Received May 27, 2021, accepted June 15, 2021, date of publication June 25, 2021, date of current version July 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3092368

Selecting Relevant Genes From Microarray Datasets Using a Random Forest Model

HUI XIA, YASEMIN M. AKAY, (Senior Member, IEEE), AND METIN AKAY, (Fellow, IEEE)
Department of Biomedical Engineering, University of Houston, Houston, TX 77204, USA

Corresponding author: Metin Akay (makay@uh.edu)

This work was supported by the University of Houston.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Animal Care and Use Committee (IACUC) and the University of Houston Animal Care Operations (ACO).

ABSTRACT Recent studies have demonstrated microarray expression data can be used to identify gene regulatory pathways. However, one of the major challenges is to utilize the large microarray data (genes and micro-RNAs) to have an efficient computational model. Therefore, there is an urgent need to reduce the dimensionality of these large sets using machine learning methods without compromising the accuracy. This requires an appropriate machine learning algorithm to select the significant features from these large datasets. Therefore, in this study, we use a supervised method based on a Random Forest to identify significant features from three microarray datasets from prenatal nicotine, alcohol, and nicotine and alcohol exposure groups in two different cell types (dopamine and non-dopamine neurons). Our approach was computationally efficient to reduce the dimensionality of extremely large microarray datasets. Furthermore, our results indicated that using only the top 20% of features was sufficient to confirm the genetic pathways previously identified when using all of the features in the model.

INDEX TERMS Feature selection, microarray, random forest.

I. INTRODUCTION

Microarrays enable the global screening of gene expression profiles by quantifying the changes in the regulation of thousands of genes [1]. Recently, microarrays have been adopted to identify the gene regulation pathways [2] using supervised or unsupervised machine learning methods. In practice, the large number of features limits the model reliability and in many cases, may cause overfitting [3]. To improve the efficiency of the gene regulatory network modelling, the dimensionality of the features including messenger RNAs (mRNA, genes) and microRNAs (miRNAs) needs to be reduced [4].

There are two different approaches including unsupervised and supervised methods to reduce the dimensionality of complex datasets. In unsupervised learning, having a large size data and features negatively affects the computational performance of the underlying learning algorithm. The Hill Climb (HC) unsupervised learning algorithm for dimensionally reduction has been widely used in practice to improve its computational efficiency [5].

Thus, a feature selection step is necessary before modelling microarray data using machine learning algorithms. Several feature selection methods have been proposed to identify important genes using unsupervised models including statistical clustering [6]–[9], consensus group [10], particle swarm [6], [11], coefficient correlation [6], [12], [13], and principal component analysis (PCA) [14]–[16], with the classifiers such as support-vector machine (SVM) [10], [17], [18], Neural Networks [13], k-nearest neighbors (KNN) [9], [11], and K-means [14]. These methods have been found to be successful in reducing the dimensionality features.

Biological systems are inherently complex and nonlinear [19], [20]. Therefore, the extraction of relevant features and determining the related pathways using these large datasets can be challenging. Furthermore, this requires the use of non-linear modelling approaches since linear feature selection methods are inadequate. One nonlinear approach involves tree models [21], which are computationally fast and have been widely adopted as an effective and efficient feature selection solution in cancer diagnostics [22], [23], cancer classification [24], [25], image-based medical

The associate editor coordinating the review of this manuscript and approving it for publication was Henry Hess.

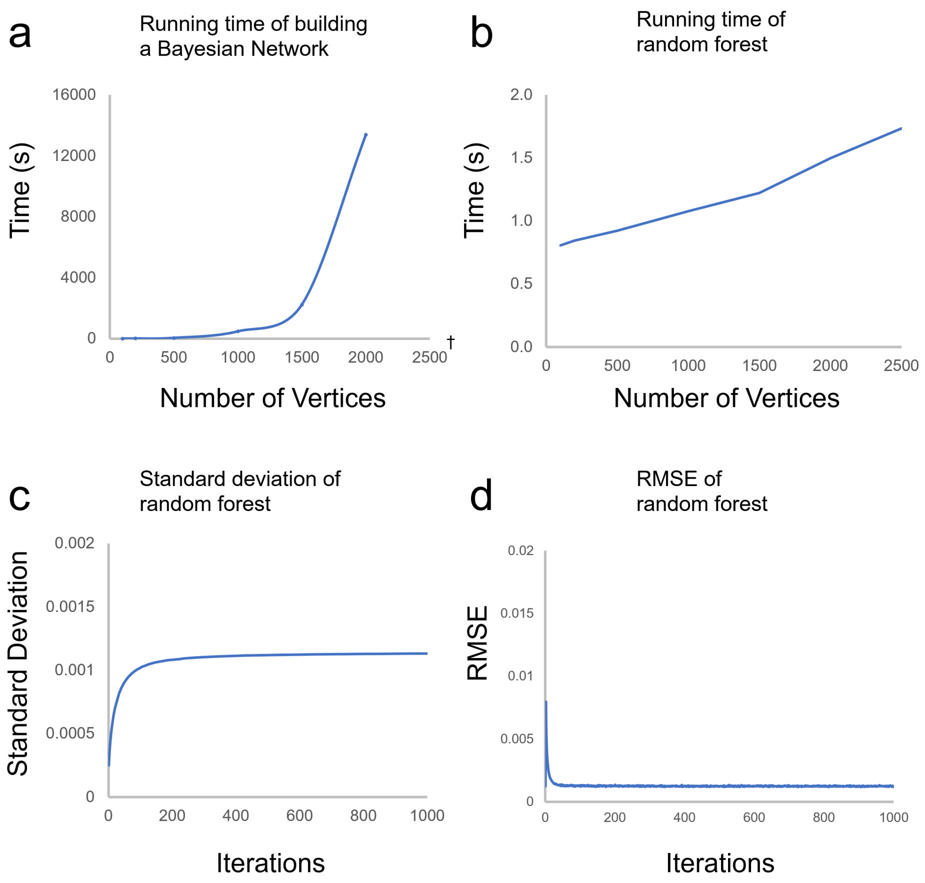

FIGURE 1. Gini index estimation in random forest is based on representative features (i.e., differentially expressed genes (DEGs) and differentially expressed miRNAs (DEmiRs)). (a) Bayesian Network Running time over various number of features using bnlearn. †Using 2500 features, a Bayesian Network cannot be built within 24 hours (b) Running time of performing one iteration of impurity-based feature importance algorithm with Random Forest. The feature importance algorithm was updated using variant random seeds until the (c) standard deviation and (d) root-mean-square-error (RMSE) converges.

diagnostics [26], drunk-drive detection [27] and spectrometry data analysis [28].

In this paper, we performed feature selection using the supervised Random Forest classifier over a collection of expression data (differentially expressed genes (DEGs) and differentially expressed miRNAs (DEmiRs)) obtained from 13 microarrays. These data were previously generated by our group to investigate perinatal exposure to alcohol, nicotine, and both nicotine and alcohol during rat gestational development [29], [30]. This final dataset for this study consisted of 5523 genes and miRNAs for the alcohol exposure, 7863 for nicotine, and 5613 for co-exposure (i.e., nicotine and alcohol) dataset. To implement the Random Forest, we assigned the data to three labels (i.e., nicotine/alcohol/dopamine-cell (DA) as shown in **Table 1**). Among these three labels, samples labelled as "nicotine" and/or "alcohol" identified whether the pup was prenatally exposed to nicotine and/or alcohol, respectively. The DA label was used to identify dopamine cells. After feature selection, we performed pathway enrichment analysis over both the feature-reduced datasets and the original datasets. We found comparable genetic regulation pathways using both methods of modelling the datasets.

II. METHODS

A. ANIMAL EXPERIMENTS

The microarray data was collected from dopaminergic and non-dopaminergic neurons obtained from the rat ventral tegmental area (VTA). All experiments were performed in accordance with the protocols approved by the Institutional Animal Care and Use Committee (IACUC) and the University of Houston Animal Care Operations (ACO). The detailed protocol has been published previously [29]–[32]. Briefly, pregnant Sprague–Dawley (SD) rats (Charles River, Wilmington, MA, USA) were maintained at standard conditions and were given an *ad libitum* diet. Rats were implanted with a subcutaneous osmotic minipump (Alzet, Cupertino, CA) containing either nicotine (at levels to stimulate moderate smoking CITE) or saline. A liquid diet of ethanol was gradually

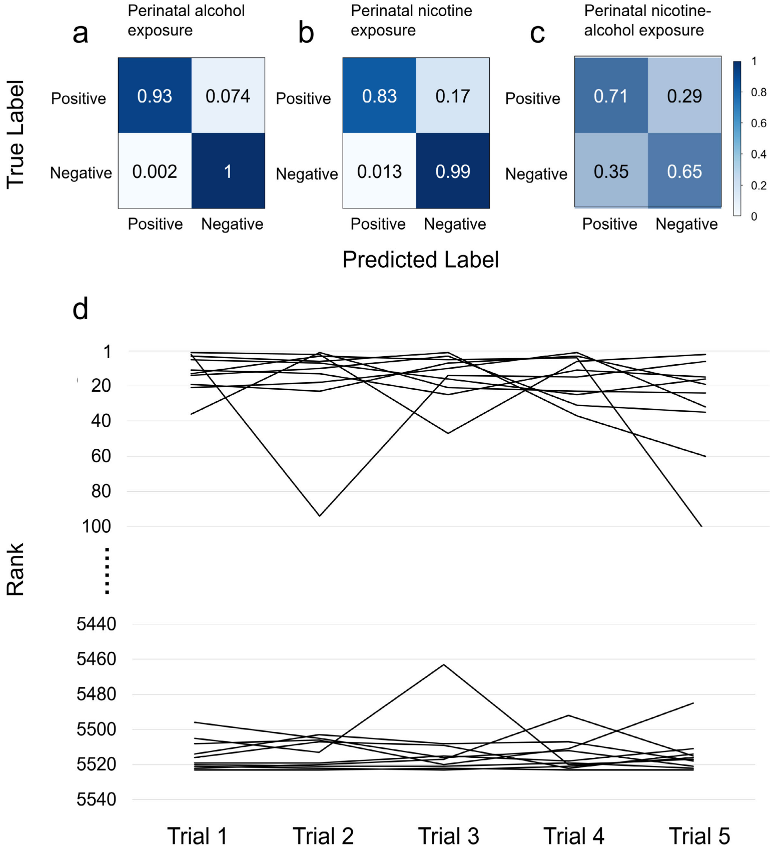

FIGURE 2. (a)-(c) Confusion matrix for Random Forest based feature importance selection on the DEGs and DEmiRs from the perinatal alcohol exposure versus the negative control. (a) sample from embryo related with perinatal alcohol exposure, (b) sample from embryo related with perinatal nicotine exposure, and (c) sample from DA cells exposed to XX. The test set consisted of 5000 data points randomly selected from the entire data set. Rows correspond to predicted classes and columns correspond to actual classes. The overall low-resolution classification of the algorithm is 0.68 ± 0.15. (d) tracks the rank of ten most important and ten least important features (DEGs and DEmiRs) from five machine learning trials over the perinatal alcohol dataset, among more than 5000 features.

introduced to the pregnant mothers to produce blood alcohol concentrations similar to what is observed in children with fetal alcohol spectrum disorders [33].

The fetus was continuously exposed via the placenta to nicotine and/or alcohol ingested by the mother from gestational day 6 (G6) to delivery (around G21–22). After birth, the pups were still exposed to nicotine and/or alcohol via the rat mother's milk.

The brain tissue samples from the pups were pooled for each litter for either the alcohol, nicotine, nicotine-alcohol, and saline treated groups for a total n = 13 litter groups?. The samples were then dissociated, pelleted, fixed, and labelled with conjugated primary antibodies neuronal marker, NeuN/Alexa Fluor 488 (NeuN/AF488, ab190195, Abcam, Cambridge, MA, USA), and tyrosine hydroxylase/ phycoerythrin (TH/PE, ab209921, Abcam). The labelled cells were sorted on an (LSR II) FACS Aria (BD Biosciences, San Jose, CA, USA) flow cytometer to identify dopamine (DA) and non-dopamine (NDA) neurons.

Following sorting, the total RNA of the cells was extracted using a miRNeasy Micro Kit (Qiagen, Hilden, Germany). The expression level of mRNA and miRNA was accessed using Agilent Sureprint mRNA and miRNA microarrays (Santa Clara, CA, USA). The raw microarray dataset was then collected from the resulting images using the Feature Extraction Software v12.0.1.

B. RANDOM FOREST

One major challenge to perform feature selection over microarray data is the inability to rely on a feature's (e.g., gene or microRNA) expression level because is difficult to determine its relevance. Therefore, when performing feature selection, we selected a subset of features, which improved the performance (in terms of running time or accuracy) of the machine learning algorithm. This method is typically considered a nondeterministic polynomial hard (NP-hard) problem [34]. This challenge becomes even more pronounced when analyzing microarray data, in part, due to the large

TABLE 1. Representative structure of the data (X) and the multi-labelling (y) of the microarray samples used in this study. For the data, each column represents the expression level (adjusted against negative control) of a differentially expressed gene or miRNA. For each sample set (i.e., AlvS, NivS, and NiAlvS), 5523-7863 genes and microRNAs are included/evaluated. Each row represents a sample, which is named by their respective treatment method, and the origin of the measured cell group. DA Cell: dopamine cells. ADA: dopamine cells exposed to alcohol; NADA: dopamine cells exposed to both nicotine and alcohol. SDA: dopamine cells treated with saline (control). NAND: non-dopamine cells exposed to both nicotine and alcohol. SND: non-dopamine cells treated with saline (control). NDA: dopamine cells exposed to nicotine. NND: non-dopamine cells exposed to nicotine.

Sample	X					y		
	Zfp521	Pou6f1	...	rno-miR-99b-5p	rno-miR-9a-5p	Alcohol	Nicotine	DA Cell
ADA1	4.438	8.378	...	5.488	9.731	1	0	1
ADA2	4.401	8.125	...	5.559	9.713	1	0	1
NADA1	3.558	6.579	...	5.463	9.390	1	1	1
NADA2	3.441	6.432	...	5.378	9.302	1	1	1
SDA1	1.041	2.657	...	5.767	11.378	0	0	1
NAND1	4.628	7.273	...	5.025	8.503	1	1	0
NAND2	3.985	7.342	...	5.079	8.472	1	1	0
SND2	3.485	9.474	...	5.567	10.542	0	0	0
SND4	3.745	9.115	...	5.563	10.479	0	0	0
NDA3	3.526	6.605	...	5.864	12.353	0	1	1
NND3	3.297	7.946	...	5.569	10.897	0	1	0
NDA4	3.309	6.485	...	5.874	12.183	0	1	1
NND2	3.294	7.979	...	5.549	10.828	0	1	0

dimensionality of the features, which can lead to overfitting or lower efficiency. To overcome this challenge, one feature selection method incorporates labels to the data, which then converts the unsupervised feature selection process to a supervised one.

Herein, we propose a method that converts an unsupervised feature selection to supervised using the Random Forest classifier to analyze microarray data. We combined 13 microarray datasets and created labels that reflected their respective experimental conditions. That is, whether the sample was marked as exposed to alcohol, whether the sample has been exposed to nicotine, and whether the sample was a dopamine cell. We then calculated the Gini index as the feature importance in which p_i represents the relative frequency of the feature in the dataset, and c represents the number of classes.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Using this supervised method, the Gini of each feature (miRNA or gene) was used to quantify the likelihood of the Random Forest classifier to branch the data into subgroups.

III. RESULTS

A. DATA MULTI-LABELING

The genes and miRNAs used in these datasets were the DEGs and DEmiRs from Keller, *et al.* [32]. These DEGs and DEmiRs were identified from microarray data which compared gene and miRNA expression profiles of VTA DA neurons between treatment groups and their respective controls. This approach has been previously described in more detail [29], [32]. Briefly, this method is based on a q-value <0.001 (adjusted p-value using Benjamini-Hochberg (BH) correction) and absolute log2 fold change >1 (≥ 1 for upregulation, ≤ -1 for downregulation) [32]. The treatment groups were compared to the saline control and labeled as: alcohol (AlvS), nicotine (NivS), or both nicotine and alcohol (NiAlvS). This approach enabled us to identify large scale gene and miRNA expression profiling in the target neurons of the interested brain area (i.e., the VTA).

Table 1 demonstrates that we organized the samples using three labels: 1) perinatal alcohol exposure, 2) perinatal nicotine exposure, and 3) DA cells to represent 4 different conditions (nicotine, alcohol, co-exposure and saline) for each neuron type (DA vs NDA). For each sample set, we performed five trials using different random seeds. In each trial, we performed 5000 iterations that calculated the feature importance using the Random Forest classifier. The feature importance was calculated as the average of all 5000 iterations. The hyperparameter of the Random Forest classifier was optimized using grid search. We used 5000 iterations for each trial to ensure converge of the average feature importance.

B. SELECTING RELEVANT FEATURES EFFICIENTLY

We built a Bayesian network to represent the running-time advantage of feature selection. The Bayesian network was buit using bnlearn [35] and the expression levels of the features as vertices in a directed acyclic graph (DAG); the relationships among the expression levels were predicted as arcs that connected the vertices. A hill climbing (HC) algorithm adds one arc per iteration and was used to learn the net-

TABLE 2. KEGG pathways found using the top 20% important DEGs, and the corresponding genes identified in pathway analysis following perinatal nicotine exposure, following perinatal alcohol exposure, and following perinatal nicotine and alcohol exposure. *: Pathways that have been reported in our group's previous publications.

Perinatal nicotine exposure		Perinatal alcohol exposure		Perinatal nicotine-alcohol exposure	
Term	**p-value**	**Term**	**p-value**	**Term**	**p-value**
rno03010:Ribosome	1.58E-10	rno03010:Ribosome	3.99E-09	rno03010:Ribosome	6.87E-05
rno05034:Alcoholism	7.32E-04	rno04612:Antigen processing and presentation	9.18E-04	rno04612:Antigen processing and presentation	3.18E-03
rno04141:Protein processing in endoplasmic reticulum *	2.58E-03	rno04961:Endocrine and other factor-regulated calcium reabsorption	1.63E-03	rno04141:Protein processing in endoplasmic reticulum	3.96E-03
rno05322:Systemic lupus erythematosus	3.28E-03	rno05031:Amphetamine addiction	2.45E-03	rno04145:Phagosome	5.57E-03
rno04612:Antigen processing and presentation	4.12E-03	rno04141:Protein processing in endoplasmic reticulum	3.96E-03	rno05416:Viral myocarditis	6.17E-03
rno04022:cGMP-PKG signaling pathway *	4.17E-03	rno04144:Endocytosis *	5.75E-03	rno04150:mTOR signaling pathway	7.93E-03
rno00061:Fatty acid biosynthesis	4.67E-03	rno05034:Alcoholism *	5.97E-03	rno04713:Circadian entrainment *	9.96E-03
rno04360:Axon guidance *	6.23E-03	rno05416:Viral myocarditis	6.17E-03	rno05332:Graft-versus-host disease	1.11E-02
rno04710:Circadian rhythm	1.77E-02	rno04260:Cardiac muscle contraction	8.28E-03	rno04921:Oxytocin signaling pathway *	1.12E-02
rno01212:Fatty acid metabolism	2.04E-02	rno05332:Graft-versus-host disease	1.11E-02	rno04514:Cell adhesion molecules (CAMs)	1.19E-02
rno04921:Oxytocin signaling pathway	2.55E-02	rno04514:Cell adhesion molecules (CAMs)	1.19E-02	rno05034:Alcoholism *	1.42E-02
rno04380:Osteoclast differentiation *	3.46E-02	rno05330:Allograft rejection	1.50E-02	rno04911:Insulin secretion	1.45E-02
rno04145:Phagosome	4.17E-02	rno05168:Herpes simplex infection	1.50E-02	rno05330:Allograft rejection	1.50E-02
rno04978:Mineral absorption	4.90E-02	rno04142:Lysosome	2.00E-02	rno04940:Type I diabetes mellitus	2.27E-02
rno00310:Lysine degradation	5.05E-02	rno04940:Type I diabetes mellitus	2.27E-02	rno05320:Autoimmune thyroid disease	2.42E-02
rno04330:Notch signling pathway	5.05E-02	rno05320:Autoimmune thyroid disease	2.42E-02	rno04260:Cardiac muscle contraction	2.58E-02
rno05012:Parkinson's disease	5.09E-02	rno05322:Systemic lupus erythematosus	2.54E-02	rno04972:Pancreatic secretion	2.65E-02
rno05410:Hypertrophic cardiomyopathy (HCM)	5.40E-02	rno04921:Oxytocin signaling pathway *	2.65E-02	rno04961:Endocrine and other factor-regulated calcium reabsorption	2.98E-02
rno05168:Herpes simplex infection	5.51E-02	rno04713:Circadian entrainment *	2.79E-02	rno05031:Amphetamine addiction	3.14E-02
rno05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC) *	5.69E-02	rno04145:Phagosome *	2.82E-02	rno04261:Adrenergic signaling in cardiomyocytes	3.32E-02
rno05203:Viral carcinogenesis	6.12E-02	rno00310:Lysine degradation	4.35E-02	rno05231:Choline metabolism in cancer *	3.44E-02
rno04514:Cell adhesion molecules (CAMs)	6.16E-02	rno04964:Proximal tubule bicarbonate reclamation	4.68E-02	rno05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)	3.81E-02
rno04142:Lysosome	6.41E-02	rno04360:Axon guidance *	4.94E-02	rno04022:cGMP-PKG signaling pathway *	3.95E-02

work structure.As illustrated in **Figure 1a**, the running time after building a Bayesian Network increases exponentially when the feature size (also described as number of vertices) increases. This method will consume a PC (using 32.0 GB RAM and Intel Core i7-10875H CPU) and take >24 hours to perform one iteration of Bayesian network prediction when 2500 features are used. We implemented impurity-based feature importance in Random Forest to calculate the relative importance of the DEGs and the DEmiRs. When training a decision tree using a Random Forest, the feature importance

TABLE 2. ***(Continued.)*** **KEGG pathways found using the top 20% important DEGs, and the corresponding genes identified in pathway analysis following perinatal nicotine exposure, following perinatal alcohol exposure, and following perinatal nicotine and alcohol exposure. *: Pathways that have been reported in our group's previous publications.**

Perinatal nicotine exposure		Perinatal alcohol exposure		Perinatal nicotine-alcohol exposure	
Term	**p-value**	**Term**	**p-value**	**Term**	**p-value**
rno05414:Dilated cardiomyopathy	6.81E-02	rno04380:Osteoclast differentiation	5.15E-02	rno03013:RNA transport	4.26E-02
rno05166:HTLV-I infection	7.19E-02	rno04728:Dopaminergic synapse *	5.15E-02	rno00310:Lysine degradation	4.35E-02
rno04911:Insulin secretion	7.60E-02	rno05203:Viral carcinogenesis *	5.65E-02	rno04964:Proximal tubule bicarbonate reclamation	4.68E-02
rno04919:Thyroid hormone signaling pathway	7.71E-02	rno04010:MAPK signaling pathway *	5.73E-02	rno04142:Lysosome *	4.74E-02
rno04024:cAMP signaling pathway *	8.00E-02	rno04916:Melanogenesis *	7.60E-02	rno04380:Osteoclast differentiation	5.15E-02
rno03013:RNA transport	8.23E-02	rno04310:Wnt signaling pathway	8.05E-02	rno04728:Dopaminergic synapse*	5.15E-02
rno05215:Prostate cancer	8.44E-02	rno05414:Dilated cardiomyopathy	9.29E-02	rno04010:MAPK signaling pathway	5.73E-02
rno05200:Pathways in cancer *	8.48E-02	rno00061:Fatty acid biosynthesis	9.39E-02	rno04024:cAMP signaling pathway	6.05E-02
rno05416:Viral myocarditis	9.33E-02	rno04710:Circadian rhythm	9.96E-02	rno05168:Herpes simplex infection	6.06E-02
rno00071:Fatty acid degradation *	9.39E-02			rno04970:Salivary secretion	6.35E-02
				rno05200:Pathways in cancer *	6.47E-02
				rno05205:Proteoglycans in cancer *	7.30E-02
				rno05410:Hypertrophic cardiomyopathy (HCM)	7.74E-02
				rno05010:Alzheimer's disease	7.78E-02
				rno04144:Endocytosis *	8.00E-02
				rno04270:Vascular smooth muscle contraction	8.52E-02
				rno04723:Retrograde endocannabinoid signaling	8.95E-02
				rno05210:Colorectal cancer	8.96E-02
				rno04071:Sphingolipid signaling pathway	9.15E-02
				rno05414:Dilated cardiomyopathy	9.29E-02
				rno00061:Fatty acid biosynthesis	9.39E-02

can be measured by how much the feature will decrease the weighted impurity (or information entropy) over various trees [36]. This process runs in polynomial time. As shown in **Figure 1b**, the running time of calculating the feature importance using the Random Forest model increases linearly against the feature size. The short running time facilities the generation of stable feature importance results. As shown in **Figure 1c and 1d**, the standard deviation and root-mean-square error (RMSE) of the importance of all features converges after a large number of iterations. It takes typically 2500 and 1000 iterations for the standard deviation and the RMSE to plateau (which is defined as the slope of the moving average of 1000 data points to approach 0), respectively. To ensure a stable result and to reduce overfitting, we used 5000 iterations with 5-fold cross-validations in the feature importance calculation for all samples.

C. SELECTING REPRODUCABLE FEATURES

Using the experimental conditions as the prediction labels, we then performed a grid search over various hyper-parameters in the Radom Forest model for each of the datasets (i.e. AlvS, NivS, and NiAlvS), to ensure precise predictions. The details on the hyperparameter search can be found in the Methods section. **Figures 2a-2c** illustrate the confusion matrix for Random Forest based feature importance selection on the DEGs and DEmiRs from the AlvS group. The hyperparameter-tuned model yielded a 93% true positive rate and 100% true negative rate when predicting whether a neuron had been perinatally exposed to alcohol. For nicotine-treated neurons, the true positive rate was 83% and the true negative rate was 99%. Finally, this model had a 71% true positive rate and 65% true negative rate when predicting on whether a neuron was a DA neuron. The subset accuracy (i.e., the percentage rate of all three labels to be correctly predicted) was 68 ± 15%. For the NiAlvS and the NivS data group, the hyperparameter-tuned model yielded similar classification accuracy (data not shown)

This method is rather stable. As shown in **Figure 2d**, out of the 5000 features in this perinatal alcohol dataset, we tracked the rank of the ten most important (i.e., highest feature importance score) and the ten least important features from five machine learning trials. The important features obtained a high importance score in all five trials, while the less-important features reproducibly obtained a lower score in all five trials.

D. COMPARISON WITH EXISTING KEGG PATHWAYS

To examine the effectiveness of our feature selection method, we selected the top 20% most important DEGs and DEmiRs to estimate gene pathways using the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways. The results are summarized in **Table 2.** Then, we compared the identified pathways with those identified using the full DEGs and DEmiRs data sets in [29], [31]. We noticed that we were able to detect multiple pathways that have been identified in our previous publications using this proposed method. We identified enriched pathways including Protein processing in endoplasmic reticulum, cGMP-PKG signaling, Osteoclast differentiation, Axon guidance, Arrhythmogenic right ventricular cardiomyopathy (ARVC), cAMP signaling, and Pathways in cancer from the NivS dataset ($p < 0.05$), similar to our group's previous report which used all the DEGs and DEmiRs [31]. We also identified enriched pathways including Endocytosis, Alcoholism, Oxytocin signaling, Circadian entrainment, Phagosome, Axon guidance, Dopaminergic synapse, Viral carcinogenesis, MAPK signaling pathway and Melanogenesis from the AlvS dataset ($p < 0.05$). Using the NiAlvS dataset, we identified enriched pathways including Circadian entrainment, Oxytocin signaling pathway, Dopaminergic synapse, Alcoholism, Choline metabolism in cancer, cGMP-PKG signaling pathway, Lysosome, Pathways in cancer, Proteoglycans in cancer, and Endocytosis ($p < 0.05$], in agreement with [29].

E. VALIDATION OF RESULTS WITH EXISTING MODELS

We also analzyed the same top 20% DEGs and DEmiRs data sets in the NivS, AlvS and NiAlvS groups using the commonly used parametric (Pearson [12]) and nonparametric (Spearman [36]) corellation methods and the Random KNN method [37] to identify potential gene pathways as identified using the KEGG pathways.

Using the top 20% DEGs and DEmiRs data sets in the NivS group, the Pearson correlation method was able to identify a few enriched pathways including cGMP-PKG signalling, Osteoclast differentiation, Axon guidance, Arrhythmogenic right ventricular cardiomyopathy (ARVC) and cAMP signalling ($p < 0.05$). However, the Random KNN was able to identify protein processing in endoplasmic reticulum ($p < 0.05$). The Spearman correlation method failed to identify any pathways.

Using the top 20% DEGs and DEmiRs data sets in the AlvS group, the Pearson correlation method was able to identify two pathways including alcoholism and viral carcinogenesis ($p < 0.05$). Additionally, the Spearman correlation was able to identify four pathways including Endocytosis, Phagosome, axon guidance, and viral carcinogenesis ($p < 0.05$). The Random KNN was unable to identify any pathways.

Using the top 20% DEGs and DEmiRs data sets in the AlNivS group, the Pearson correlation method was able to identify three pathways including alcoholism choline metabolism in cancer, and proteoglycans in cancer ($p < 0.05$). The Spearman correlation was able to identify four pathways including cGMP-PKG signalling pathway, lysosome, proteoglycans in cancer, and endocytosis ($p < 0.05$). Finally, the Random KNN identified two pathways including pathways in cancer and proteoglycans in cancer ($p < 0.05$).

These results indicate that the Random Forest method performed better that the Pearson, Spearman and Random KNN methods on the same top 20% DEGs and DEmiRS data sets in the NivS, AlvS and the NiAlvS groups to estimate gene pathways using the KEGG pathways. The Random Forest method was able to identify more pathways identified using the all DEGs and DEmiRS data sets in each group [29], [31].

IV. DISCUSSION

Building a mathematical model over a large set of features could be a computationally-challenging problem. For example, when building a Bayesian network, the expression level of the DEGs and the DEmiRs are treated as vertices in a Directed acyclic graph (DAG), and the relationships (to be predicted) among the DEGs and the DEmiRs will be represented by arcs that connect the vertices. Assume that we start to predict the Bayesian Network model with an empty DAG (that is, a DAG that contains all the vertices but no arcs), and that we use a hill climbing (HC) algorithm to add one arc per iteration, this requires $O(N^2)$ model comparisons. Thus, the overall time complexity of the HC algorithm is to the scale of $O(cN^3)$ model comparisons [5]. The NP-hardness of learning Bayesian networks have been generally accepted by the researchers in the machine-learning community [39]. Similarly, in most machine learning studies using microarray datasets, feature selection is needed.

Other well-used methods to perform gene selection include permutation-based feature selection, which randomly reshuffles the data and balances the influence of each feature to the model performance. Compared with the Gini-based feature selection, the permutation algorithm is usually significantly slower [40], [41]. The Gini-based method may not be effective when the potential predictor variables vary in their scale of measurement or their number of categories [42]. However, this may not be applicable to the analysis of microarray data, which is homogeneous.

Notably, our methods have suggested several pathways that were also previously identified by our group using the enriched KEGG pathways (e.g., the Dopaminergic synapse pathway from the AlvS and the NiAlvS dataset) [29]. The Dopaminergic synapse pathway describes the release of DA neurotransmitter. According to the major hypothesis of drug reinforcement, the reinforcing effect of addiction is believed to be conveyed through the activation of the meso-corticolimbic DA system. Stimulation of VTA DA neurons via alcohol and/or nicotine administration results in the release of DA in the NAc, which is believed to describe how the DA synapse pathway plays a role in the reinforcing effect [29], [43]. The enrichment of this pathway in our current results demonstrates that perinatal alcohol and/or nicotine exposure leads to genetic alterations in VTA DA neurons that are in accordance with addiction mechanisms.

Additionally, we found the Axon guidance pathway from the NivS and the AlvS datasets, which was highlighted in our group's recent work [29]. Axon guidance is an important pathway that regulates the migration of an axon is directed to a specific target. Following perinatal alcohol and/or nicotine exposure, we found Semaphorin 3F, 4A, 4D, 4G, and 6D, which are genes that belong to the Semaphorin axon guidance pathway [44].

V. SUMMARY

In this study, we proposed a fast and reproducible method using a Random Forest classifier to perform impurity-based feature selection over a microarray datasets with a large dimensionality. Using our method, we successfully identified the Glutamatergic synapse and the Axon guidance pathways, which were previously reported to be enriched following perinatal nicotine or alcohol-nicotine exposure. In the future, we are interested in performing additional unsupervised mathematical studies, such as Bayesian Network analysis, over the selected important genes and miRNAs.

ACKNOWLEDGMENT

The authors would like to thank Dr. Charlotte Mae K. Waits for editing the article.

REFERENCES

[1] T. Lenoir and E. Giannella, "The emergence and diffusion of DNA microarray technology," *J. Biomed. Discovery Collaboration*, vol. 1, no. 1, pp. 1–39, 2006.

[2] C. Yoo, V. Thorsson, and G. F. Cooper, "Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data," in *Proc. Biocomput.*, Dec. 2001, pp. 498–509.

[3] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 5, pp. 971–989, Sep. 2016.

[4] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinf.*, vol. 2015, pp. 1–13, Jun. 2015.

[5] M. Scutari, C. Vitolo, and A. Tucker, "Learning Bayesian networks from big data with greedy search: Computational complexity and efficient implementation," *Statist. Comput.*, vol. 29, no. 5, pp. 1095–1108, Sep. 2019.

[6] K.-S. Lin and C.-F. Chien, "Cluster analysis of genome-wide expression data for feature extraction," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3327–3335, Mar. 2009.

[7] C. Maugis, G. Celeux, and M.-L. Martin-Magniette, "Variable selection for clustering with Gaussian mixture models," *Biometrics*, vol. 65, no. 3, pp. 701–709, Sep. 2009.

[8] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *J. Amer. Stat. Assoc.*, vol. 105, no. 490, pp. 713–726, 2010.

[9] L.-Y. Chuang, C.-H. Yang, and C.-H. Yang, "Tabu search and binary particle swarm optimization for feature selection using microarray data," *J. Comput. Biol.*, vol. 16, no. 12, pp. 1689–1703, Dec. 2009.

[10] S. Loscalzo, L. Yu, and C. Ding, "Consensus group stable feature selection," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2009, pp. 567–576.

[11] L.-Y. Chuang, H.-W. Chang, C.-J. Tu, and C.-H. Yang, "Improved binary PSO for feature selection using gene expression data," *Comput. Biol. Chem.*, vol. 32, no. 1, pp. 29–38, Feb. 2008.

[12] P. Rudra, W. J. Shi, P. Russell, B. Vestal, B. Tabakoff, P. Hoffman, K. Kechris, and L. Saba, "Predictive modeling of miRNA-mediated predisposition to alcohol-related phenotypes in mouse," *BMC Genomics*, vol. 19, no. 1, pp. 1–12, Dec. 2018.

[13] R. Xu, S. Damelin, B. Nadler, and D. C. Wunsch, "Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps," *Artif. Intell. Med.*, vol. 48, nos. 2–3, pp. 91–98, Feb. 2010.

[14] R. Luss and A. d'Aspremont, "Clustering and feature selection using sparse principal component analysis," *Optim. Eng.*, vol. 11, no. 1, pp. 145–157, Feb. 2010.

[15] C. Boutsidis, M. W. Mahoney, and P. Drineas, "Unsupervised feature selection for principal components analysis," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2008, pp. 61–69.

[16] Y. B. Kim and J. Gao, "Unsupervised gene selection for high dimensional data," in *Proc. 6th IEEE Symp. Bioinf. BioEng. (BIBE)*, Oct. 2006, pp. 227–234.

[17] Q. Shen, Z. Mei, and B.-X. Ye, "Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification," *Comput. Biol. Med.*, vol. 39, no. 7, pp. 646–649, Jul. 2009.

[18] A. J. Ferreira and M. A. T. Figueiredo, "Efficient feature selection filters for high-dimensional data," *Pattern Recognit. Lett.*, vol. 33, no. 13, pp. 1794–1804, Oct. 2012.

[19] Y. H. Yang, "Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Res.*, vol. 30, no. 4, p. 15e, Feb. 2002.

[20] C. Workman, L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielser, H. H. Saxild, C. Nielsen, S. Brunak, and S. Knudsen, "A new non-linear normalization method for reducing variability in DNA microarray experiments," *Genome Biol.*, vol. 3, no. 9, pp. 1–16, 2002.

[21] H. Deng and G. Runger, "Feature selection via regularized trees," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–8.

[22] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. Mayer, and H. W. Mewes, "Gene selection from microarray data for cancer classification-a machine learning approach," *Comput. Biol. Chem.*, vol. 29, no. 1, pp. 37–46, 2005.

[23] H. Aydadenta, "A clustering approach for feature selection in microarray data classification using random forest," *J. Inf. Process. Syst.*, vol. 14, no. 5, pp. 1167–1175, 2018.

[24] A. Anaissi, P. J. Kennedy, and M. Goyal, "Feature selection of imbalanced gene expression microarray data," in *Proc. 12th ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distrib. Comput.*, Jul. 2011, pp. 73–78.

[25] S. Turgut, M. Dagtekin, and T. Ensari, "Microarray breast cancer data classification using machine learning methods," in *Proc. Electr. Electron., Comput. Sci., Biomed. Eng. Meeting (EBBT)*, Apr. 2018, pp. 1–3.

[26] D. Paul, R. Su, M. Romain, V. Sébastien, V. Pierre, and G. Isabelle, "Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier," *Computerized Med. Imag. Graph.*, vol. 60, pp. 42–49, Sep. 2017.

[27] Z. Li, H. Wang, Y. Zhang, and X. Zhao, "Random forest-based feature selection and detection method for drunk driving recognition," *Int. J. Distrib. Sensor Netw.*, vol. 16, no. 2, 2020, Art. no. 1550147720905234.

[28] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinf.*, vol. 10, no. 1, pp. 1–16, 2009.

[29] T. Kazemi, S. Huang, N. G. Avci, C. M. K. Waits, Y. M. Akay, and M. Akay, "Investigating the influence of perinatal nicotine and alcohol exposure on the genetic profiles of dopaminergic neurons in the VTA using miRNA–mRNA analysis," *Sci. Rep.*, vol. 10, no. 1, pp. 1–23, Dec. 2020.

[30] S. D. Ramachandran, K. Schirmer, B. Münst, S. Heinz, S. Ghafoory, S. Wölfl, K. Simon-Keller, A. Marx, C. I. Øie, M. P. Ebert, H. Walles, J. Braspenning, and K. Breitkopf-Heinlein, "*In vitro* generation of functional liver organoid-like structures using adult human cells," *PLoS ONE*, vol. 10, no. 10, Oct. 2015, Art. no. e0139345, doi: 10.1371/journal.pone.0139345.

[31] R. F. Keller, A. Dragomir, F. Yantao, Y. M. Akay, and M. Akay, "Investigating the genetic profile of dopaminergic neurons in the VTA in response to perinatal nicotine exposure using mRNA-miRNA analyses," *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, Dec. 2018.

[32] R. F. Keller, T. Kazemi, A. Dragomir, Y. M. Akay, and M. Akay, "Comparison between dopaminergic and non-dopaminergic neurons in the VTA following chronic nicotine exposure during pregnancy," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, Dec. 2019.

[33] A. R. Patten, C. J. Fontaine, and B. R. Christie, "A comparison of the different animal models of fetal alcohol spectrum disorders and their use in studying complex behaviors," *Frontiers Pediatrics*, vol. 2, p. 93, Sep. 2014.

[34] A. L. Blum and R. L. Rivest, "Training a 3-node neural network is NP-complete," *Neural Netw.*, vol. 5, no. 1, pp. 117–127, Jan. 1992.

[35] M. Scutari, "Learning Bayesian networks with the bnlearn R package," 2009, *arXiv:0908.3817*. [Online]. Available: http://arxiv.org/abs/0908.3817

[36] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[37] C. Wissler, "The spearman correlation formula," *Science*, vol. 22, no. 558, pp. 309–311, Sep. 1905.

[38] S. Li, E. J. Harner, and D. A. Adjeroh, "Random KNN feature selection—A fast and stable alternative to random forests," *BMC Bioinf.*, vol. 12, no. 1, pp. 1–11, Dec. 2011.

[39] M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of Bayesian networks is NP-hard," *J. Mach. Learn. Res.*, vol. 5, pp. 1–44, Jan. 2004.

[40] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, May 2010.

[41] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the Gini importance?" *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, Nov. 2018.

[42] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinf.*, vol. 8, no. 1, pp. 1–21, Dec. 2007.

[43] R. C. Pierce and V. Kumaresan, "The mesolimbic dopamine system: The final common pathway for the reinforcing effect of drugs of abuse?" *Neurosci. Biobehav. Rev.*, vol. 30, no. 2, pp. 215–238, Jan. 2006.

[44] T. W. Yu and C. I. Bargmann, "Dynamic regulation of axon guidance," *Nature Neurosci.*, vol. 4, no. S11, pp. 1169–1176, Nov. 2001.

HUI XIA received the B.S. degree in molecular biology and biochemistry from China Agricultural University, China, in 2008, and the M.S. and Ph.D. degrees in biomedical engineering from Louisiana Tech University, in 2012 and 2013, respectively. In 2013, he joined the University of Houston as a Postdoctoral Fellow. His research interests include machine learning, brain modeling, BioMEMS, and microfluidics.

YASEMIN M. AKAY (Senior Member, IEEE) received the B.S. degree in pharmaceutical sciences from Hacettepe University, Ankara, Turkey, in 1980, and the M.S. and Ph.D. degrees in biomedical engineering from Rutgers University, Piscataway, NJ, USA, in 1991 and 1998, respectively. She was a Postdoctoral Fellow with the Department of Physiology and Pharmacology, Dartmouth Medical School, and the Department of Physiology and Biophysics, School of Medicine, Boston University. She is an Instructional and a Research Associate Professor with the Department of Biomedical Engineering, Cullen College of Engineering, University of Houston, Houston, TX, USA. Her current research interests include molecular neuro-engineering, neural growth, and neurodegeneration. She was an Assistant Editor of the IEEE Book Series, from September 2001 to May 2004. She has been a Managing Editor of the *Wiley Encyclopedia of Biomedical Engineering*, since May 2004.

METIN AKAY (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from Boğaziçi University, Istanbul, Turkey, in 1981 and 1984, respectively, and the Ph.D. degree from Rutgers University, Piscataway, NJ, USA, in 1990. He is currently the Founding Chair of the Department of Biomedical Engineering, University of Houston. His current research interests include investigation of nicotine and alcohol addiction at the molecular, cellular, and system levels during maturation and the development of brain chips for precision medicine. He was the Founding Chair of the Annual International Summer School on BIO-X sponsored by the National Science Foundation (NSF) and technically co-sponsored by the IEEE EMBS, Satellite Conference on Emerging Technologies in biomedical engineering. He is the Founding Chair of the International IEEE Conference on Neural Engineering, in 2003, and the President of the IEEE EMBS.