Decision Tree Models for Developing Molecular Classifiers for Cancer Diagnosis

Alexandru Floares
Artificial Intelligence Department
SAIA & OncoPredict & Cancer Institute
Cluj-Napoca, Romania
Email: alexandru.floares@saia.ro

Adriana Birlutiu
Artificial Intelligence Department
SAIA & OncoPredict
Cluj-Napoca, Romania
Email: adriana.birlutiu@saia.ro

***Abstract*—The aim of this study is to propose a methodology for developing intelligent systems for cancer diagnosis and evaluate it on bladder cancer. Owing to recent advances in high-throughput experiments, large data repositories are now freely available for use. However, the process of extracting information from these data and transforming it into clinically useful knowledge needs to be improved. Consequently, the research focus is shifting from merely data production towards developing methods to manage and analyze it. In this study, we build classification models that are able to discriminate between normal and cancer samples based on the molecular biomarkers discovered. We focus on transparent and interpretable models for data analysis. We built molecular classifiers using decision tree models in combination with boosting and cross-validation to distinguish between normal and malign samples. The approach is designed to avoid overfitting and overoptimistic results. We perform experimental evaluation on a data set related to the urothelial carcinoma of the bladder. We identify a set of tumor microRNAs biomarkers, which integrated in an ensemble of decision tree classifiers, can discriminate between normal and cancer samples with the best published accuracy.**

I. Introduction

Significant effort has been concentrated in the last years for acquiring high-throughput molecular biology and healthcare data with the goal of understanding the mechanisms behind various biological processes and diseases. As a result, large data repositories like GEO [1] and ArrayExpress [2] are now freely available on the Internet. However, the process of extracting information from these data and transforming it in useful knowledge is still lingering behind and needs further improvement. This is why the research focus is shifting from simply acquiring data towards developing methods to manage and analyze it and extracting meaningful knowledge out of it.

Despite significant efforts, cancer is still a lethal disease with a high mortality rate. Bladder cancer is the fifth most common malignancy accounting for about 5-7% of all new diagnosed malignancies in men, and about 2-2.5% in women [3]. The biology of bladder cancer is incompletely understood, making the management of this disease difficult. Research in cancer and gene expression has increased over the last decade with high-throughput experiments investigating thousands of molecules in parallel. Important goals are early and accurate diagnosis and prognosis and monitoring the therapeutic response in a personalized way. These can be achieved by analyzing alterations in gene sequences, mRNA and microRNA expression levels, protein structure or function. These alterations form cancer biomarkers. In particular, recent evidence suggests a regulatory role for microRNA in cancer [4]–[6]. MicroRNAs are short noncoding RNA molecules that post-transcriptionally modulate protein expression. Alterations in mRNA expression appear to be important for carcinogenesis [7], [8]. There are several studies, which investigate the role of microRNAs in bladder cancer [9]–[11]; it was observed that altered microRNA expression contributes to bladder cancer carcinogenesis [12]. Furthermore, microRNA profile may be used to identify key tumorigenic pathways [9] or clinical outcome [13], [14].

Several research efforts have been directed towards discovering molecular markers or gene expression signatures for classifying and predicting disease outcome in various cancers [15]–[17]. The results of most of these studies are lists of selected and ranked molecules capable of discriminating between one or more clinical outcomes. Although these lists are useful and interesting, it has become increasingly stringent to go one step forward and use these lists to develop accurate clinical decision support systems for cancer diagnosis, prognosis and treatment. It is important to emphasize the difference between such systems, i.e., between classifiers and lists of genes. Behind any classifier, there is a mathematical or logical relationship between the inputs (genes) and the outputs (diagnosis or prognosis) to be predicted. With the same list of genes, different mathematical relationships could give totally different performances. Unfortunately, developing accurate predictive models is more difficult than identifying differentially expressed genes. Nevertheless, the list of differentially expressed genes is the starting point of biomarkers discovery from data, as a necessary step in developing clinical decision support systems.

Computational intelligence and machine learning techniques are indispensable tools for building efficient and accurate intelligent systems for clinical diagnosis of various diseases, including cancer. Investigating which algorithms and computational methods are best suited has recently become an important research topic and is the subject of the work presented here. For example, a classifier for predicting bladder cancer progression developed and presented in [18] achieved a

predictive accuracy of about 75% on the data used to develop it but only 66% on the external validation. However, this accuracy proved to be robust and it is an important progress from the list of genes in progression prediction.

Previously [19], [20], we introduced the i-Biomarker concept, representing an intelligent system (indicated by the prefix "i"), which could function in a similar manner to a biomarker. A biomarker is defined as: "a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic response to a therapeutic intervention." (NIH official definition, see [21]). The output of an i-Biomarker indicates the same things as a biomarker, but its sensibility and sensitivity is usually much higher than those of the biomarkers on which it is based. Both biomarkers and their relationships with the diagnosis or prognosis (i-Biomarker output) are discovered from data, using computational intelligence methods. However, well established biomarkers could and should be also integrated.

Because the medical community prefers transparent and interpretable predictive models, favorite computational intelligence methods are decision trees, Bayesian networks, neuro-fuzzy, or genetic programming. Transparent models can be used for generation of knowledge and hypotheses from data, and they can help in scenario analysis in the context of computer-based decision support. Moreover, one can develop panels of i-Biomarkers, as ensemble of classifiers. These could be even more accurate and robust, being based on a larger list of molecular markers, and more complex relationships between them and the output, nevertheless, the output is still unique and easy to interpret.

In this study, we use microRNA data related to the urothelial carcinoma of the bladder for developing a panel of i-Biomarkers that are able to discriminate between normal and cancer samples. We focus on classification models that are transparent and interpretable, such as boosted C5.0 decision trees. The advantage of using such models is that the reasoning process behind the model is clearly evident when browsing the tree, and this allows the extraction of knowledge that can help in understanding the mechanisms behind the disease. To our knowledge, boosted C5.0 decision trees are for the first time used to develop intelligent systems capable to discriminate between normal and cancer samples based on tumor microRNA, and the 100% accuracy of the panel of i-Biomarkers is the best published.

The rest of this paper is organized as follows: next, we discuss the methods used; further, we present the experimental results, we end with discussions and conclusions.

II. Methods

In this section, we will first discuss data preprocessing, microRNA biomarkers discovery from data, and their integration in intelligent systems, i-Biomarkers implemented as decision trees. Next, the methods of developing panels of i-Biomarkers, as ensemble of decision trees, will be described. The general strategy of building i-Biomarkers will be applied here to C5.0 decision trees and boosting, as ensemble method.

A. Data Preprocessing

Microarray data in general, and data used in this study in particular, is characterized by a larger number of features compared to the number of samples or patients. This characteristic of genomic data has to be taken into account when building classification models that generalize well to new observations and which avoid overfitting. For mRNA data the difference between features and samples is very big, of the order of thousands: data contains thousands of genes and only a few tens of samples. For microRNAs the difference between features and observations is smaller than in the case of mRNA, i.e., of the order of hundreds. Besides the curse of dimensionality, microarray data are noisy and characterized by missing values.

An exploratory data analysis was performed in order to identify missing values, outliers and extreme values, followed by a two-step feature selection.

- The first step was performed in such a way as to avoid using information about the classes –Cancer or Normal– of the samples. This is crucial for avoiding overfitting in the later stages of classifiers development. We selected only the microRNAs with a certain percentage of missing values. The missing values percentage was chosen considering that decision trees cope well with moderate missing values percentages. Thus, we did not eliminate potentially relevant genes, with a reasonable percentage of missing values. The choice for decision trees motivated why the outliers and extreme values were not eliminated, as decision trees are robust to both. Then, we used the so called unspecific filters. Only genes satisfying certain criteria, such as a standard deviation and a coefficient of variation (the ratio of the standard deviation to the mean) higher than a chosen threshold were selected.
- A second feature selection is performed by the built-in feature selection capability of the decision tree algorithms that are used (this property will be discussed in the next section).

B. i-Biomarkers as C5.0 Decision Tree

For i-Biomarkers implementation, we use transparent and interpretable models, and in particular, we focus on decision tree models. The advantage of using these so-called white-box models is that they can be used for generation of knowledge and hypotheses from data. Moreover, they are favored by the medical community instead of the so-called black-box models which do not make clear the reasoning process behind the technique.

Decision tree models [13] are classification systems that can predict or classify future observations based on a set of decision rules. Decision tree algorithms perform as follows: they examine all the fields of data to find the one that gives the best classification or prediction by splitting the data into subgroups. The process is applied recursively, splitting subgroups into smaller and smaller units until the tree is finished (as defined by certain stopping criteria). The splitting procedure is performed to maximize the purity of

the classification at a given level, and it is different for each decision tree algorithm. The resulting decision tree can be used as a visual and analytical decision support tool. A general method for building decision trees is described below:

1) The process starts with an empty tree and the entire training set.
2) The following steps are iteratively repeated until a stopping criterion is reached:
   a) At the current node, if the stopping criterion is met then the node becomes terminal (leaf node). Examples of stopping criteria: all training samples belong to the same category –Cancer or Normal; or the minimum number of samples has been reached.
   b) Otherwise, the attribute (microRNA) which best splits the training samples into the two categories is determined and becomes a decision node. From the decision node, a branch is created and the samples are partitioned accordingly.

One of the main advantages of using decision trees is that the reasoning process behind the model is visible when browsing the tree. This is in contrast to black box modeling techniques, in which the internal logic can be difficult to work out. Another advantage of using decision tree models is that the process will automatically include in its rule only the features that really matter in making a decision and features which are not important or relevant will be ignored. This can yield very useful information about the data and can be used to reduce the data to relevant features before training another learning technique, such as a neural network. Summarizing, the main advantages in developing i-Biomarkers for cancer diagnosis using decision trees are:

1) *Feature selection capability*: only the relevant microRNAs are selected for building the classification model. At every decision node all the inputs are evaluated, but only the feature offering the best split is chosen for further use.
2) *Robustness*: decision trees are robust to outliers and extreme values and their accuracy is not decreased significantly by missing information.
3) *Flexibility*: the algorithm implies a limited number of tunable parameters, which makes it very flexible (these tunable parameters will be discussed in the next section for the C5.0 decision tree algorithm).
4) *Computational complexity*: building and running of decision trees are fast in comparison to other learning methods.
5) *Interpretability*: the resulting model given as a binary tree graphic is easy to interpret and in particular useful for the medical community.

Certain disadvantages appear however when using decision trees. For example, the samples used to train the classifier are reduced after every split, thus, when reaching high dimensions this can lead to overfitting. Also, the greedy search strategy (local optima) implies that small changes in the data (e.g., sampling fluctuations) can result in big changes of the fully developed tree.

We tested different decision tree algorithms such as: CART [22], CHAID [23] and C5.0 which is the latest version of the C4.5 decision tree [24]. Although all the decision trees performed well, the C5.0 algorithm gave the best results and was further used in this study. The splitting criterion used by C5.0 is based on the maximum information gain (difference in entropy). Furthermore, C5.0 presents several tunable parameters which can be adjusted to obtain the best performance, such as:

- *Pruning severity.* Trees are pruned in two stages: First, a local pruning stage, which examines subtrees and collapses branches to increase the accuracy of the model. Second, a global pruning stage considers the tree as a whole, and weak subtrees may be collapsed.
- *Minimum records per child branch.* The size of subgroups can be used to limit the number of splits in any branch of the tree, thus controlling the growth of the tree. A branch of the tree will be split only if two or more of the resulting sub-branches would contain at least a certain number of records from the training set. The default value is 2, an increase of this value helps to prevent overtraining with noisy data.
- *Winnow features*. C5.0 examines the usefulness of the predictors before starting to build the model. Predictors that are found to be irrelevant are then excluded from the model-building process. This can be helpful for models with many variables and can help to prevent overfitting.

C. Panel of i-Biomarkers as Boosted C5.0 Trees.

The performance of decision trees, and in general of the supervised learning algorithms, can be improved by using ensemble methods. Ensemble is a technique for combining many weak learners with the goal of producing a strong learner which will achieve a better predictive performance than any of the constituent learners. Examples of ensemble methods are: Bayes optimal classifier, bagging, boosting, Bayesian model averaging, etc.

Boosting [25] can be successfully applied to improve the accuracy of learning algorithms, and in particular of C5.0 decision tree. Boosting is an ensemble method which works by repeatedly developing classifiers on various distributions over the training data, and then combining these classifiers into a single composite model. Variety is created by assigning different weights to samples according to whether they were easier or harder to classify correctly (see [26] for details). The following steps describe a general algorithm for boosting applied to the decision tree classifier:

1) Assign equal weights to the samples.
2) Repeat the following steps for a given number n of trials ($j = 1, 2, \ldots, n$):
   a) Develop an i-Biomarker j (C5.0 decision tree) using the computed weights.
   b) Give higher weights to the samples more difficult to classify correctly.

c) Give lower weights to the samples easier to classify accurately.

3) Combine the n i-Biomarkers obtained by assigning higher weights to the earliest i-Biomarkers, and lower weights to the last i-Biomarkers.

Fig. 1 (adapted from [27]) shows the basic framework for ensemble methods classification. The approach is to compose an ensemble with n i-Biomarkers (decision trees), each i-Biomarker trained on a data set derived from the original data set. Each i-Biomarker is used for making predictions on the samples from the test data set. The votes of individual i-Biomarkers are integrated in a final decision output (final diagnosis).

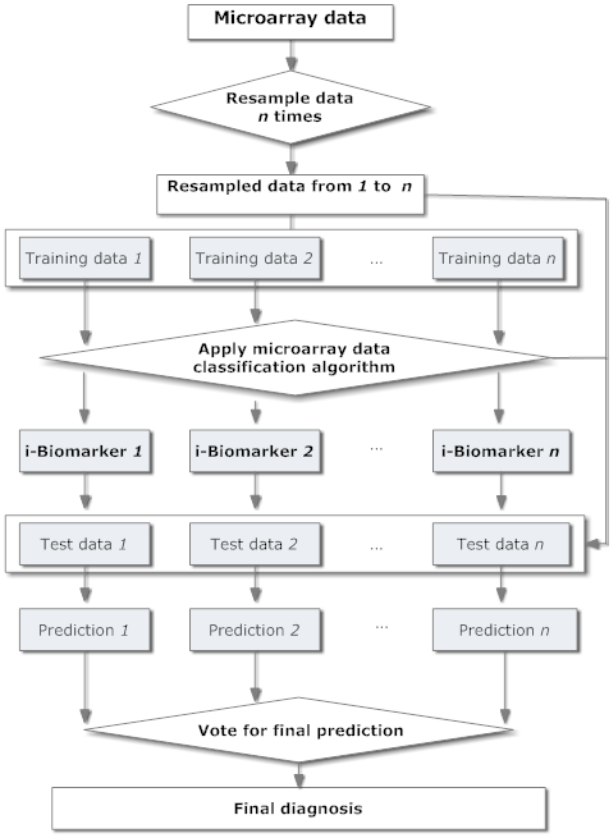

Fig. 1. Ensemble method classification flow chart. The approach is to compose an ensemble with n i-Biomarkers (decision trees), each i-Biomarker trained on a data set derived from the original data set. Each i-Biomarker is used for making predictions on the samples from the test data set. The votes of individual i-Biomarkers are integrated in a final decision (diagnosis).

Misclassification costs. For cancer diagnosis it is particularly important to take into account the different misclassification costs. Misclassifying observations of one class has more severe consequences than misclassifying observations of another class. Failure to identify a patient with cancer (false negative) has far more severe consequences than misidentifying a normal individual as having cancer (false positive). High costs should be assigned to misidentifying malignant as benign and lower costs to misidentifying benign as malignant.

III. Results and Discussions

A. Data

We used the data set made available on the Internet by Catto et al., and published in [12]. It consists of 78 samples out of which 72 were taken from patients and 6 belong to cell lines; we further considered only the samples obtained from the patients. The patient samples were classified into two groups: normal and cancer. The normal group contains a total of 20 samples and were obtained from 10 subjects with normal urothelium (taken distant to any tumor) and from 10 disease-free controls. The cancer group contains the rest of 52 samples and were sub-classified into three subgroups: *(i)* low-grade non-muscle invasive– 22 samples, *(ii)* high-grade non-muscle invasive– 12 samples and *(iii)* muscle invasive– 18 samples. Since in this study we focus only on a diagnostic problem– Cancer vs. Normal– we transformed the initial multi-class classification problem into a binary classification problem; we do not make any distinction between the classifications within the groups, considering only two classes.

For all samples, both normal and malignant, 322 microRNAs have been examined, and expression values for these have been extracted. Enriched small and total RNA were extracted using the mirVana kit (Ambion) according to the manufacturer's protocol. The expression of microRNAs and 3 small nucleolar RNA molecules was determined using preprinted microfluidic cards (Human miR v1.0, Applied Biosystems). Relative microRNA concentrations were calculated with respect to the median of three reference RNA molecules. Median data centering was performed.

B. Data Preprocessing

From the total set of genes a subset was selected according to an unspecific filter containing the following criteria *(i)* percentage of missing values smaller than 70%, *(ii)* standard deviation larger than 2 and *(iii)* coefficient of variation (ratio of standard deviation to mean) larger than 0.1. There is no information leakage from output when performing this filtering which is crucial for avoiding overfitting. The number of remaining features is 278.

The selected subset of data did not have extreme values and contained a few outliers. These outliers were not eliminated because decision trees are robust in this situation (we considered values as extreme when they exceed 5 standard deviations from the mean and outliers the values which exceed 3 standard deviations from the mean).

C. Protocol

We used a V-fold cross-validation procedure which was chosen since the number of samples in the data set is relatively small ($n = 72$) and partitioning the data into training and testing resulted in overfitting. With various settings tested, a typical result was 100% accuracy on training data and only about 70% accuracy on testing data. The cross-validation procedure is illustrated in Fig. 2. Each fold contains the same (or nearly the same) k number of samples obtained by dividing the total number of samples n to the number of folds V.

At each fold, the C5.0 algorithm performed a selection of genes and an i-Biomarker j was developed using the temporary training data set consisting of $n - k$ patients. The temporary testing data set containing the remaining k patients was used to evaluated the accuracy of the i-Biomarker. The final result is the fully developed i-Biomarker and the estimated error computed as the mean of each i-Biomarker j's error on the test set.

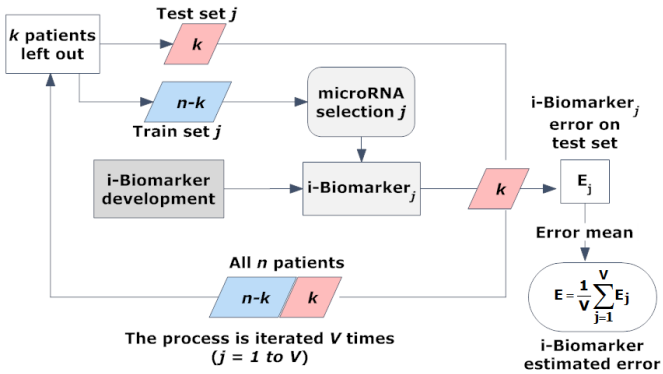

Fig. 2. Development of an i-Biomarker with a V-fold cross-validation procedure. The i-Biomarker is for a binary outcome: Cancer vs. Normal. From the initial data set with n samples, k samples are left out for creating a temporary testing set and a training set with $n - k$ samples. A different microRNA selection j is performed on the training set and a different i-Biomarker j is developed in each iteration of cross-validation. The i-Biomarker j developed is used to classify the left out samples in the testing set. Final result: the fully developed i-Biomarker and the estimated error computed as the mean of each i-Biomarker j's error on the test set.

We applied the cross-validation procedure using different values for V: $V = 2$, $V = 3$, $V = 10$, $V = 30$, and $V = 72$. The last setting of $V = 72$ is the special case of cross-validation in which the number of folds is equal to the number of samples, also known as leave-one-out cross-validation. This procedure uses in each iteration a single sample from the original data as the validation set and the remaining samples form the training set. This is repeated such that each observation in the sample is used once as the validation data.

D. Experimental Results

The experimental setup was chosen with the goal of obtaining the i-Biomarkers with the following characteristics:

- *Accuracy: high*. Accuracy is computed as the percentage of correctly classified samples from the total number of samples in the training set.
- *Robustness: high*. By robustness we understand obtaining a similar accuracy on external validation (we refer to external validation as a totally different set of patient samples). Usually, smaller trees generalize better, therefore robustness is mainly influenced by the minimum number of records per child branch.
- *Model complexity: low*. Model complexity is expressed in the number of trees and the number of nodes. The complexity of a single i-Biomarker is simply the number of nodes of its decision tree representation. The complexity of a panel of i-Biomarkers is related to the number of trees in the ensemble and the number of nodes of the trees.

We experimented with several settings of the number of folds in the cross-validation. Our experiments show that the evaluation measures did not vary with number of folds. We further considered cross-validation with the number of folds set to 10 and leave-one-out cross validation. We will show results obtained with 10 folds cross-validation.

We empirically established that for the current data set the minimum number of records per child branch should be equal to 6. This number was chosen to address the size of the data set with only 72 samples. Other settings for the C5.0 decision tree: we used a pruning severity of 75 and we did not use Winnow feature since we noticed that its presence did not affect the performance.

We first tested with a single decision tree (i-Biomarker) for which we obtained an accuracy of 95.83%; the number of false negative samples is 2 and the number of false positive is 1. In the clinical context, one aims at obtaining a false negative rate as low as possible, therefore we set different costs for misclassification for the two classes. We experimented with different settings for the misclassification costs; the most appropriate costs which we use for further experimentation are 1.1 for the false negative class and 1 for the false positive. The accuracy obtained in this case increased to 97.22%, with no false negative and 2 false positives, which is clinically acceptable.

We tested the robustness measured by the accuracy when increasing the minimum number of records per branch. We noticed that when increasing the maximum number of records per branch from 2 to 3 the accuracy decreased from 95.83% to 93.06%. We expect that on a totally different set of patients the external validation will result in a decrease of performance even more.

Ensemble methods are well-known to increase the robustness. The accuracy increases with the number of decision trees in the ensemble. With a lower number of decision trees, for example 2, the accuracy obtained was of 95.83%, while with a higher number of trees, for example 10, the accuracy increased to 100%. For medical community, it is important to mention that the panel of i-Biomarkers (the ensemble of decision trees) can be overlapping with respect to the list of genes contained by each i-Biomarker. The number of trees in the ensemble determines the number of microRNAs selected for building the i-Biomarker panel. This is illustrated in Table I.

TABLE I
NUMBER OF MICRORNAS SELECTED AS A FUNCTION OF THE NUMBER OF DECISION TREES IN THE ENSEMBLE.

Number of i-Biomarkers	2	3	5	10	20
Number of microRNAs	3	5	8	17	25

TABLE II
SUMMARY FOR THE I-BIOMARKERS (DECISION TREES) FORMING THE PANEL OF I-BIOMARKERS (THE ENSEMBLE OF CLASSIFIERS). FOR EACH I-BIOMARKER (ROW) IS SHOWN THE ACCURACY, THE SIZE (NUMBER OF NODES), AND THE LIST OF GENES.

i-Biomarker	Accuracy	Size	List of genes
i-Biomarker 1	94.80%	3	*hsa-miR-133b, hsa-miR-303, hsa-miR-512-2p*
i-Biomarker 2	87.77%	1	*hsa-miR-218*
i-Biomarker 3	91.68%	2	*hsa-miR-133b, hsa-miR-146b*
i-Biomarker 4	94.50%	3	*hsa-miR-133b, hsa-miR-34c, hsa-miR-550*
i-Biomarker 5	88.53%	2	*hsa-miR-218, hsa-miR-622*
i-Biomarker 6	83.79%	2	*hsa-miR-10a, hsa-miR-143*
i-Biomarker 7	99.53%	4	*hsa-miR-133b, hsa-miR-650, hsa-miR-21, hsa-miR-575*
i-Biomarker 8	94.69%	3	*hsa-miR-30a-3p, hsa-miR-155,hsa-miR-218*
i-Biomarker 9	81.06%	1	*hsa-miR-93*
i-Biomarker 10	92.81%	3	*hsa-miR30a-3p, hsa-miR-135b, hsa-miR-383*

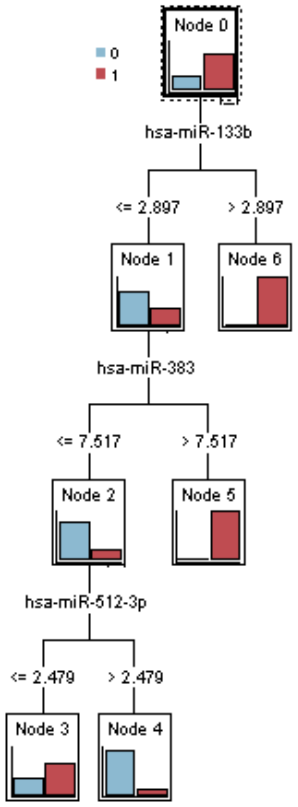

Fig. 3. i-Biomarker represented by a decision tree. The samples are classified in the terminal nodes of the tree: cancer (red rectangles) or normal (blue rectangles). For a new sample, we observe the values of the three genes and compare them with the threshold values identified at each node. For example, if the value of *hsa-miR-133b* for the new sample is higher than the threshold value 2.897, then the sample is classified as cancer (class 1, the red color); otherwise, we compare the value of *hsa-miR-383* with the threshold value equal to 7.517: if it is higher, the sample is included in the normal cases; else, *hsa-miR-512-3p* and the associated threshold of 2.479 is investigated and depending on this the sample is classified in one of the two classes.

Based on our evaluation criteria, the best i-Biomarker panel obtained (not shown due to European patent law) is characterized by the highest accuracy (100%), best robustness (minimum 6 records per child branch), lowest complexity (10 decision trees in the ensemble). For illustrative purposes, we show a summary for the i-Biomarkers (decision trees) forming a panel of i-Biomarkers (the ensemble of classifiers) in Table II. The characteristics of this panel are minimum 2 records per child branch (the robustness in this case is not the optimal one), and 10 decision trees in the ensemble and the accuracy obtained is of 100%. For each i-Biomarker is shown the accuracy, the size (number of nodes), and the list of genes. The individual decision trees in the ensemble have an average depth of 4; this is equivalent with the ensemble providing a short list of genes based on which new samples can be classified. The accuracy of individual trees in the ensemble is between 83.79% and 99.53%. The individual decision trees can be used separately for diagnosis, but the accuracy and robustness obtained in this case are lower than in the case of ensemble. One of the resulting i-Biomarkers is represented by the decision tree shown in Fig. 3. The samples are classified in the terminal nodes of the tree: Cancer (red rectangles) or Normal (blue rectangles). The decision tree is easy to interpret. For a new sample, we observe the values of the three genes and compare them with the threshold values identified at each node. For example, if the value of *hsa-miR-133b* for the new sample is higher than the threshold value 2.897, then the sample is classified as cancer (class 1, the red color); otherwise, we compare the value of *hsa-miR-383* with the threshold value equal to 7.517: if it is higher, the sample is included in the normal cases; else, *hsa-miR-512-3p* and the associated threshold of 2.479 is investigated and depending on this the sample is classified in one of the two classes.

IV. CONCLUSIONS AND FUTURE WORK

We proposed a methodology for developing transparent, interpretable, and highly accurate intelligent clinical decision support systems for cancer diagnosis, using an ensemble of decision trees in a knowledge discovery from data approach

especially designed to avoid overfitting and overoptimistic results. Special attention was paid to increase the systems robustness, to avoid a significant accuracy decrease on different patients data sets (external validation), and to false negative rate (cancer samples misclassified as normal) minimization, which is desired in a clinical context. The study presented in this paper is the first step in developing an integrated system which can be realistic in a clinical setting for bladder cancer diagnosis based on plasma microRNA and bladder cancer prognosis based on tumor microRNA.

In future work, the molecules selected will be placed on pathways and networks to understand the complex molecular interactions involved. The high accuracy of the classifiers also means that a relevant set of genes was discovered and as a consequence the relevant pathways, deregulated in bladder cancer, will be identified. The long-term goal will be to define for each pathway, targets that are amenable to drug discovery efforts, and that can be validated in cell culture models.

ACKNOWLEDGMENT

The authors would like to thank to professor Liana Adam from Urology Department, MD Anderson Cancer Center, Texas, USA, for useful discussion about the biology and diagnosis potential of microRNA in bladder cancer, and to Ioana Ilea and Simina Crisan for their help in preparing the manuscript.

REFERENCES

[1] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva, "Ncbi geo: archive for functional genomics data sets 10 years on," *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D1005–D1010, 2011.

[2] H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, N. Kurbatova, M. Lukk, J. Malone, R. Mani, E. Pilicheva, G. Rustici, A. Sharma, E. Williams, T. Adamusiak, M. Brandizi, N. Sklyar, and A. Brazma, "Arrayexpress updatean archive of microarray and high-throughput sequencing-based functional genomics experiments," *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D1002–D1004, 2011.

[3] N. Howlader, A. Noone, M. Krapcho, R. A. R. N. Neyman, and et al., "SEER Cancer Statistics Review 1975-2008 National Cancer Institute," *Bethesda*, p. posted to the SEER web site, 2011.

[4] G. A. Calin and C. M. Croce, "MicroRNA signatures in human cancers," *Nature Reviews Cancer*, vol. 6, no. 11, pp. 857–866, 2006.

[5] Y. Han, J. Chen, X. Zhao, C. Liang, Y. Wang, L. Sun, Z. Jiang, Z. Zhang, R. Yang, J. Chen, and et al., "MicroRNA expression signatures of bladder cancer revealed by deep sequencing," *PLoS ONE*, vol. 6, no. 3, p. 6, 2011.

[6] S. Bandyopadhyay, R. Mitra, U. Maulik, and M. Q. Zhang, "Development of the human cancer microrna network," *Silence*, vol. 1, no. 1, p. 6, 2010.

[7] L. He, M. Thomson, M. Hemann, and et al, "A microrna polycistron as a potential human oncogene," *Nature*, vol. 435, pp. 828–33, 2005.

[8] P. Voorhoeve, C. le Sage, M. Schrier, and et al., "Genetic screen implicates mirna372 and mirna373 as onco-genes in testicular germ cell tumors," *Cell*, vol. 124, no. 6, pp. 1169–81, 2006.

[9] S. Volinia, G. A. Calin, C.-G. G. Liu, S. Ambs, A. Cimmino, F. Petrocca, R. Visone, M. Iorio, C. Roldo, M. Ferracin, R. L. Prueitt, N. Yanaihara, G. Lanza, A. Scarpa, A. Vecchione, M. Negrini, C. C. Harris, and C. M. Croce, "A microRNA expression signature of human solid tumors defines cancer gene targets." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 7, pp. 2257–2261, Feb. 2006. [Online]. Available: http://dx.doi.org/10.1073/pnas.0510565103

[10] F. Gottardo, C. G. Liu, M. Ferracin, G. A. Calin, M. Fassan, P. Bassi, C. Sevignani, D. Byrne, M. Negrini, F. Pagano, and et al., "Micro-rna profiling in kidney and bladder cancers." *Urologic Oncology*, vol. 25, no. 5, pp. 387–392, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17826655

[11] L. A. Neely, K. M. Rieger-Christ, B. S. Neto, A. Eroshkin, J. Garver, S. Patel, N. A. Phung, S. McLaughlin, J. A. Libertino, D. Whitney, and et al., "A microrna expression ratio defining the invasive phenotype in bladder tumors." *Urologic Oncology*, vol. 28, no. 1, pp. 39–48, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18799331

[12] W. Catto, S. Miah, H. Owen, and et. al, "Distinct microrna alterations characterize high and low grade bladder cancer," *Cancer Research*, vol. 69, no. 21, pp. 84 772–81, 2008.

[13] A. J. Schetter, S. Y. Leung, J. J. Sohn, K. A. Zanetti, E. D. Bowman, N. Yanaihara, S. T. Yuen, T. L. Chan, D. L. Kwong, G. K. Au, C. G. Liu, G. A. Calin, C. M. Croce, and C. C. Harris, "MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma." *JAMA*, vol. 299, no. 4, pp. 425–436, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1001/jama.299.4.425

[14] S.-L. Yu, H.-Y. Chen, G.-C. Chang, C.-Y. Chen, H.-W. Chen, S. Singh, C.-L. Cheng, C.-J. Yu, Y.-C. Lee, H.-S. Chen, T.-J. Su, C.-C. Chiang, H.-N. Li, Q.-S. Hong, H.-Y. Su, C.-C. Chen, W.-J. Chen, C.-C. Liu, W.-K. Chan, W. J. Chen, K.-C. Li, J. J. W. Chen, and P.-C. Yang, "MicroRNA Signature Predicts Survival and Relapse in Lung Cancer," *Cancer Cell*, vol. 13, no. 1, pp. 48–57, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.ccr.2007.12.008

[15] L. J. Van't Veer and R. Bernards, "Enabling personalized cancer medicine through analysis of gene-expression patterns." *Nature*, vol. 452, no. 7187, pp. 564–570, 2008.

[16] J. M. Llovet, Y. Chen, E. Wurmbach, S. Roayaie, M. I. Fiel, M. Schwartz, S. N. Thung, G. Khitrov, W. Zhang, A. Villanueva, and et al., "A molecular signature to discriminate dysplastic nodules from early hepatocellular carcinoma in hcv cirrhosis." *Gastroenterology*, vol. 131, no. 6, pp. 1758–1767, 2006.

[17] Z. Walther and D. Jain, "Molecular pathology of hepatic neoplasms: Classification and clinical significance," *Pathology Research International*, 2011.

[18] L. Dyrskjt, K. Zieger, F. X. Real, N. Malats, A. Carrato, C. Hurst, S. Kotwal, M. Knowles, P.-U. Malmstrm, M. De La Torre, and et al., "Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: a multicenter validation study." *Clinical Cancer Research*, vol. 13, no. 12, pp. 3545–3551, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17575217

[19] A. G. Floares, C. Floares, O. Vermesan, T. Popa, M. Williams, S. Ajibode, L. Chang-Gong, D. Lixia, W. Jing, T. Nicola, D. Jackson, C. Dinney, and L. Adam, "Intelligent clinical decision support systems for non-invasive bladder cancer diagnosis," in *Proceedings of the 7th international conference on Computational intelligence methods for bioinformatics and biostatistics*, ser. CIBB'10. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 253–262.

[20] A. Floares, O. Balacescu, C. Floares, L. Balacescu, T. Popa, and O. Vermesan, "Mining knowledge and data to discover intelligent molecular biomarkers: Prostate cancer i-biomarkers," in *Soft Computing Applications (SOFA), 2010 4th International Workshop on*, july 2010, pp. 113 –118.

[21] B. D. W. Group, "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," *Clin Pharmacol Ther.*, vol. 69, no. 3, pp. 89–95, 2001.

[22] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

[23] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, vol. 29, no. 2, pp. 119–127, 1980.

[24] J. Quinlan, *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, 1993.

[25] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," in *In Proceedings of the Thirteenth International Conference on Machine Learning*, 1996, pp. 148–156. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.3868

[26] G. Seni and J. F. Elder, "Ensemble methods in data mining: Improving accuracy through combining predictions," *Statistics*, vol. 2, no. 1, pp. 1–126, 2010.

[27] H. Hong, "Accurate and robust algorithms for microarray data classification," Ph.D. dissertation, University of Southern Queensland, 2008.