Permutation-based Identification of Important Biomarkers for Complex Diseases via Black-box Models

Xinlei Mi[1], Baiming Zou[2], Fei Zou[2], & Jianhua Hu[1*]

April 24, 2020

1 Department of Biostatistics, Columbia University

2 Department of Biostatistics, University of North Carolina at Chapel Hill

* Corresponding author: jh3992@cumc.columbia.edu

Abstract

Study of human disease remains challenging due to convoluted disease etiologies and complex molecular mechanisms at genetic, genomic, and proteomic levels. Many machine learning-based methods, including deep learning and random forest, have been developed and widely used to alleviate some analytic challenges in complex human disease studies. While enjoying the modeling flexibility and robustness, these model frameworks suffer from non-transparency and difficulty in interpreting the role of each individual feature due to their intrinsic black-box natures. However, identifying important biomarkers associated with complex human diseases is a critical pursuit towards assisting researchers to establish novel hypotheses regarding prevention, diagnosis and treatment of complex human diseases. Herein, we propose a Permutation-based Feature Importance Test (PermFIT) for estimating and testing the feature importance, and for assisting interpretation of individual feature in various black-box frameworks, including deep neural networks, random forests, and support vector machines. PermFIT (available at `https://github.com/SkadiEye/deepTL`) is implemented in a computationally efficient manner, without model refitting for each permuted data. We conduct extensive numerical studies under various scenarios, and show that PermFIT not only yields valid statistical inference, but also helps to improve the prediction

accuracy of black-box models with top selected features. With the application to the Cancer Genome Atlas (TCGA) kidney tumor data and the HITChip atlas BMI data, PermFIT clearly demonstrates its practical usage in identifying important biomarkers and boosting performance of black-box predictive models.

Introduction

With the advancement of high-throughput technologies, massive amounts of high-dimensional omics data have been generated and made available through large public databases due to great data sharing efforts by the research community, such as The Cancer Genome Atlas (TCGA) [1]. These data are valuable in elucidating the molecular mechanisms of disease phenotypes [2, 3]. However, study of complex human disease remains challenging due to convoluted disease etiologies and underlying intricate molecular mechanisms at genetic, genomic, and proteomic levels. Many popular machine learning algorithms, such as non-linear kernel support vector machines (SVMs), random forests (RFs), and deep neural networks (DNNs) in artificial intelligence areas, have been developed to build more powerful predictive models for biomedical and bio-omics data regarding clinical outcomes, e.g. drug response [4], and medical imaging classification [5]. While enjoying the modeling flexibility and robustness, these model frameworks suffer from non-transparency and difficulty in interpreting the role of each individual feature due to their intrinsic black-box natures, compared with those more interpretable parametric models, such as linear regressions, logistic regressions, and decision trees. Nonetheless, identifying important biomarkers associated with complex human disease is a critical pursuit towards assisting researchers to establish novel hypotheses regarding prevention, diagnosis and treatment of complex human diseases. Accurate identification of important biomarkers associated with complex human disease not only provides valuable insights into their underlying genetic architecture and disease etiology but also offers great potentials for early disease diagnosis, improved precision medicine, innovative treatment development, and accurate disease risk and progression prediction [6, 7].

To address the non-transparency in the association study between disease outcomes and predictors using black-box machine learning models, the feature importance score strategy has been proposed and extensively investigated [8–14], including surrogate models [8], conditional randomization tests (CRTs) [11], knockoff

models (i.e. model-X) [11, 13], and permutation-based feature importance [9]. Surrogate modeling methods approximate the black-box models by using explanatory surrogate models, such as linear models or decision trees. While enjoying the great flexibility in choosing the surrogate models, the feature importance is still restricted to the selected explanatory models that might be misspecified [14]. Both CRT and model-X knockoff were proposed in Candes et al. [11], while CRT is less preferred due to its expensive computational cost. Model-X knockoff is more computationally efficient in performing feature importance test via constructing knockoff features. Recently, model-X knockoff was adopted for DNN [13] models. Tansey et al. [12] proposed the holdout randomization test (HRT) to reduce the computational cost of CRT via avoiding model re-fitting.

The overall disadvantage of CRT, HRT and model-X knockoff is that they all depend on the assumption of a known covariance structure [11]. When the covariance structure is not accurately estimated, their performance could be severely impacted [15]. The permutation-based feature importance learning method, another popular approach for feature selection, measures the change of prediction errors due to the shuffling of a feature. The larger the increase of prediction errors is, the more impact a feature makes on the outcome of interest. However, unlike CRT, HRT, or model-X knockoff, permutation-based feature selection does not require prior knowledge of feature distributions and thus it is more statistically robust. Several permutation-based feature importance methods have been proposed, with applications mainly on random forests and DNNs [9, 10, 16]. These methods either do not conduct any statistical inference or can not offer valid inference on the feature importance. For example, Putin et al. [16] applied permutation-based importance scores to DNNs to identify biomarkers associated with human aging, but provided no formal statistical testing. Notably, Altmann et al. [10] proposed a corrected permutation-based importance score approach for random forest, which however, is difficult to be generalized to other black-box model frameworks.

To overcome the aforementioned challenges, we propose a general permutation-based feature importance test (abbreviated as PermFIT), for black-box models, which takes advantage of (i) permutation test coupled with cross-fitting to obtain a valid importance score test that properly controls the type-I error; and (ii) selecting important features from PermFIT to further improve the accuracy of the black-box predictive models. We implement PermFIT for the following black-box models, including DNN, RF, and SVM. More

specifically, PermFIT first approximates the function that maps features to the outcome, based on which, PermFIT then evaluates the importance score of each feature, defined as the expected increase of prediction errors due to the permutation of the feature. Computationally, the PermFIT framework does not require re-fitting the black-box models. In order to reduce the bias of important score estimation from the potential model overfitting, we adopt cross-fitting to ensure the validity of the test statistics. PermFIT is motivated by two benchmark data: the Reverse Phase Protein Arrays (RPPAs) data from three kidney cancer studies in The Cancer Genomic Atlas (TCGA) and the HITChip Atlas microbiome data regarding body mass index (BMI). However, PermFIT has broad applicability to a wide variety of biomedical data and beyond.

Results

To evaluate the performance of PermFIT, we first conduct comprehensive simulation studies under various scenarios with different sample sizes and correlation structures among features. Moreover, it is applied to two real-world datasets: the Reverse Phase Protein Arrays (RPPA) data from three kidney cancer studies in TCGA and the HITChip Atlas microbiome data. We apply PermFIT to three commonly used black-box machine learning methods: DNN [17, 18], RF [9], and SVM [19], denoted as PermFIT-DNN, PermFIT-RF and PermFIT-SVM, respectively. We also compare PermFIT with two other existing popular feature selection methods for RF and SVM: RF importance evaluation of Breiman [9] (denoted as Vanilla-RF, i.e. an ensemble approach based on decision trees), and SVM with recursive feature elimination [20] (denoted as RFE-SVM). RFE-SVM utilizes an importance score to rank input features, from which top features are selected. For each feature, Vanilla-RF provides an importance score estimate and its associated standard error, with which the statistical significance of the feature importance can be tested. We evaluate these methods as follows: i) we apply each method to the training data with all the input features, estimate the feature importance scores, p-values, and assess the type-I error; ii) we refit each model with its corresponding top ranked important variables, and re-evaluate its goodness-of-fit and prediction improvement.

Simulation Studies

We examine the performance of the proposed methods with the following simulation scenarios. First, we generate the continuous data from the following model,

$$Y = X_1 + 2\log\left(1 + 2X_{p_0+1}^2 + (X_{2p_0+1} + 1)^2\right) + X_{3p_0+1}X_{4p_0+1} + \epsilon,$$

where X is a p-dimensional random variable drawn from $\mathrm{MVN}(0, \Sigma)$, $p = 10p_0$, $p_0 = 10$, $\Sigma = \mathrm{diag}\{\Sigma_1, ..., \Sigma_{10}\}$, is a block-diagonal matrix, $\Sigma_1 = ... = \Sigma_{10} = \{\sigma_{ij}\}_{0<i,j\leq p_0}$, are $p_0 \times p_0$ matrices, $\sigma_{ij} = 1$ for $i = j$ and $\sigma_{ij} = \rho$ for $i \neq j$, and $\epsilon \sim \mathrm{N}(0, 1)$. N independent observations are drawn from the distribution of (Y, X) in the training set and 10,000 in the test set, which is used to evaluate model fitting performance. To mimic the real-world data, we introduce correlations among variables by blocks, and let one variable from each of the first 5 blocks have a signal. We define S_0 and S_1 as the sets that contain all the null features that are correlated and uncorrelated with the causal features, respectively, i.e. $S_0 = \{X_j : j \leq 5p_0 \text{ and } j \neq 1, p_0+1, 2p_0+1, 3p_0+1, 4p_0+1\}$, $S_1 = \{X_j : j > 5p_0\}$. We consider various simulation settings with different values of $\rho \in \{0, 0.2, 0.5, 0.8\}$, and $N \in \{1000, 5000\}$. Each simulation scenario is replicated 100 times.

The results are displayed in Figure 1 and Table 1. Figure 1a displays detailed feature importance scores generated from each method that we consider. Note that the estimated importance scores from PermFIT methods and Vanilla RF are in the same scale, while the one from RFE-SVM is not [20]. For X_1 whose effect is linear, the importance scores from PermFIT-DNN and PermFIT-SVM are higher, compared with those from RF-based framework due to the restricted tree-based modeling nature of RF. In addition, the RF-based framework can barely detect the interaction between X_{3p_0+1} and X_{4p_0+1} because the split rule in tree-based methods is less effective in dealing with such interactions. Expectedly, as the within-block correlation ρ increases, the estimated importance scores from all methods deviate further away from their estimands. However, PermFIT-SVM remains high power in detecting the true positive features. As ρ increases, it's noticeable that Vanilla-RF and PermFIT-SVM tend to identify the null features that are correlated with the causal features. Compared with Vanilla-RF, PermFIT-RF has fewer false positive discoveries. Overall, PermFIT-DNN provides the most precise and stable importance measure in differentiating the true positive

from null features.

The frequency (percentage) of the important variables detected by each method is presented in Table 1. For Vanilla-RF and PermFIT methods that provide p-values, the significance level is controlled at 0.05, while for RFE-SVM, the top 10 features with the largest importance scores are selected. First of all, at $\rho = 0$, PermFIT controls the rate of significance findings across all null features at around 0.05, suggesting that the type-I error is well controlled by PermFIT, while Vanilla-RF has the type-I error of 0.09, nearly double of PermFIT. When ρ increases to 0.5 or 0.8, RFE-SVM tends to select the null features that are correlated with the true causal features, or those in S_0 more frequently than X_{3p_0+1} and X_{4p_0+1}, the two causal variables that interact with each other, demonstrating its limited capability in detecting variables with interaction effects when correlation exists. In contrast, PermFIT-SVM is capable of identifying X_{3p0+1} and X_{4p0+1} consistently at a much higher frequency than all the null features. Compared with PermFIT-RF, Vanilla-RF has a higher power in detecting X_{3p0+1} and X_{4p0+1}, but also produces remarkably more false positive findings among features in S_0. For example, as ρ goes to 0.8 and $N = 1000$, it results in $> 80\%$ false positive rate in S_0, suggesting a far inferior feature selection performance. In all these scenarios, PermFIT-DNN can consistently identify causal features while controlling false positive findings at a much lower rate than those of Vanilla-RF, PermFIT-RF, and PermFIT-SVM.

Posterior to important feature selection, the prediction performance of the comparing models almost all gets improved. Figure 1b displays the mean squared prediction error (MSPE) of each model, (i) with full input features, respectively denoted as DNN, RF and SVM; and (ii) with top selected features from each method (except for RFE-SVM) at the significance level of 0.1, and top 10 features from RFE-SVM. Selected features help boosting the prediction accuracy of all models except the RFE-SVM, across all simulation scenarios. However, at $\rho = 0.8$ corresponding to high correlation among within-block input features, RFE-SVM fails to improve the model fitting over SVM because of feature selection failure, in particular, on X_{3p0+1} and X_{4p0+1}; its inferior performance to PermFIT-SVM is clearly observed. Moreover, PermFIT-RF outperforms Vanilla-RF in terms of MSPE, because the latter yields more false positives and cannot effectively reduce the feature dimension. We note that PermFIT-DNN consistently outperforms all other methods in comparison, due to its high success rate in identifying true positive features while maintaining a

ρ	Variable	$N = 1000$					$N = 5000$				
		PermFIT DNN	Vanilla RF	PermFIT RF	RFE SVM	PermFIT SVM	PermFIT DNN	Vanilla RF	PermFIT RF	RFE SVM	PermFIT SVM
0	X_1	100	100	100	100	100	100	100	100	100	100
	X_{p_0+1}	100	100	100	100	94	100	100	100	100	100
	X_{2p_0+1}	100	100	100	100	99	100	100	100	100	100
	X_{3p_0+1}	100	64	26	41	96	100	100	98	43	100
	X_{4p_0+1}	100	71	39	41	91	100	100	98	36	100
	S_0	5.4	9.4	5.3	6.5	6.1	5.0	9.3	4.9	7.1	5.0
	S_1	5.5	8.9	5.1	6.5	5.7	4.5	9.0	5.1	6.0	5.1
0.2	X_1	100	100	100	100	100	100	100	100	100	100
	X_{p_0+1}	100	100	100	100	100	100	100	100	100	100
	X_{2p_0+1}	100	100	100	100	100	100	100	100	100	100
	X_{3p_0+1}	100	68	36	27	98	100	100	94	0	100
	X_{4p_0+1}	100	69	35	22	98	100	100	93	0	100
	S_0	6.1	17.5	6.7	13.2	7.9	5.7	29.0	7.5	15.6	12.8
	S_1	5.5	10.6	5.1	1.2	5.2	4.9	11.3	4.8	0.0	4.7
0.5	X_1	100	100	100	100	100	100	100	100	100	100
	X_{p_0+1}	100	100	100	100	100	100	100	100	100	100
	X_{2p_0+1}	100	100	100	100	100	100	100	100	100	100
	X_{3p_0+1}	100	69	35	0	96	100	100	96	0	100
	X_{4p_0+1}	100	68	28	0	99	100	100	98	0	100
	S_0	9.3	59.3	14.1	15.6	15.9	9.5	87.9	33.8	15.6	35.3
	S_1	6.7	16.9	4.6	0.0	6.0	7.0	38.9	4.2	0.0	11.8
0.8	X_1	100	100	100	100	100	100	100	100	100	100
	X_{p_0+1}	100	100	100	100	100	100	100	100	100	100
	X_{2p_0+1}	100	100	100	100	100	100	100	100	100	100
	X_{3p_0+1}	100	90	54	0	96	100	100	99	0	100
	X_{4p_0+1}	100	88	49	0	98	100	100	99	0	100
	S_0	18.8	88.3	31.6	15.6	27.7	20.6	100.0	74.9	15.6	68.2
	S_1	9.7	37.3	4.0	0.0	10.5	9.2	93.8	3.2	0.0	34.8

Table 1: Simulation Results on Continuous Outcomes. We report the percentage of the important variables detected by each method, for five true causal features: X_1, X_{p_0+1}, X_{2p_0+1}, X_{3p_0+1}, X_{4p_0+1}, and two null feature sets: S_0 and S_1.

considerably low false positive rate at the same time.

We further conduct a simulation study on binary outcomes generated from the following model:

$$P(Y = 1|X) = \text{expit}\left(4X_1 + 8\log\left(1 + 2X_{p_0+1}^2 + (X_{2p_0+1} + 1)^2\right) + 4X_{3p_0+1}X_{4p_0+1} - 11\right),$$

where $\text{expit}(x) = 1/(1 + \exp(-x))$. All the other data structures, including X, are generated in the same way as in the continuous case. Similar conclusions are observed with the details presented in Table S1 and Figure S1 in the Supplement.

Figure 1: Simulation Results on Continuous Outcomes. a. Mean with 95% confidence interval bar of estimated feature importance for the five true causal features: X_1, X_{p_0+1}, X_{2p_0+1}, X_{3p_0+1}, X_{4p_0+1}, and two null feature sets: S_0 and S_1. b. Mean squared prediction error (MSPE) for methods in comparison. DNN, RF or SVM: specific modeling with all features; PermFIT-DNN, Vanilla-RF, PermFIT-RF, RFE-SVM or PermFIT-SVM: specific modeling after feature selection.

TCGA Kidney Tumor Data.

A large collection of clinical and multiple omics data have been made publicly available by TCGA research project [1]. In our analysis, we included three studies of kidney-related cancer types from TCGA research network: kidney renal clear cell carcinoma (KIRC, $N_1 = 537$), kidney renal papillary cell carcinoma (KIRP, $N_2 = 291$), and kidney chromophobe (KICH, $N_3 = 113$). We defined long-term survivor (LTS) as patients who survived more than five years after diagnosis, and short-term survivor (STS) as patients who died within 5 years. We aimed to predict the probability of a patient being in the LTS group and to identify significant biomarkers that contribute to classification of the LTS/STS status. We included 188 LTS and 178 STS subjects with the known survival status in our analysis. We focused our analysis on expression data of 118 proteins extracted from reverse phase protein arrays (RPPAs)—a highly sensitive, reproducible, and high-throughput proteomic method for protein expression profiling [21].

The negative $\log_{10}$(p-value)s and the estimated importance scores from each method are presented in Figure 2 and Supplementary Figure S3. Vanilla-RF and PermFIT models control the FDR at 0.1, and RFE-SVM selects 10 features with the largest importance scores. We notice that moderate correlations generally exist among the proteins (see Supplementary Figure S2). However, six proteins, SRC, RAF1, RB1, RPS6, YWHAZ and EGFR, are highly correlated and clustered together by hierarchical clustering in Figure 2. Among them, EGFR, YWHAZ RPS6, RB1 and SRC are identified by Vanilla-RF, and RPS6, RB1 and SRC are selected by RFE-SVM, while none of these biomarkers are selected by any PermFIT procedures. According to our observations in simulation studies, both Vanilla-RF and RFE-SVM tend to identify false positive biomarkers in the presence of high correlation among features, casting some doubts on the validity of their biomarker selection results.

Since the underlying genetic truth is unknown, we alternatively use the model performance improvement estimated via 5-fold cross-validation, randomly repeated for 100 times (see Figure 3a-b; Supplementary Table S2) as a surrogate measure for evaluating the relative quality of the selected features. Similar to the simulation study, we set the feature inclusion criteria on p-values smaller than 0.1 for Vanilla-RF and PermFIT methods, and top 20 features for RFE-SVM. PermFIT-RF improves the accuracy from 0.694 to 0.732 on average, while Vanilla-RF only improves to 0.713. Moreover, PermFIT-SVM elevates the accuracy

from 0.69 to 0.744, outperforming RFE-SVM (0.709). Similar to the simulation results, PermFIT-DNN achieves the highest accuracy. The same conclusion is further confirmed by area under the ROC curve (AUC) results. In summary, it is evident that PermFIT procedures consistently perform more efficient and accurate feature selection across various black-box frameworks.

On the identified biomarkers, four genes - CDKN1A, EIF4EBP1, INPP4B and SERPINE1 - are claimed by all the three PermFIT methods to be significantly associated with the survival status. Interestingly, all four genes have been reported to be cancer related. Especially, INPP4B, identified as the most significant biomarker by all the three methods (p-value $= 1.3E-05$ by PermFIT-DNN, $9.1E-07$ by PermFIT-RF, and $4.5E-05$ by PermFIT-SVM), encodes inositol polyphosphate-4-phosphatase, type II, a dual specificity phosphatase. Low INPP4B is recently reported to be associated with shorter survival in kidney clear cell, liver hepatocellular, and bladder urotheleal carcinomas, and with long survival in pancreatic adenocarcinoma [22]. It is also related to acute myeloid leukemia, breast cancer and bladder cancer [23–25]. SERPINE1 encodes plasminogen activator inhibitor-1, which plays an important role in various diseases, in particular, kidney pathology and renal cell cancer [26–28]. In addition, the CDKN1A encoded protein, CDK-interacting protein 1, was reported as a prognostic marker for renal cell cancer [29], and has an effect on kidney cancer cell death [30] as well as kidney cancer survival [31]. Similarly, EIF4EBP1 affects disease progression in renal cell carcinoma [32].

Moreover, the DNA repair protein XRCC1, identified by PermFIT-DNN and PermFIT-SVM, is shown to be associated with bladder cancer [33]. ANXA7, identified by PermFIT-DNN and PermFIT-RF, is reported to be associated with prostate cancer and breast cancer [34, 35], and its encoded protein has an impact on prostate cancer and breast cancer [36, 37]. Furthermore, MYH9 and NRG1 are identified by PermFIT-DNN. Myosin-9, encoded by MYH9, has been discussed for its role as a tumor suppressor [38], and NRG1 is also reported to be related to multiple cancer types [39, 40]. Lastly, PermFIT-RF identifies a novel gene, STK11 whose role in kidney cancer is unknown, however, it has been reported that inactivation of STK11 in lung adenocarcinomas is a common event [41].

HITChip Atlas Data.

In the HITChip Atlas study, the data was collected from 1006 adults in 15 western countries [42] by using the HITChip, and it is publicly available in R library "microbiome" [43]. Besides demographic and clinical variables, the HITChip Atlas data includes microbiome measurements from 130 taxonomic groups summarized at the genus level, which covered major types of human intestinal microbiota bacterial diversity. Many of the 130 taxonomic groups are highly correlated, which is reflected in the correlation heat map and the hierarchical clustering dendrogram (see Figure 4 and Supplementary Figure S4). We then investigated the importance of demographic factors, including gender and nationality, together with 129 microbial genus (1 was removed due to the use of compositional values), in predicting the baseline BMI level. Our analysis includes 900 subjects with BMI measurements.

The feature selection and biomarker identification criteria remain the same as those in the TCGA example. The improvement of the model performance from variable selection is presented in Figure 3c-d and Supplementary Table S2. Besides the MSPE, we report the Pearson correlation between the predicted and the true values. We notice that, with variable selection, RFE-SVM results in a worse performance and Vanilla-RF barely improves the RF model. High correlation among microbiome features leads to large inflation in importance scores from Vanilla-RF, corresponding to high false positive rate, as a result, fails to improve the model performance. We observe that the reduced model with selected features from RFE-SVM performs worse than the full model. Again, this is likely due to the fact that highly correlated biomarkers are falsely selected by RFE-SVM. For instance, *Streptococcus mitis et rel*, *Streptococcus bovis et rel* and *Streptococcus intermedius et rel*, highly correlated to each other, are among top 10 biomarkers identified by RFE-SVM. In contrast, PermFIT procedures yield remarkable improvements in all these black-box models, reflected in both MSPE and correlation.

Figure 4 and Supplementary Figure S5 show the negative $\log_{10}$(p-value)s and the importance scores estimated from each method. Among all the features, as expected, age is identified as the most significant factor. The nationality is also selected by PermFIT-DNN and PermFIT-SVM. Among all the microbiome features, *Megasphaera elsdenii* is identified by all the PermFIT methods. *M. elsdenii* is shown by prior studies as one of the ruminal and intestinal lacate- and sugar-fermenting species [44]. *M. elsdenii* is also

reported to massively reside in patients with an increase in BMI after bariatric surgery [45]. In addition, *Eggerthella lenta* is identified by PermFIT-SVM. *E. lenta* is not well-studied, but its potential role as an emerging pathogen has been increasingly recognized in years [46]. Lastly, uncultured clostridiales is identified by PermFIT-RF.

Discussions

Black-box models are difficult to distinguish the contribution of individual input features, though they enjoy the more robustness and flexibility in modeling complex human diseases as compared with parametric models. In this paper, we introduce PermFIT, a computationally efficient permutation-based feature importance test with applicability to various black-box models such as DNN, RF and SVM, to identify important features. The type-I error of the associated test is evaluated and shown to be properly controlled except when the null and true positive features are extremely correlated—a situation where all competing methods fail, while PermFIT-DNN and RFE-SVM perform the best. On the selection of true positive features, in the presence of high correlation among input features, PermFIT, in particular PermFIT-DNN yields much better performance, compared with Vanilla-RF and RFE-SVM. Also, as demonstrated by the applications to TCGA kidney cancer data and HITChip Atlas BMI data, PermFIT procedures further show the superior performance over Vanilla-RF and RFE-SVM. The latter two procedures severely suffer from false positive or negative findings, leading to inferior prediction performance with top selected features. In contrast, feature selection via PermFIT procedures remarkably improves the performance of the black-box predictive models. However, it is worth pointing out that the prediction improvement of PermFIT is restricted to the capability of each black-box model framework. For example, RF is relatively inefficient in modeling interaction terms, thus the performance of PermFIT-RF may be limited for complex traits with strong gene-gene interactions. Overall, PermFIT coupled with DNN consistently shows superior empirical performance.

The proposed analytical tool, PermFIT, is computationally efficient and has broad applicability in addressing real-world problems. It can be implemented and incorporated into all types of black-box models with different types of outcomes, and without the need of model refitting. PermFIT provides researchers a useful tool for deciphering complex genetic architecture and disease etiologies of complex traits.

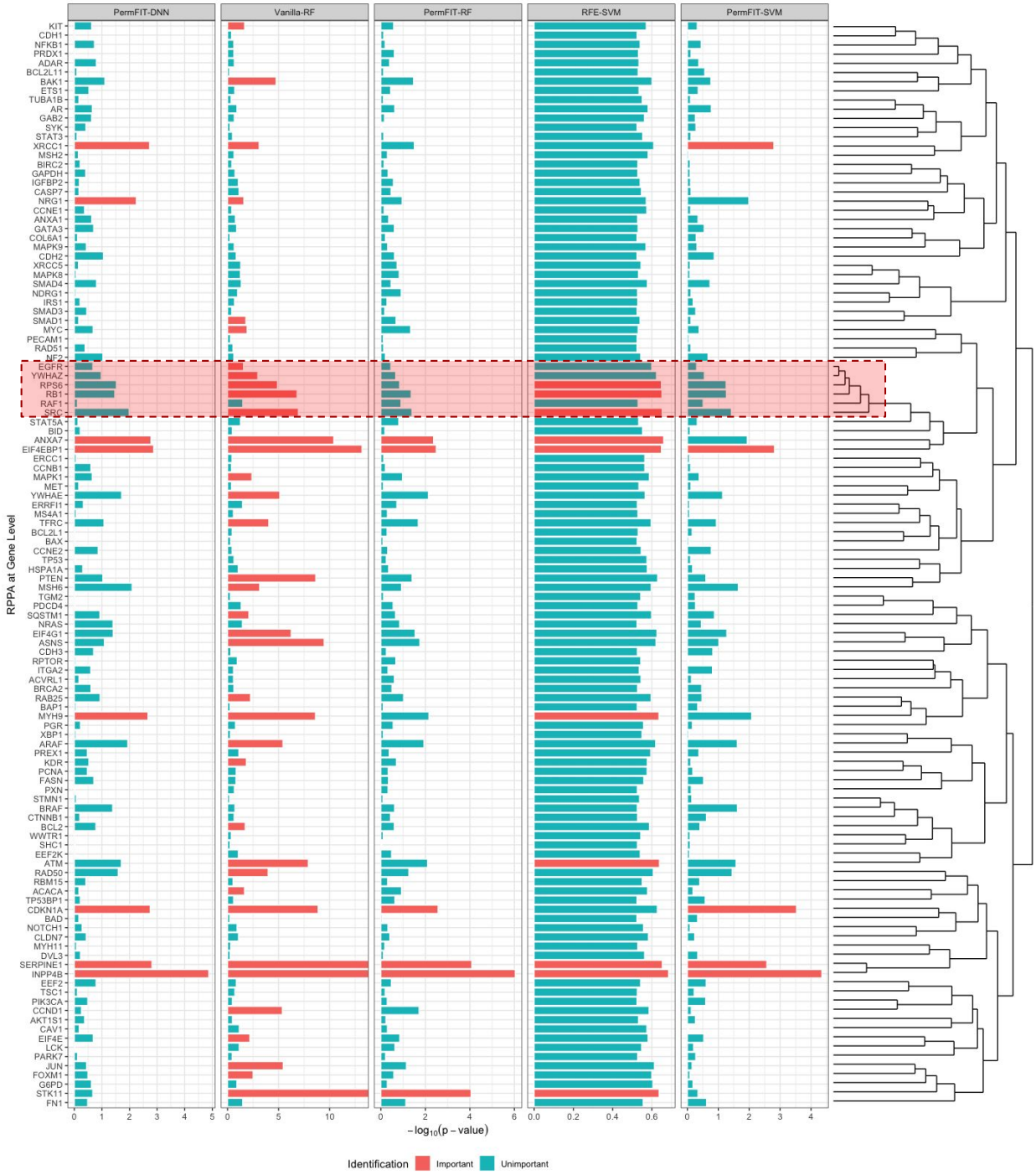

Figure 2: Negative $\log_{10}$p-values for TCGA kidney cancer data. Important features selected by each method is marked in red. Since RFE-SVM does not produce a p-value, its importance is presented instead, and 10 features with top importance scores are marked. The highly correlated features (see details from the dendrogram on the right) selected by RFE-SVM and Vanilla-RF, but not by PermFIT methods, are highlighted.

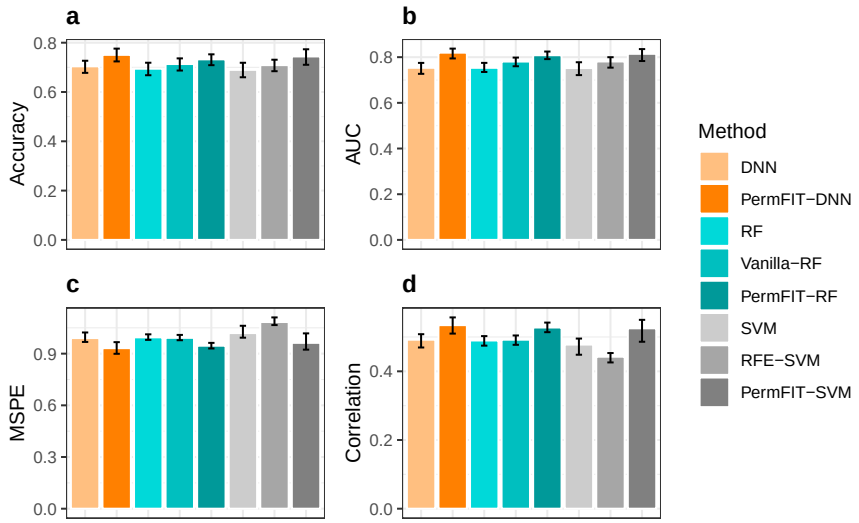

Figure 3: Model Performance Improvement from Feature Selection. a, b: 5-fold cross-validated prediction accuracy and AUC for TCGA kidney cancer data. c, d: 5-fold cross-validated MSPE and Pearson correlation (between true outcome and prediction) for HITChip Atlas data. 5-fold cross-validation evaluation is randomly repeated for 100 times.

Methods

PermFIT

We start the case with a continuous outcome. Let $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, where $X = (X_1, ..., X_p)$ is a p-dimensional covariate vector, the observation of the outcome variable Y is a continuous scalar, $\mathrm{E}(Y|X) = \mu(X)$, $\mu(\cdot)$ is an unknown mapping from $\mathcal{X}$ to $\mathcal{Y}$, the residues $\epsilon = Y - \mu(X)$ is independent of X and $0 < \sigma^2 = \mathrm{E}(\epsilon^2) < \infty$.

Feature Importance Scoring via Permutation. We define the feature importance score M_j for X_j, the jth feature in X $(j = 1, ..., p)$, as the expected squared difference between $\mu(X)$ and $\mu\left(X^{(j)}\right)$, where $X^{(j)} = (X_1, ..., X_{j-1}, X_{j'}, X_{j+1}, ..., X_p)$ is equal to X with the jth covariate replaced by a random vector $X_{j'}$ whose elements are independently drawn from the distribution of X_j. The importance score M_j can be expressed as,

$$M_j = \mathrm{E}_{X, X_{j'}} \left[\mu(X) - \mu\left(X^{(j)}\right)\right]^2. \tag{1}$$

The M_j is zero only when $\mu(X) \equiv \mu\left(X^{(j)}\right)$ on $\mathcal{X}$, implying that the jth element of X does not have any impact on $\mu(X)$; and is non-zero otherwise.

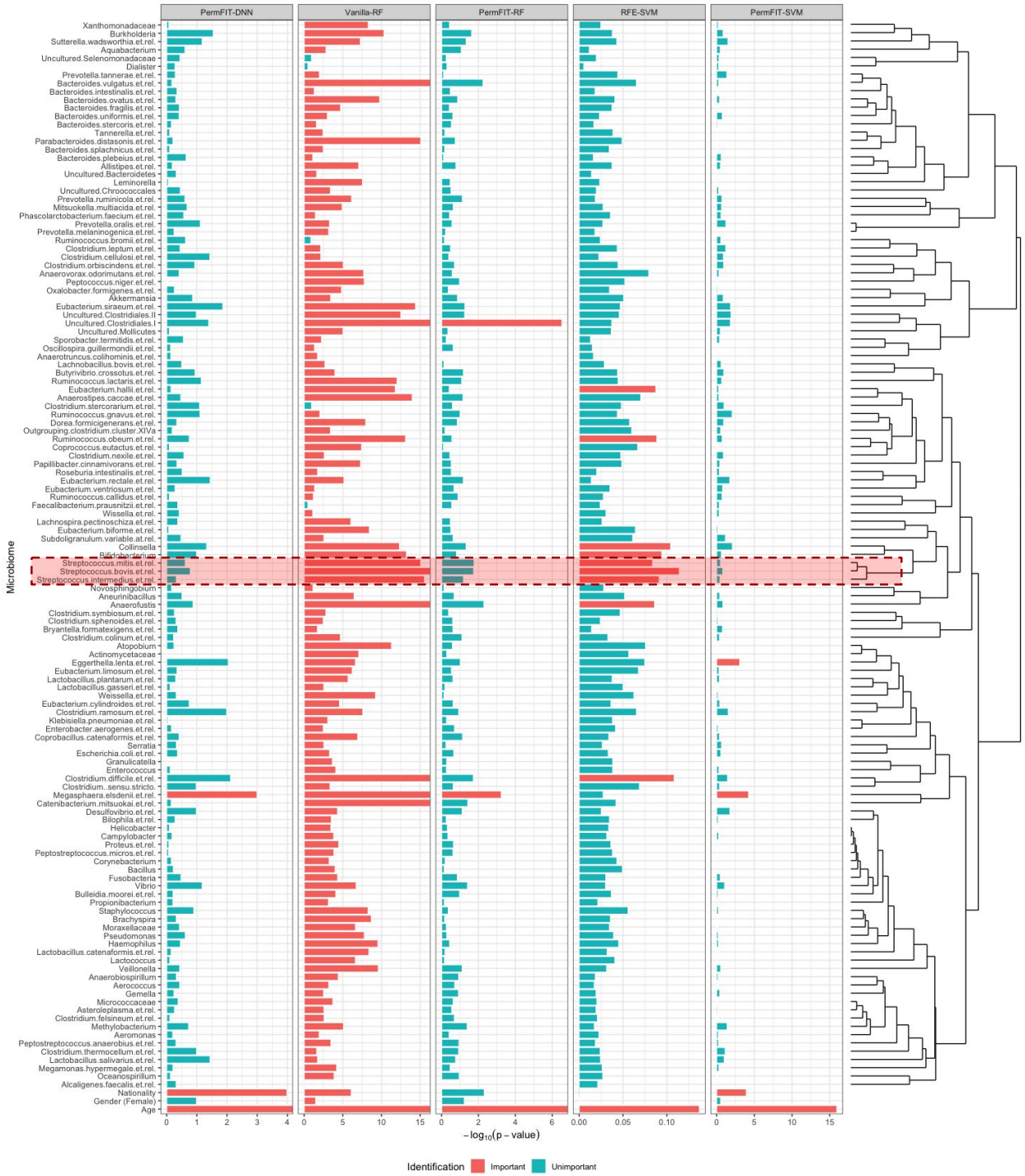

Figure 4: Negative $\log_{10}$p-values for HITChip Atlas data. Important features selected by each method is marked in red. Since RFE-SVM does not produce a p-value, its importance is presented instead, and 10 features with top importance scores are marked. The highly correlated features (see details from the dendrogram on the right) selected by RFE-SVM and Vanilla-RF, but not by PermFIT methods, are highlighted.

To obtain a clear understanding of M_j, we take the linear model as an example where $\mu(X) = X\boldsymbol{\beta} + \beta_0$, with $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ consisting of p parameters. Under the linear assumption, (1) becomes,

$$M_j = \mathrm{E}_{X,X_{j'}}(X_j - X_{j'})^2\beta_j^2 = 2\beta_j^2 \operatorname{Var}(X_j). \tag{2}$$

Here, (2) is proportional to the squared standardized coefficient, which has been recognized as a popular measure of variable importance in multiple linear regression.

Furthermore, M_j can be simply decomposed as follows,

$$M_j = \mathrm{E}_{X,X_{j'}}\left[Y - \mu\left(X^{(j)}\right)\right]^2 - \mathrm{E}_X[Y - \mu(X)]^2. \tag{3}$$

Ideally, given the true form of $\mu(\cdot)$, from (3), M_j could be estimated through permutation. Let $(Y_i, X_{i1}, ..., X_{ip})$, $(i = 1, ..., N)$ be N independent observations drawn from the distribution of $(Y, X_1, ..., X_p)$. A permutation on one covariate $X_j = (X_{1j}, ..., X_{Nj})$ is to randomly sample the elements in X_j without replacement to generate a permuted version of $X_j' = (X_{s_1,j}, ..., X_{s_N,j})$. The empirical permutation importance score is then,

$$M_j^{(P)} = \frac{1}{N}\sum_{i=1}^{N}\left[\left\{Y_i - \mu\left(X_{i\cdot}^{(j)}\right)\right\}^2 - \{Y_i - \mu(X_{i\cdot})\}^2\right], \tag{4}$$

where $X_{i\cdot} = (X_{i1}, ..., X_{ip})$ and $X_{i\cdot}^{(j)} = (X_{i1}, ..., X_{i,j-1}, X_{s_i,j}, X_{i,j+1}, ..., X_{ip})$. Let $M_j^{(P)} = \frac{1}{N}\sum_{i=1}^{N} M_{ij}^{(P)}$, where $M_{ij}^{(P)} = \left\{Y_i - \mu\left(X_{i\cdot}^{(j)}\right)\right\}^2 - \{Y_i - \mu(X_{i\cdot})\}^2$, then

$$\mathrm{E}\left[M_j^{(P)}\right] = \mathrm{E}\left[M_{ij}^{(P)}\right] = \frac{N-1}{N}M_j.$$

When N is large, M_j could be well approximated by $M_j^{(P)}$. Besides, $\operatorname{Var}\left[M_j^{(P)}\right] \approx \frac{1}{N}\operatorname{Var}\left[M_{ij}^{(P)}\right]$ which can be approximated by the empirical variance of $M_{ij}^{(P)}$.

Let $\widehat{\mu}(\cdot)$ be the fitted black-box approximator to $\mu(\cdot)$, according to (4), we propose to estimate M_j^P by

$$\widehat{M}_j^{(P)} = \frac{1}{N}\sum_{i=1}^{N}\left[\left\{Y_i - \widehat{\mu}\left(X_{i\cdot}^{(j)}\right)\right\}^2 - \{Y_i - \widehat{\mu}(X_{i\cdot})\}^2\right]. \tag{5}$$

If feature X_j is not associated with Y, then $\mu\left(X_{i\cdot}^{(j)}\right) = \mu(X_{i\cdot})$ with corresponding $M_j^{(P)} = 0$, and Equation (5) becomes,

$$\widehat{M}_j^{(P)} = \frac{1}{N}\sum_{i=1}^{N}\left[\left\{\mu(X_{i\cdot}) - \widehat{\mu}\left(X_{i\cdot}^{(j)}\right)\right\}^2 - \{\mu(X_{i\cdot}) - \widehat{\mu}(X_{i\cdot})\}^2 + 2\epsilon_i\left\{\widehat{\mu}(X_{i\cdot}) - \widehat{\mu}\left(X_{i\cdot}^{(j)}\right)\right\}\right]. \tag{6}$$

With the universal consistency, the three terms are expected to converge to zero as N goes to infinity. However, for data with a finite sample size, the black-box model $\widehat{\mu}(\cdot)$ may become overfit, leading to $\left\{\mu(X_{i\cdot}) - \widehat{\mu}\left(X_{i\cdot}^{(j)}\right)\right\}^2 > \{\mu(X_{i\cdot}) - \widehat{\mu}(X_{i\cdot})\}^2$ in estimating $M_j^{(P)}$. To overcome the bias issue, we employ cross-fitting strategy to separate the input data as the training and validation sets, with one set being utilized to obtain $\widehat{\mu}(\cdot)$ and the other set to estimate $\widehat{M}_j^{(P)}$. Let $\widehat{\mu}_T(\cdot)$ be the estimate of $\mu(\cdot)$ from the training set, and $\mathcal{D}_V = \{Y_i, X_{i\cdot}\}_{i=1}^{N_V}$ be the validation set,

$$\widehat{M}_j^{(P)} = \frac{1}{N_V}\sum_{i=1}^{N_V}\left[\left\{Y_i - \widehat{\mu}_T\left(X_{i\cdot}^{(j)}\right)\right\}^2 - \{Y_i - \widehat{\mu}_T(X_{i\cdot})\}^2\right], \tag{7}$$

$$\widehat{\mathrm{Var}}\left[\widehat{M}_j^{(P)}\right] = \frac{1}{N_V}\sum_{i=1}^{N_V}\left[\left\{Y_i - \widehat{\mu}_T\left(X_{i\cdot}^{(j)}\right)\right\}^2 - \{Y_i - \widehat{\mu}_T(X_{i\cdot})\}^2 - \widehat{M}_j^{(P)}\right]^2. \tag{8}$$

The one-sided p-value can be obtained by assuming normality. To increase the power of important feature identification, K-fold cross-fitting can be adopted. Here, we randomly divide the data into K folds, denoted as $V_1, ..., V_K$. For each of V_k, $k = 1, ..., K$, $\overline{V}_k$ denote the complementary set of V_k, which is used to fit the model $\widehat{\mu}_k(\cdot)$. Then

$$\widehat{M}_j^{(P,CV)} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in V_k}\left[\left\{Y_i - \widehat{\mu}_T\left(X_{i\cdot}^{(j)}\right)\right\}^2 - \{Y_i - \widehat{\mu}_k(X_{i\cdot})\}^2\right], \tag{9}$$

$$\widehat{\mathrm{Var}}\left[\widehat{M}_j^{(P,CV)}\right] = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in V_k}\left[\left\{Y_i - \widehat{\mu}_T\left(X_{i\cdot}^{(j)}\right)\right\}^2 - \{Y_i - \widehat{\mu}_k(X_{i\cdot})\}^2 - \widehat{M}_j^{(P,CV)}\right]^2. \tag{10}$$

The algorithm of PermFIT with cross-fitting is illustrated in Algorithm 1.

Algorithm 1 Algorithms for PermFIT

```
Randomly divide the data into K folds.
for k = 1 to K do
    Denote the data in k^th fold as V_k and the rest of the data as V̄_k.
    Build the black-box model with V̄_k, denoted as μ̂_k(·).
    for j = 1 to p do
        Calculate M̂_ij^(P,CV) for subjects in D_k.
    end for
end for
for j = 1 to p do
    Calculate M̂_j^(P,CV) and estimate V̂ar[M̂_j^(P,CV)]. Calculate p-value by assuming nomality.
end for
```

Binary Outcome. For binary outcome $Y \in \{0, 1\}$, we have $\mu(X) = \mathrm{E}(Y|X) = \Pr(Y = 1|X)$ and define M_j as the expectation of binomial deviance,

$$M_j = \mathrm{E}_{X,X_{j'}} \left[Y \log \left(\frac{\mu(X)}{\mu(X^{(j)})} \right) + (1 - Y) \log \left(\frac{1 - \mu(X)}{1 - \mu(X^{(j)})} \right) \right].$$

The empirical estimate of M_j can be similarly obtained by plugging in the estimate of $\mu(X^{(j)})$ and $\mu(X)$ as in the continuous data scenario.

DNN with Bootstrap Aggregating

In this paper, we use feedforward and fully-connected deep neural networks (DNNs) to approximate function $\mu(\cdot)$. The DNN model contains L hidden layers of $(n_1, ..., n_L)$ hidden nodes that transform the initial input covariates X to the estimation of the continuous outcome Y. Let $\boldsymbol{\theta}$ denote all the parameters in the DNN model, we essentially have the fitted DNN, $\widehat{\mu}(X; \boldsymbol{\theta})$, by minimizing the empirical risk function,

$$\arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \ell\{Y_i, \mu(X_i; \boldsymbol{\theta})\} + \lambda \Omega(\boldsymbol{\theta}),$$

where $\ell(\cdot, \cdot)$ is the loss function dependent on the outcome type, $\Omega(\boldsymbol{\theta})$ is a penalty on $\boldsymbol{\theta}$ and λ is a hyperparameter that controls the degree of regularization, via mini-batch stochastic gradient descent algorithm and Adam [47] to adjust the learning rate.

To increase the robustness and accuracy of DNNs, bootstrap aggregating (bagging) is applied [48]. Besides, due to the randomness of initial parameters, some DNNs may not converge to a stable solution, hence,

perform poorly. In neural network ensembles, it is argued that "many could be better than all", meaning that using a subset of bagged DNNs that well fit the data could be better than using all bagged DNNs [49, 18]. Therefore, we adopt the scoring system to select the optimal subset of DNNs in the bagging procedure, following Mi et al. [18]. DNN with bagging has been implemented in the R package "deepTL" (available at `https://github.com/SkadiEye/deepTL`). According to Mi et al. [18], for all the reported numerical analysis in this paper, we set bagging size to 100, batch size to 50, the number of hidden layers to 4 with $50, 40, 30, 20$ hidden nodes at each layer subsequently, penalty weight λ to $1E-4$, and reclified linear units as the activation function.

RF and SVM

RF is implemented via R package "randomForest". Vanilla-RF importance and its standard error is generated from the "randomForest" function with 1000 trees. SVM is implemented through R package "e1071" with Radial kernels used. The hyper-parameters in SVMs are searched via 5-fold cross-validation. RFE-SVM was implemented with the function "rfe" in R package "caret".

The Simulation Study and Real Data Applications

In the simulation studies, PermFIT is performed by randomly splitting the samples into training (80%) and validation (20%) sets, and the importance score is estimated via (7) and (8). In real applications, PermFIT is conducted with 5-fold cross-fitting through (9) and (10). To eliminate the impact from the randomness of cross-fitting and other random factors in model fitting, we repeat each method 100 times and report the mean and standard deviation of MSPE, Pearson correlation, AUC or accuracy, and the median of the importance scores and p-values. Features presented in figures are ordered by hierarchical clustering, which is implemented in "hclust" function in R package "stats", where the dissimilarity is set to one minus the Pearson correlation.

For the TCGA kidney cancer application, RPPAs at gene level are analyzed. We first remove the proteins that are not common across all three TCGA datasets (KIRC, KIRP and KICH). In addition, we remove the proteins with perfect multicollinearity, after which 118 are kept for further analysis.

For the HITChip Atlas data, the BMI level was originally grouped into six groups: underweight, lean, overweight, obese, severeobese and morbidobese, which we transform into numerical levels from 1 to 6 in our analysis. Total 900 subjects are left for the analysis after subjects with missing BMI are excluded. Missing information on nationality is grouped into a new group named "Unknown". Missing values in the microbiome data are simply imputed with the median values across all samples. The analysis on the microbiome data is based on the compositional values but we remove the cell proportion from the last group due to the sum to 1 constraint on the compositional values, after which a log-transformation is applied to the remaining compositions.

Data Availability

The TCGA datasets are available at the LinkedOmics website (`http://linkedomics.org`), among which three studies, KIRC, KICH, and KIRP, are used. The HITChip Atlas data is available in R package "microbiome". We provide final datasets used in our analysis.

Code Availability

PermFIT is implemented in our R package "deepTL" (`https://github.com/SkadiEye/deepTL`). We also provide source code for replicating the simulation studies and real data applications.

Acknowledgement

The work was partially supported by the National Institute of Health Grants R01AI143886, R01CA219896, and CCSG P30 CA013696.

Competing Interests

The authors declare that they have no competing financial interests.

References

[1] Cancer Genome Atlas Research N et al. The cancer genome atlas pan-cancer analysis project. Nature Genetics, 45(10):1113–1120, 2013.

[2] D Hanahan and RA Weinberg. Hallmarks of cancer: the next generation. Cell, 114(5):646–674, 2011.

[3] M Wilhelm, J Schlegl, H Hahne, et al. Mass-spectrometry-based draft of the human proteome. Nature, 509(7502):582–587, 2014.

[4] Erik Gawehn, Jan A Hiss, and Gisbert Schneider. Deep learning in drug discovery. Molecular Informatics, 35(1):3–14, 2016.

[5] Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. Machine learning for medical imaging. Radiographics, 37(2):505–515, 2017.

[6] Aihua Zhang, Hui Sun, Guangli Yan, Ping Wang, Ying Han, and Xijun Wang. Metabolomics in diagnosis and biomarker discovery of colorectal cancer. Cancer Letters, 345(1):17–20, 2014.

[7] Ashley J Vargas and Curtis C Harris. Biomarker development in the precision medicine era: lung cancer as a case study. Nature Reviews Cancer, 16(8):525, 2016.

[8] Mark Craven and Jude W Shavlik. Extracting tree-structured representations of trained networks. In Advances in Neural Information Processing Systems, pages 24–30, 1996.

[9] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[10] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10):1340–1347, 2010.

[11] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(3):551–577, 2018.

[12] Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M Blei. The holdout randomization test: Principled and easy black box feature selection. arXiv preprint arXiv:1811.00645, 2018.

[13] Yang Lu, Yingying Fan, Jinchi Lv, and William Stafford Noble. Deeppink: reproducible feature selection in deep neural networks. In Advances in Neural Information Processing Systems, pages 8676–8686, 2018.

[14] Christoph Molnar. Interpretable Machine Learning. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[15] Enguerrand Horel and Kay Giesecke. Computationally efficient feature significance and importance for machine learning models. arXiv preprint arXiv:1905.09849, 2019.

[16] Evgeny Putin, Polina Mamoshina, Alexander Aliper, Mikhail Korzinkin, Alexey Moskalev, Alexey Kolosov, Alexander Ostrovskiy, Charles Cantor, Jan Vijg, and Alex Zhavoronkov. Deep biomarkers of human aging: application of deep neural networks to biomarker development. Aging (Albany NY), 8(5):1021–1033, 2016.

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436–444, 2015.

[18] Xinlei Mi, Fei Zou, and Ruoqing Zhu. Bagging and deep learning in optimal individualized treatment rules. Biometrics, 75(2):674–684, 2019.

[19] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. Neural Processing Letters, 9(3):293–300, 1999.

[20] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3):389–422, 2002.

[21] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary Oncology, 19(1A):68–77, 2015.

[22] Irakli Dzneladze. Pan-Cancer Study of INPP4B Reveals its Unexpected Oncogene-Like Role and Prognostic Significance. PhD thesis, 2017.

[23] Ying Sun, Huan Ding, Xinguang Liu, Xiaoqing Li, and Li Li. Inpp4b overexpression enhances the antitumor efficacy of parp inhibitor ag014699 in mda-mb-231 triple-negative breast cancer cells. Tumor Biology, 35(5):4469–4477, 2014.

[24] Iawen Hsu, Chiuan-Ren Yeh, Spencer Slavin, Hiroshi Miyamoto, George J Netto, Mesut Muyan, Xue-Ru Wu, Edward M Messing, Elizabeth A Guancial, and Shuyuan Yeh. Estrogen receptor alpha prevents bladder cancer development via inpp4b inhibited akt pathway in vitro and in vivo. Oncotarget, 5(17): 7917–7935, 2014.

[25] Irakli Dzneladze, John F Woolley, Carla Rossell, Youqi Han, Ayesha Rashid, Michael Jain, Jüri Reimand, Mark D Minden, and Leonardo Salmena. Subid, a non-median dichotomization tool for heterogeneous populations, reveals the pan-cancer significance of inpp4b and its regulation by evi1 in aml. PloS One, 13(2):e0191510, 2018.

[26] Allison A Eddy. Plasminogen activator inhibitor-1 and the kidney. American Journal of Physiology-Renal Physiology, 283(2):F209–F220, 2002.

[27] Sylwia Małgorzewicz, Ewa Skrzypczak-Jankun, and Jerzy Jankun. Plasminogen activator inhibitor-1 in kidney pathology. International Journal of Molecular Medicine, 31(3):503–510, 2013.

[28] Rainer Hofmann, Antonie Lehmer, Rudolf Hartung, Cordula Robrecth, Markus Buresch, and Franziska Grothe. Prognostic value of urokinase plasminogen activator and plasminogen activator inhibitor-1 in renal cell cancer. The Journal of Urology, 155(3):858–862, 1996.

[29] Robert H Weiss, Alexander D Borowsky, David Seligson, Pei-Yin Lin, Lisa Dillard-Telm, Arie S Belldegrun, Robert A Figlin, and Allan D Pantuck. p21 is a prognostic marker for renal cell carcinoma: implications for novel therapeutic approaches. The Journal of Urology, 177(1):63–69, 2007.

[30] Hiromi Inoue, Sung Hee Hwang, Aaron T Wecksler, Bruce D Hammock, and Robert H Weiss. Sorafenib attenuates p21 in kidney cancer cells and augments cell death in combination with dna-damaging chemotherapy. Cancer Biology & Therapy, 12(9):827–836, 2011.

[31] Mohd Saif Zaman, Varahram Shahryari, Guoren Deng, Sobha Thamminana, Sharonjot Saini, Shahana Majid, Inik Chang, Hiroshi Hirata, Koji Ueno, Soichiro Yamamura, et al. Up-regulation of microrna-21 correlates with lower kidney cancer survival. PloS One, 7(2):e31060, 2012.

[32] Lee Campbell, Bharat Jasani, David FR Griffiths, and Mark Gumbleton. Phospho-4e-bp1 and eif4e over-expression synergistically drives disease progression in clinically confined clear cell renal cell carcinoma. American Journal of Cancer Research, 5(9):2838–2848, 2015.

[33] LZ Akhmadishina, IR Giliazova, LR Kutlyeva, GF Korytina, OV Kochetova, MF Urmantsev, SM Izmaĭlova, AA Izmaĭlov, GB Kunsbaeva, AA Zagidullin, et al. Dna repair xrcc1, xpd genes polymorphism as associated with the development of bladder cancer and renal cell carcinoma. Genetika, 50 (4):481–490, 2014.

[34] Meera Srivastava, Lukas Bubendorf, Lisa Nolan, Mirta Glasman, Ximena Leighton, Georgina Miller, Wilfred Fehrle, Mark Raffeld, Ofer Eidelman, Olli P Kallioniemi, et al. Anx7 as a bio-marker in prostate and breast cancer progression. Disease Markers, 17(2):115–120, 2001.

[35] Meera Srivastava, Yelizaveta Torosyan, Mark Raffeld, Ofer Eidelman, Harvey B Pollard, and Lukas Bubendorf. Anxa7 expression represents hormone-relevant tumor suppression in different cancers. International Journal of Cancer, 121(12):2628–2636, 2007.

[36] Andrew B Smitherman, James L Mohler, Susan J Maygarden, and David K Ornstein. Expression of annexin i, ii and vii proteins in androgen stimulated and recurrent prostate cancer. The Journal of Urology, 171(2):916–920, 2004.

[37] Meera Srivastava, Lukas Bubendorf, Mark Raffeld, Christoph Bucher, Jochen Torhorst, Guido Sauter, Cara Olsen, Olli P Kallioniemi, Ofer Eidelman, and Harvey B Pollard. Prognostic impact of anx7-gtpase in metastatic and her2-negative breast cancer patients. Clinical Cancer Research, 10(7):2344–2350, 2004.

[38] Daniel Schramek, Ataman Sendoel, Jeremy P Segal, Slobodan Beronja, Evan Heller, Daniel Oristian, Boris Reva, and Elaine Fuchs. Direct in vivo rnai screen unveils myosin iia as a tumor suppressor of squamous cell carcinomas. Science, 343(6168):309–313, 2014.

[39] Astrid De Boeck, Patrick Pauwels, Karen Hensen, Jean-Luc Rummens, Wendy Westbroek, An Hendrix, Dawn Maynard, Hannelore Denys, Kathleen Lambein, Geert Braems, et al. Bone marrow-derived mesenchymal stem cells promote colorectal cancer progression through paracrine neuregulin 1/her3 signalling. Gut, 62(4):550–560, 2013.

[40] Huai-En Huang, Suet-Feung Chin, Christophe Ginestier, Valérie-Jeanne Bardou, José Adélaïde, N Gopalakrishna Iyer, Maria J Garcia, Jessica C Pole, Grace M Callagy, Stephen M Hewitt, et al. A recurrent chromosome breakpoint in breast cancer at the nrg1/neuregulin 1/heregulin gene. Cancer Research, 64(19):6840–6844, 2004.

[41] Montserrat Sanchez-Cespedes, Paola Parrella, Manel Esteller, Shuji Nomoto, Barry Trink, James M Engles, William H Westra, James G Herman, and David Sidransky. Inactivation of lkb1/stk11 is a common event in adenocarcinomas of the lung. Cancer Research, 62(13):3659–3662, 2002.

[42] Leo Lahti, Jarkko Salojärvi, Anne Salonen, Marten Scheffer, and Willem M De Vos. Tipping elements in the human intestinal ecosystem. Nature Communications, 5:4344, 2014.

[43] Leo Lahti and Sudarshan Shetty. Microbiome r package, 2012-2019.

[44] MILAN Marounek, KATERINA Fliegrova, and STANISLAV Bartos. Metabolism and some characteristics of ruminal strains of megasphaera elsdenii. Applied Environmental Microbiology, 55(6):1570–1573, 1989.

[45] Alessandro Federico, Marcello Dallio, Salvatore Tolone, Antonietta Gerarda Gravina, Vania Patrone, Marco Romano, Concetta Tuccillo, Anna Licia Mozzillo, Vincenzo Amoroso, Gabriella Misso, et al. Gastrointestinal hormones, intestinal microbiota and metabolic homeostasis in obese patients: effect of bariatric surgery. In Vivo, 30(3):321–330, 2016.

[46] Bradley J Gardiner, TM Korman, and Ralph K Junckerstorff. Eggerthella lenta bacteremia complicated by spondylodiscitis, psoas abscess, and meningitis. Journal of Clinical Microbiology, 52(4):1278–1280, 2014.

[47] Richard H Byrd, Gillian M Chin, Jorge Nocedal, and Yuchen Wu. Sample size selection in optimization methods for machine learning. Mathematical Programming, 134(1):127–155, 2012.

[48] Leo Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

[49] Z. H. Zhou, J. X. Wu, and W. Tang. Ensembling neural networks: Many could be better than all. Artificial Intelligence, 137(1-2):239–263, 2002. ISSN 0004-3702. doi: 10.1016/s0004-3702(02)00190-x. URL <GotoISI>://WOS:000176015900007.