Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Heterogeneous Data Fusion and Selection in High-volume Molecular and Imaging Datasets

Konstantinos Moutselos, Ilias Maglogiannis
Department of Computer Science and Biomedical Informatics University of Central Greece
Lamia, Greece
kmouts@ucg.gr, imaglo@ucg.gr

Aristotelis Chatziioannou
Institute of Biology, Medicinal Chemistry and Biotechnology
National Hellenic Research Foundation
Athens, Greece
achatzi@eie.gr

Abstract—**In this work, two disparate datasets, concerning the study of the same physiological type of cutaneous melanoma but derived from different donors, one of image (dermatoscopy) and the other of molecular (trascriptomic expression) origin are utilized, so as to form an expanded in description depth, integrative dataset. Four different imputation methods are employed in order to derive the unified dataset, prior the application of backward selection together with ensemble classifiers (random forests). The various imputation schemes applied, manage to emulate the effect of biological noise on the unified dataset, adding realistic signal variation. Thus, they immunize the discovery process in the integrative dataset, from false positive artifacts, which do not have a true differential effect. The results suggest that the expansion of the feature space through the data integration and the exploitation of elaborate imputation schemes in general, aid the classification task, imparting stability as regards the derivation of the putative classifiers.**

Keywords- data integration; feature selection; random forest; biomarker inference; cutaneous melanoma

I. INTRODUCTION

Integration of multi-modal and multi-scale data is of known importance in the context of personalized medicine and the electronic health records. In the context of Virtual Physiological Human (VPH), an integrated framework should promote the interconnection of predictive models pervading different scales, with different methods, characterized by different granularity, to form topological networks, which consolidate system level information in a holistic context [1].

Information-fusion algorithms can be distinguished to those based in combination of data (COD) or combination of interpretations (COI). COD methods fuse features from each source into a single feature vector prior the classification, while COI methods classify the data from each source independently and then aggregate the results. Rohlfing et. al [2] compared the two methods to combine information sources in different biomedical image analysis applications, while Haapanen and Tuominen [3] followed a COD approach for the combination of satellite image and aerial photograph features for forest variable estimation. On the other hand, Jesneck et al [4], on a COI path, optimized a decision-fusion technique to combine heterogeneous breast cancer data. Lee et al [5], proposed a Generalized Fusion Framework (GFF) for homogenous data representation and subsequent fusion in the meta-space using dimensionality reduction techniques. Such meta-space representation approaches, which transform data into a homogeneous space allowing for direct fusion of heterogeneous data, are embedding projections and kernel space projections [6].

GFF algorithms assume that we have raw data from sources $S_i(x_1, x_2, ...x_k)$, where $x_1, x_2, ...x_k$ represent the k observations in a study and i represents one of the N data sources, $i \in \{1,2,...,N\}$. While this could be the case for specific studies or electronic patient records, most available databases contain data from a single layer of dissection and for different patients. The number of observations from each source S_i differs. Nonetheless, the datasets formed from each source encapsulate information regarding the same disease, and it is an open question how these interconnections could be exploited.

In this work, two disparate datasets, concerning the study of the same physiological type of cutaneous melanoma, one of image (dermoscopy) and the other of molecular (trascriptomic expression) origin, are utilized to compose an integrative dataset. We refer to these data as *separate* datasets. Moreover, the unified dataset is formed through the application of various imputation methods and further subjected to processing by a feature selection algorithm (random forests). Integration of separate datasets regarding the same disease can be a promising avenue, which may ultimately promote the derivation of reliable composite biomarkers, for various diseases.

A. Cutaneous Melanoma

Cutaneous melanoma (CM) is considered a complex multigenic and multifactorial disease that involves both environmental and genetic factors. It represents the most aggressive and lethal skin neoplasm, and its incidence and mortality have been increasing worldwide. CM tumorigenesis is often explained as a progressive transformation of normal melanocytes to nevi that subsequently develop into primary cutaneous melanomas (PCM). However, the molecular pathways involved have not yet been clearly elucidated [7].

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

Despite successes in the definition of genomic markers or gene signatures for other kinds of cancers (such as breast cancer), there no such progress related to malignant melanoma has been noted.

The microarray studies that have been performed on CM by different groups so far, employ different microarray platforms in highly heterogeneous patient cohorts and pathological sample collections [8]. These differences perplex in-between comparisons and result in a reduced cohort size and phenotypic diversity for the unified datasets, since independent cohorts from different studies are hard to be integrated [9].

Regarding the clinical diagnostic methods for diagnosis of melanoma, there are several standard approaches for analysis and diagnosis of lesions. For example the Menzies scale, the Seven-point scale, the Total Dermoscopy Score based on the ABCD rule, and the ABCDE rule (Asymmetry, Border, Color, Diameter, Evolution). In these methods, digital images can serve as a basis for the medical analysis and diagnosis of lesions under consideration. As there is a general lack of precision in human interpretation of image content, advanced computerized techniques can assist doctors in the diagnostic process [10]. A review of image acquisition and feature extraction methods utilized in the literature regarding existing classification systems can be found in [11].

B. Feature selection

Feature selection techniques do not alter the original representation of the variables, but merely select a subset of them, in contrast to other dimensionality reduction techniques like those based on projection (e.g. principal components analysis) or compression (e.g. using information theory). Thus, they preserve the original semantics of the variables, offering the advantage of interpretability by a domain expert [12].

In this work, the feature selection that was applied, concerns a wrapper type technique (sequential backward elimination) exploiting the random forest algorithm [13], an ensemble classifier which utilizes ensembles of decision trees. In addition, a multivariate filter was used as an option to reduce the co-linearity among features of the microarray dataset, prior to the application of the wrapper method. This filtering together with the imputation procedure, is a departure from a merely COD method, towards a GFF approach, although here no further transformation is applied to the feature vectors.

The random forest algorithm, among other ensemble learning methods, is reported to be successful in variance reduction, which is associated with overfitting [14]. In addition, we utilized the option of stratifying the bootstrapped samples with equal number of cases per class [15]. This is compatible with the Balanced Random Forest (BRF) approach, which is computationally more efficient with large imbalanced data, since each tree only uses a small portion of the training set to grow. Additionally, it is less vulnerable to noise (mis-labeled class) than the Weighted Random Forest (WRF), where a heavier penalty is placed on misclassifying the minority class [16]. BRF alleviated the class imbalance problem, which represents a fundamental problem in disease diagnosis, where the disease cases are rare as compared with normal populations.

II. Materials & Methods

A. Multi-Modal Data Fusion of Separate Datasets

1) Image data

The dataset derived from skin lesion images contained 972 instances of nevus skin lesions and 69 melanoma cases. Three types of features are analyzed: Border Features which cover the A and B parts of the ABCD-rule of dermatology, Color Features which correspond to the C rules and Textural Features, which are based on D rules. The total number of features assessed was 31 from the initial set of 32 (one feature was removed as having zero variation across the samples). The relevant pre-processing steps applied at all features, is described in [17].

2) Microarray data

The respective microarray dataset data was selected from the Gene Expression Omnibus (GEO) [18], GDS1375. In that experiment, total RNA isolated from 45 primary melanoma, 18 benign skin nevi, and 7 normal skin tissue specimens were analyzed on an Affymetrix Hu133A microarray containing 22,000 probe sets [19]. The dataset contains the values of MAS5-calculated signal intensities after global scaling the average intensity to 600.

The data retrieval from GEO was performed using GEOquery [20] and processed with limma [21] R packages from the Bioconductor project [22], following the main steps as listed in the R script produced by the GEO2R tool [23]. The input contrast levels were differentially expressed genes between melanoma versus skin and melanoma versus nevus. 1701 genes from a linear model fit were extracted setting FDR for multiple testing adjustment, p-value 0.001 and 2-fold changes as thresholds. As a normalization step, after taking the logarithms of the values, the mean values of normal skin were subtracted from the rest of the data, and the normal skin columns were removed from the table.

3) Data integration

The two tables containing the microarray and image data were merged to one block sparse matrix with dimensions 1104 rows x 1734 columns, marking the not available values as NA. The rows contain the microarray and image data samples, and the columns microarray and image features plus one binary response variable (0 for nevus and 1 for melanoma). All the programming of the workflow was implemented in R [24].

4) Missing values imputation

Although there are several software packages implementing advanced imputation methods [25], they mostly require that a subset of existing feature values is available per class to permit imputation of the missing ones. However this is not case in this study, as the integrative dataset is formed through the fusion of two separate datasets, which may refer to the same phenotypes but which measurements do not incorporate information about the feature space of the other set. The data are only linked by their phenotypic class description. So when the features of the two datasets are cross-tabulated, they compose an expanded feature space, where the missing values are imputed column wise, according to the imputation method

applied. In this study we considered four simple imputation methods applied per feature and per class:

- "mean value" imputation
- "normal random" imputation
- "uniform" imputation
- "bootstrap" imputation

In the second case, after deriving the mean value (m) and standard deviation (sd) of each feature (ignoring the NA values) per class, we randomly imputed the missing values by sampling from a normal distribution population, having as parameters: (m, sd). The "uniform" imputation is conducted by sampling uniformly within the range of each feature per class, and the "bootstap" imputation by sampling from the bootstrap distribution of each variable separately per class. The two last imputation methods behave similar to the way random forests construct synthetic data, in order to derive a similarity measure [15].

For the efficient execution of the imputations, the `plyr` R package was used [26].

B. Feature Selection

The setup of the in-silico experiment encompassed the examination of the reported selected feature subsets when: a) applying a co-linearity removal filter to the microarray dataset prior to the execution of the selection algorithm (marked as: Filtered/Unfiltered), and b) setting a 95% tolerance threshold to the best obtained performance criterion (Tolerance/Best). The tolerance in the performance method allows the selection of a subset size that is small enough but without sacrificing too much performance, and can produce good results where there is a plateau of good performance for larger subset sizes. The combination of these two parameters (prior filtering and tolerance threshold) resulted in the examination of four distinct cases.

For each of the four cases, a 10-fold cross-validation procedure was adopted with 50 repetitions on six different datasets: the microarray data alone (marked as om), the unified dataset produced by the mean imputations (m.i), the unified dataset following normal random imputation for the missing values (nr.i), the unified dataset by the "uniform" imputations (u.i), the unified dataset by the "bootstrap" imputations (b.i) and the image data alone (oi). In all the repetitions, the nr.i, u.i and b.i datasets were re-estimated, thus providing more sampling variations. Prior to the application of the repetitions, the datasets were centered and scaled as a pre-processing step on the predictors.

The feature selection workflow was built using the R package `caret` (classification and regression training) [27]. The search algorithm employed in caret uses the recursive feature elimination method on predefined sets of predictors, and in this study the length of the variable subsets was defined as [1 to 10, 15, 20, 25, 30, 35, 40, 45, 50], except for the image only data where the subsets were [1 to 10, 15, 20, 25, 30, 31].

For each of the 50 repetitions, the optimum subset number of predictors was recorded, along with the names of the predictors, and the performance attained. As performance metric the area under the ROC curve (auc) was set. The auc of a classifier is equivalent to the probability that the classifier

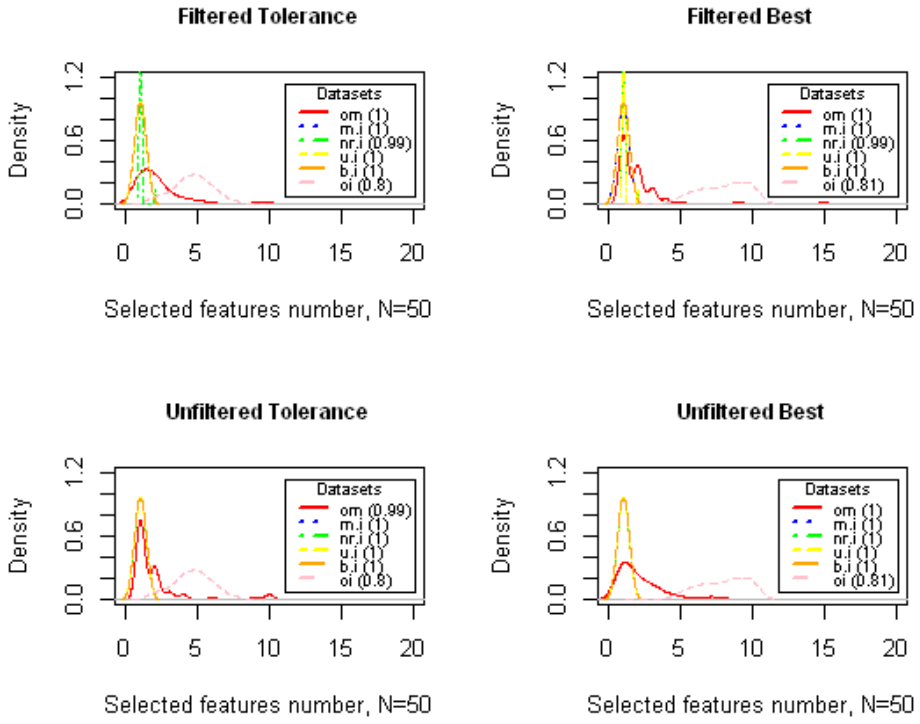

Figure 1. Density plots of the optimum features number from 50 repetitions. The six datasets are: only microarray (om), mean imputation (m.i), normal random imputation (nr.i), uniform imputation (u.i), bootstrap imputation (bi),and only image (oi). In parentheses are the medians of the obtained performances (auc) for each dataset.

will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This is equivalent to the Wilcoxon test of ranks, and also it is closely related to the Gini coefficient [28].

III. Preliminary Results and Discussion

The results of the trials are depicted in Fig. 1.

TABLE I. TOP FEATURES (GENES) SELECTED AFTER 50 REPETITIONS OF THE 10-FOLD CROSS-VALIDATION MODELING FOR THE BEST-FILTERED CASE IN EACH OF THE TREE DATASETS

Feature (om)	*Freq. (om)*	*Feature (m.i)*	*Freq. (m.i)*	*Feature (nr.i)*	*Freq. (nr.i)*
CDC37L1	47	NEIL1	4	CDC37L1	49
RRAS2	34	IFI16	3	RRAS2	2
SLC7A8	18	CTDSPL	2	-	-
HPCAL1	14	DLK2	2	-	-
IFT81	8	NADK	2	-	-
SSBP2	6	OR2A9P	2	-	-
GIPC2	5	PIK3C2G	2	-	-
CTDSPL	3	-	-	-	-

om: only-image dataset, m.i: mean imputation dataset, nr.i: normal random imputation dataset

TABLE II. TOP FEATURES SELECTED AT THE TOLERANCE-FILTERED CASE

Feature (om)	*Freq. (om)*	*Feature (m.i)*	*Freq. (m.i)*	*Feature (nr.i)*	*Freq. (nr.i))*
CDC37L1	45	PARD3	5	CDC37L1	40
RRAS2	25	ACOT9	3	RRAS2	6
SLC7A8	17	CYP4F3	3	HPCAL1.1	2
HPCAL1.1	10	FZD10	3	SSBP2	2
IFT81	6	NEIL1	3	-	-
GIPC2	5	ACADL	2	-	-
CTDSPL	4	MTUS1	2	-	-
NEIL1	4	PER3	2	-	-
SSBP2	3	PPP2R3A	2	-	-
SMAD5OS	2	SMAD5OS	2	-	-

om: only-image dataset, m.i: mean imputation dataset, nr.i: normal random imputation dataset

Regarding the median value of optimum performances, in all cases an almost perfect score was achieved in the case of the unified datasets. Only-image dataset (oi) exhibited the lower performance and the higher subset numbers. The application of the co-linearity reduction filter reduced the dispersion of the optimal subset number. Furthermore, the execution time in the reduced dataset was 4 times faster, in proportion to the number of remaining features after the use of the filter (482 from the initial 1701 differentially expressed genes in the microarray dataset). The results on the imputed datasets demonstrated a minimization of the dispersion of the subset numbers too. In addition, the four imputed datasets presented a very similar distribution. Nevertheless this similarity was abolished, when we compared the gene sets, retrieved in each case. As shown at Tables 1 and 2, the normal random imputation dataset (nr.i) resulted in a considerably more stable selection of features compared to the mean imputation unified dataset (m.i). The same pattern is observed for the unfiltered cases too. In the unfiltered cases, the nr.i dataset exhibited far better stability in the predictors' selection outperforming even the om dataset. The features resulted from the mean imputation unified dataset presented high instability, and were considered as the least preferable option to the imputation procedure. The other two imputation methods (u.i and b.i) performed equally well to the nr.i method.

The deficiency of using only performance indicators for marker discovery has been noted in the literature [29] and this is consistent with the findings of this study. The measure of stability of feature selection results with respect to sampling top selected features at the tolerance-filtered case variations, provides higher confidence regarding the suggested signatures. In this way, they become plausible targets on a biomarker discovery venture. In this case, the imputations applied by the nr.i, u.i, and b.i methods, managed to emulate the effect of a noise factor on the unified dataset, adding realistic signal variation. Thus, they immunize the discovery process in the integrative dataset, from false positive artifacts, which do not have a true differential effect. This additional variation resulted in the retrieval of smaller, optimal subsets of features, consisting of fewer, re-occurring genes as possible biomarkers.

Notably, none of the image-derived features were present to the top selected features of the unified datasets, as seen at Tables 1 and 2. In order to assess the importance ranking of image-features, 50 repetitions of the random forest algorithm were run for the unified dataset imputed by the four methods (m.i,nr.i, u.i and b.i). Each of the resulted 50 lists of features was sorted by decreasing importance. Next, the positions of the image-features in the lists were collected and the density

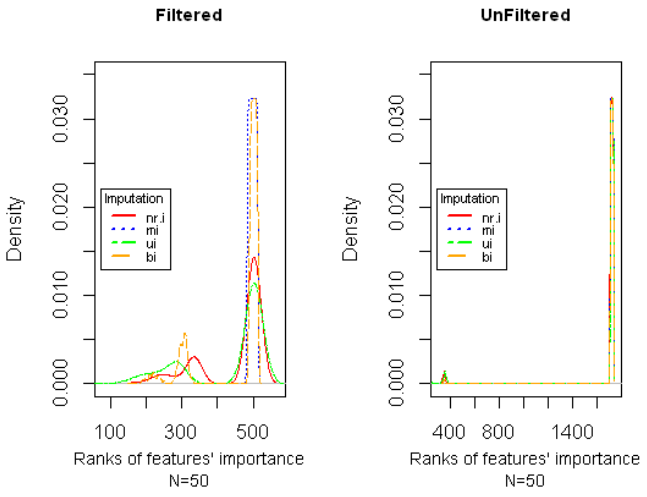

Figure 2. Density plots of importance ranks for image-derived features (ranking in the x-axis is in decreasing order of importance).

plots for the filtered/unfiltered cases are shown in Fig 2. Random forest avails four importance measures [30] and in this case the "MeanDecreaseGini" criterion was chosen. The results using the other three criteria were similar.

IV. CONCLUSIONS

In this work, two disparate datasets, concerning the study of the same physiological type of cutaneous melanoma, one of image (dermatoscopy) and the other of molecular (trascriptomic expression) origin, are fused in an integrative dataset. The majority of the image-features ranked as less important when compared to the microarray features. This implies their lower informative power with respect to the total observed variation in the integrated dataset, probably due to the fact that technical covariance but also size, leave their fingerprint in the integration process, despite the application of normalization techniques, thus inflicting their effect on the response vector of the disease. When using the nr.i, u.i, or b.i methods however, a better performance of the image features is observed, which is captured as their more frequent presence in higher positions of the classifier's vector, in discord to the results of the m.i method. Mean imputation process resulted in scoring all image features in the lowest positions of the complete feature set, considering them less informative compared to the microarray features. In this sense, it is obvious that the three imputation methods: nr.i, u.i and bi yield a more impartial effect, as can be surmised from the improved score of the image related features, providing practical value to its application in the integration process, as the simulated dataset thus derived, is a more realistic representation of the real one.

This is the first attempt, to the best of our knowledge, to assess feature selection algorithms on integrative datasets retrieved from separate sources (modalities) dissecting the same pathological mechanism.

As future work we intend to examine further the discriminative effect of the applied imputation algorithms on the unified datasets compared to the original microarray and image datasets.

ACKNOWLEDGMENT

The authors would like to thank Dr. Max Kuhn creator of the `caret` R package for his support on the functionality of this software.

REFERENCES

[1] J. W. Fenner, B. Brook, G. Clapworthy *et al.*, "The EuroPhysiome, STEP and a roadmap for the virtual physiological human," *Philos Transact A Math Phys Eng Sci,* vol. 366, no. 1878, pp. 2979-99, Sep 13, 2008.

[2] T. Rohlfing, A. Pfefferbaum, E. V. Sullivan *et al.*, "Information fusion in biomedical image analysis: combination of data vs. combination of interpretations," *Inf Process Med Imaging,* vol. 19, pp. 150-61, 2005.

[3] R. Haapanen, and S. Tuominen, "Data combination and feature selection for multi-source forest inventory," *Photogrammetric Engineering and Remote Sensing,* vol. 74, no. 7, pp. 869-880, 2008.

[4] J. L. Jesneck, L. W. Nolte, J. A. Baker *et al.*, "Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis," *Med Phys,* vol. 33, no. 8, pp. 2945-54, Aug, 2006.

[5] G. Lee, S. Doyle, J. Monaco *et al.*, "A knowledge representation framework for integration, classification of multi-scale imaging and non-imaging data: preliminary results in predicting prostate cancer recurrence by fusing mass spectrometry and histology," in Proceedings of the Sixth IEEE international conference on Symposium on Biomedical Imaging: From Nano to Macro, Boston, Massachusetts, USA, 2009, pp. 77-80.

[6] P. Tiwari, S. Viswanath, G. Lee *et al.*, "Multi-Modal Data Fusion Schemes For Integrated Classification Of Imaging And Non-Imaging Biomedical Data," in IEEE International Symposium on Biomedical Imaging (ISBI), 2011, pp. 165-168.

[7] M. Balázs, S. Ecsedi, L. Vízkeleti *et al.*, "Genomics of Human Malignant Melanoma " *Breakthroughs in Melanoma Research*, Breakthroughs in Melanoma Research Y. Tanaka, ed., InTech, 2011.

[8] J. Timar, B. Gyorffy, and E. Raso, "Gene signature of the metastatic potential of cutaneous melanoma: too much for too little?," *Clin Exp Metastasis,* vol. 27, no. 6, pp. 371-87, Aug, 2010.

[9] W. K. Martins, G. H. Esteves, O. M. Almeida *et al.*, "Gene network analyses point to the importance of human tissue kallikreins in melanoma progression," *BMC Med Genomics,* vol. 4, pp. 76, 2011.

[10] M. Ogorzałek, L. Nowak, G. Surówka *et al.*, "Modern Techniques for Computer-Aided Melanoma Diagnosis," *Melanoma in the Clinic - Diagnosis, Management and Complications of Malignancy*, Melanoma in the Clinic - Diagnosis, Management and Complications of Malignancy M. Murph, ed., InTech, 2011.

[11] I. Maglogiannis, and C. N. Doukas, "Overview of advanced computer vision systems for skin lesions characterization," *IEEE Trans Inf Technol Biomed,* vol. 13, no. 5, pp. 721-33, Sep, 2009.

[12] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics,* vol. 23, no. 19, pp. 2507-17, Oct 1, 2007.

[13] L. Breiman, "Random Forests," *Machine Learning*, 2001, pp. 5-32.

[14] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 23, no. 4, pp. 687-719, 2009.

[15] A. Liaw, and M. Wiener, "Classification and Regression by randomForest," *R News,* vol. 2, no. 3, pp. 18-22, 2002.

[16] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," http://stat-reports.lib.berkeley.edu/accessPages/666.html, 2004].

[17] M. Maragoudakis, and I. Maglogiannis, "Skin lesion diagnosis from images using novel ensemble classification techniques," in 10th IEEE EMBS International Conference on Information Technology Applications in Biomedicine, Corfu, Greece, 2010.

[18] T. Barrett, D. B. Troup, S. E. Wilhite *et al.*, "NCBI GEO: archive for functional genomics data sets--10 years on," *Nucleic Acids Res,* vol. 39, no. Database issue, pp. D1005-10, Jan, 2011.

[19] D. Talantov, A. Mazumder, J. X. Yu *et al.*, "Novel genes associated with malignant melanoma but not benign

melanocytic lesions," *Clin Cancer Res,* vol. 11, no. 20, pp. 7234-42, Oct 15, 2005.

[20] S. Davis, and P. Meltzer, "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor," *Bioinformatics,* vol. 14, pp. 1846-1847, 2007.

[21] G. K. Smyth, "Limma: linear models for microarray data," *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pp. 397-420, New York: Springer, 2005.

[22] R. C. Gentleman, V. J. Carey, D. M. Bates *et al.*, "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology,* vol. 5, pp. R80, 2004.

[23] NCBI_GEO. "GEO2R," http://www.ncbi.nlm.nih.gov/geo/info/geo2r.html.

[24] R. Development_Core_Team. "R: A Language and Environment for Statistical Computing," http://www.R-project.org.

[25] N. J. Horton, and K. P. Kleinman, "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models," *Am Stat,* vol. 61, no. 1, pp. 79-90, Feb, 2007.

[26] H. Wickham, "The Split-Apply-Combine Strategy for Data Analysis," *Journal of Statistical Software,* vol. 40, no. 1, pp. 1-29, 2011.

[27] M. Kuhn, Contributions:, S. Weston *et al.*, "caret: Classification and Regression Training," http://CRAN.R-project.org/package=caret, 2012].

[28] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.,* vol. 27, no. 8, pp. 861-874, 2006.

[29] Z. He, and W. Yu, "Stable feature selection for biomarker discovery," *Comput Biol Chem,* vol. 34, no. 4, pp. 215-25, Aug, 2010.

[30] L. Breiman, "Manual on setting up, using, and understanding random forests v3.1," http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf, 2002].