung	95.07	97.04	95.07	95.57	
Ovarian	98.81	100	100	100	

Fig. 2. The accuracy of machine learning methods with feature selection

In this study, the Colon cancer dataset was further analysed for validating our results from the medical point of view. In the Colon cancer dataset, the selected top four informative genes with highest gain ratio values were *TGFBR2, CSRP1, GUCA2B and MYL9*. Out of these four genes, *TGFBR2* gene has been reported in human Colon tumours as it is found in 97% of tumour samples in the mutated gene position [37]. However, there is no any direct indication presenting that those other three genes were associated with Colon tumours [38]. The details description of the selected genes are given in Table X. The roles of the other genes need to be further investigated to identify the cancers.

TABLE VIII. THE COMPARISON OF AVERAGE RUNS TIME (IN SECOND) WITH AND WITHOUT FEATURE SELECTION

Dataset	Without Feature Selection (s)	Proposed Method (With Feature Selection) (s)
Colon	2.480	0.034
Leukaemia	0.138	0.028
Breast	0.778	0.067
Lung	1.358	0.55
Ovarian	1.895	0.085

TABLE IX. COMPARISONS OF THE ACCURACY OF THE PROPOSED METHOD WITH THE EXISTING FEATURE SELECTION MODELS

Dataset	**Proposed Method** (2018)	**Reference [31]** (M. Morovvat – 2016)	**Reference [33]** (Q. Su – 2017)
Colon	**95.16**	90.32	90.1
Leukaemia	**100**	**100**	79.6
Breast	89.69	**94.84**	87.4
Lung	**97.04**	96.55	90.1
Ovarian	**100**	**100**	98.5

TABLE X. DESCRIPTION OF THE COLON CANCER INFORMATIVE GENES

Datasets	Selected Informative Genes	Descriptions
Colon Cancer	*TGFBR2*	transforming growth factor-beta (TGF-β) receptor type 2
	CSRP1	Cysteine and glycine rich protein 1
	MYL9	Myosin light chain 9
	GUCA2B	Guanylate cyclase activator 2B

V. CONCLUSION

In this study, a modern filter-wrapper based feature selection approach is suggested to select more informative gene subsets for cancer classification. Wrappers take all the possible combinations of gene subsets and eventually select the best subset, which performs well for a given classifier. Thus, it requires a huge time for processing and becomes more challenging when applied to cancer microarray data directly. Therefore, in order to overcome this issue a filter based pre-processing is carried out initially before using a wrapper. Gain ratio is a filter method that can effectively eliminate irrelevant and redundant genes from microarray cancer datasets. Wrappers Subset Evaluator is a wrapper method that considers the inter gene interactions which would provide more informative knowledge for accurate classification. Five machine learning classifiers were used in this study for classification with a validation technique called 10-fold cross validation. Our proposed method was tested on 5 cancer microarray datasets. The tested results indicate that the proposed approach has acceptable level of performance in terms of accuracy and time efficiency compared to the existing methods.

REFERENCES

[1] A. Menendez, and F. D. Cos Juez, "Artificial neural networks applied to cancer detection in a breast screening programme," Mathematical and Computer Modelling, vol. 52, pp. 983-981, 2010.

[2] B. Zheng, et al., "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms". Expert Systems with Applications, 2013.

[3] R. Fakoor, F. Ladhak, and A. N. Azade, "Using deep learning to enhance cancer diagnosis and classification. Computer Science and Engineering Dept, the University of Texas at Arlington, Arlington, 2013, TX 76019 USA

[4] K. Park et al., "Robust predictive model for evaluating breast cancer survivability", Engineering Applications of Artificial Intelligence, vol. 26, pp. 2194–2205, 2013.

[5] S. Agrawal, and J. Agrawal, "Neural Network Techniques for Cancer Prediction: A Survey", 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science, vol. 60, pp. 769 – 774, 2015.

[6] J. P. Choi1, T. H. Han, and R. W. Park, "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", J Kor Soc Med Informatics, vol. 15, pp. 49-57, 2009.

[7] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015.

[8] A. Bashiri, M. Ghazisayedi, R. Safdari, L. Shahmoradi, and H. Ehtesham, **"**Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review", Iran J Public Health, Vol. 46, No.2, Feb 2017, pp.165-172

[9] Z. M. Hira, and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," vol. 2015, no. 1, 2015.

[10] M. Esposito, K. Bheemaiah, and T. Tse, "What is machine learning?," Grenoble École de Management (GEM), 04 05 2017. [Online]. Available: http://theconversation.com/what-is-machine-learning-76759. [Accessed 23.06.2017].

[11] G. Chandrashekar, and F. Sahin, "A survey on feature selection methods", Computer and Electrical Engineering; 2014, pp.16-28.

[12] V. Kumar and S. Minz, "Feature Selection: A literature Review", Smart Computing Review; 2014.

[13] D. Guana , W. Yuana, Y.K. Leea, K. Najeebullaha, and M.K. Rasela, "A Review of Ensemble Learning Based Feature Selection", IETE Technical Review; 2014.

[14] Colon Cancer, "Bioinformatics Research Group (Dataset Repository in ARFF (WEKA))", [Online] Available: http://eps.upo.es/bigs/datasets.html [Accessed: Aug. 2017].

[15] Breast Cancer Data Set, [Online] Available at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer , [Accessed on: 10.10.2017]

[16] Lung Cancer Data Set, [Online] Available at: https://archive.ics.uci.edu/ml/datasets/lung+cancer [Accessed on: 10.10.2017]

[17] T. Golub et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, (1999) 286, pp. 531–536.

[18] E. Petricoin, et al., "Use of protomic patterns in Serum to Identify Ovarian cancer" The Lancet, 359: pp. 572-577, February 2002.

[19] Ovarian Cancer dataset, [Online] http://mldata.org/repository/data/viewslug/ovarian-cancer-nci-pbsii-data/ [Accessed on: 10.02.2018]

[20] Y. Wang et al., "Gene selection from microarray data for cancer classification - A machine learning approach," Comput. Biol. Chem., vol. 29, no. 1, pp. 37–46, 2005.

[21] Introduction to Feature Selection, [Online] Available at: https://machinelearningmastery.com/an-introduction-to-feature-selection/, [Accessed on: 10.02.2018]

[22] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection," Int. J. Inf. Technol. Knowl. Manag., vol. 2, no. 2, pp. 271–277, 2010.

[23] V. B. Canedo, N. S. Maroño, A. A. Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," Inf. Sci. (Ny)., vol. 282, pp. 111–135, 2014.

[24] L. A. Torre, F. Bray, R. L. Siegel, and J. Ferlay, "Global Cancer Statistics , 2012," vol. 65, no. 2, pp. 87–108, 2015.

[25] R. Panthong and A. Srivihok, "Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm," Procedia Comput. Sci., vol. 72, pp. 162–169, 2015.

[26] N. S´anchez-Maro˜no, A. Alonso-Betanzos, and R.M. Calvo-Est´evez, "A Wrapper Method for Feature Selection in Multiple Classes Datasets", J. Cabestany et al. (Eds.): IWANN; 2009, pp. 456–463.

[27] A. Ozcift, and A. Gulten, "A Robust Multi-Class Feature Selection Strategy Based on Rotation Forest Ensemble Algorithm for Diagnosis", J Med Syst; 2012, pp.941–949.

[28] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," Expert Syst. Appl., vol. 41, no. 4 PART 2, pp. 1937–1946, 2014.

[29] T. H. Dang, T. D. Pham, H. L. Tran, and Q. Le Van, "Using dimension reduction with feature selection to enhance accuracy of tumor classification," BME-HUST 2016 - 3rd Int. Conf. Biomed. Eng. IEEE, pp. 14–17, 2016.

[30] D. Pavithra and B. Lakshmanan, "Feature selection and classification in gene expression cancer data," 2017 Int. Conf. Comput. Intell. Data Sci., pp. 1–6, 2017.

[31] M. Morovvat and A. Osareh, "An Ensemble of Filters and Wrappers for Microarray Data Classification," Mach. Learn. Appl. An Int. J., vol. 3, no. 2, pp. 1–17, 2016.

[32] M. N. F. Fajila and R. D. Nawarathna, "New Feature Selection Method for High Dimensional Gene Data," in Symposium on Statistical & Computational Modelling with Applications - 2016, 2016, pp. 66–69.

[33] Q. Su, Y. Wang, X. Jiang, F. Chen, and W. Lu, "A Cancer Gene Selection Algorithm Based on the K-S Test and CFS," Biomed Res. Int., vol. 2017, pp. 1–6, 2017.

[34] A.A. Cruz-Roa1, J.E.A. Ovalle, A. Madabhushi, and F.A.G. Osorio, "A Deep Learning Architecture for Image Representation, Visual Interpretability, and Automated Basal-Cell Carcinoma Cancer Detection", MICCAI 2013, Part II, LNCS 8150, pp. 403–410, 2013. _c Springer-Verlag Berlin Heidelberg 2013

[35] L.G. Ahmad, A.T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi and A.R. Razavi, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence", J Health Med Inform, 2013, vol. 4, pp. 124

[36] D. L. Poole, and A. K. Mackworth, Artificial Intelligence, 2017, Page-354

[37] J. S. Michael et al., "A multigene mutation classification of 468 colorectal cancers reveals a prognostic role for APC," Nature Communications, vol. 7, pp. 1-12, 2016.

[38] G. Lingyun, Y. Mingquan, L. Xiaojie and H. Daobin, "Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification," Genomics Proteomics Bioinformatics, vol. 15, pp. 389-395, 2017.