scientific reports

Check for updates

OPEN

DeepPIG: deep neural network architecture with pairwise connected layers and stochastic gates using knockoff frameworks for feature selection

Euiyoung Oh[1] & Hyunju Lee[1,2]✉

Selecting relevant feature subsets is essential for machine learning applications. Among the feature selection techniques, the knockoff filter procedure proposes a unique framework that minimizes false discovery rates (FDR). However, employing a deep neural network architecture for a knockoff filter framework requires higher detection power. Using the knockoff filter framework, we present a Deep neural network with PaIrwise connected layers integrated with stochastic Gates (DeepPIG) for the feature selection model. DeepPIG exhibited better detection power in synthetic data than the baseline and recent models such as Deep feature selection using Paired-Input Nonlinear Knockoffs (DeepPINK), Stochastic Gates (STG), and SHapley Additive exPlanations (SHAP) while not violating the preselected FDR level, especially when the signal of the features were weak. The selected features determined by DeepPIG demonstrated superior classification performance compared with the baseline model in real-world data analyses, including the prediction of certain cancer prognosis and classification tasks using microbiome and single-cell datasets. In conclusion, DeepPIG is a robust feature selection approach even when the signals of features are weak. Source code is available at https://github.com/DMCB-GIST/DeepPIG.

Since the era of big data, revolutionary improvements have been made in various fields. A deep neural network (DNN) is a plausible approach for treating complex data. While DNNs offer remarkable predictive abilities in various tasks, their "black box" nature was most concerning to many experts who needed to understand data features used to make such decisions[1,2]. Furthermore, most datasets usually contain features irrelevant to the responses of interest, leading to suboptimal training or overfitting[3,4]. In this context, identifying the crucial features contributing to a specific response and reducing the feature dimensions are essential[5].

Various feature selection and importance scoring methods have been proposed for statistics and machine learning[6]. Feature selection methods should ideally control the rate of selecting irrelevant features while maintaining high power to identify relevant features. Traditional approaches, including the Benjamini and Hochberg procedures[7–9], use p-values that reflect feature importance. While these methods are effective for simple models, they face challenges with complex models such as DNNs. In such cases, the generation of meaningful p-values that reflect feature importance becomes ambiguous[10]. Moreover, high-dimensional data incur high costs for computing the p-values.

The model-X knockoff framework for feature selection was proposed to bypass the usage of p-values without violating the false discovery rate (FDR) above a preselected level[11]. The knockoff framework starts by generating knockoff variables that mimic the arbitrary dependence structure among the original features without looking at the responses. Knockoff variables have been used as controls in feature selection by comparing the importance of the original features and their knockoff counterparts.

Recently, feature selection approaches using modified layer architectures from vanilla neural networks have been proposed, such as stochastic gates (STG)[12]. Although they achieved a high detection power in many applications, they often did not consider FDR control explicitly. In addition, most procedures for identifying significant

[1]Gwangju Institute of Science and Technology, School of Electrical Engineering and Computer Science, Gwangju 61005, South Korea. [2]Gwangju Institute of Science and Technology, Artificial Intelligence Graduate School, Gwangju 61005, South Korea. ✉email: hyunjulee@gist.ac.kr

features depend on empirical thresholds, which make them less deterministic. Meanwhile, a DNN architecture suitable for knockoff frameworks, such as DeepPINK, was proposed[13]. DeepPINK introduces a pairwise connected filter layer for the original and knockoff variables. However, it often fails to select a single feature when the signal of important features is dim.

In this study, we designed a novel architecture of DNN and a unique training scheme called pairwise connected layers and stochastic gates (DeepPIG). We combined the core layer architecture of DeepPINK and STG to achieve higher detection power while preventing FDR violations. Furthermore, we developed distinctive training algorithms to compute feature importance using the original and corresponding knockoff variables. From the experimental results, we observed enhanced feature selection performance compared with baseline models using synthetic datasets. DeepPIG exhibited good sensitivity by finding significant features with higher power, especially when the signals of the true features were weak, while ensuring that the FDR was not violated. We also conducted the real data analysis, including cancer survival prediction tasks and the previously used microbiome and single-cell datasets in the baseline model study. The features selected by DeepPIG exhibited better classification performance and robustness than that of DeepPINK. Finally, we report the identified cancer prognostic genes, frequently identified as significant genes for classifying long-term survivors of kidney, liver, and pancreatic cancers. DeepPIG selected several prognostic genes at higher frequencies than the baseline model, highlighting its robustness. Taken together, DeepPIG provides robust feature selection by enhancing selection power and maintaining strict FDR control, making it applicable to various biological datasets.

Methods

Knockoff framework

The knockoff filter procedure was introduced as a variable selection method that controls the FDR[11]. The knockoff filter method is well known for providing accurate FDR control while bypassing p-values. The knockoff procedure has two main parts: constructing knockoff variables that imitate the original variables and defining the knockoff statistics that can be taken as feature importance scores. When generating knockoff variables, the responses of interest must not be associated. Formally, knockoff variables for a set of original random variables $\mathbf{x} = (x_1, ...x_d)^T$ are defined as a new set of random variables $\widetilde{\mathrm{x}} = (\widetilde{x}_1, ...\widetilde{x}_d)^T$ that satisfy the following properties:

(1) For any subset $S \subset \{1, ...p\}$, $(x^T, \widetilde{x}^T)_{\mathrm{swap}(S)} \overset{d}{=} (x^T, \widetilde{x}^T)$, where $(x^T, \widetilde{x}^T)_{\mathrm{swap}(S)}$ is obtained by swapping the components x_j and $\widetilde{x}_j$ in $(x^T, \widetilde{x}^T)$ for each $j \in S$ and $\overset{d}{=}$ denotes equal in distribution;
(2) $\widetilde{x} \perp y \mid x$

Among the methods for constructing knockoff variables, one promising approach is to use DNN-based models such as DeepLINK[14,15]. DeepLINK takes advantage of an autoencoder with flexible nonlinear factor modeling power. Because the feature vectors generated from the autoencoder are nonlinear, one can generate knockoff variables without assuming a joint distribution of x, such as Gaussian.

Next, knockoff variables were used as controls for the original variables; therefore, original variables with a significantly stronger relationship with the response than their corresponding knockoffs were considered important features. For each feature index $j = 1, ..., d$, we defined K_j as the knockoff statistic to measure the importance of the j-th original feature. A large positive value of K_j provides evidence that the jth original feature is important, whereas small magnitudes around zero of K_j are expected to be null features. Formally, knockoff statistics K_j is a function of the augmented data matrix $[\mathbf{X}, \widetilde{\mathrm{X}}]$ and the response vector y with a function k_j which satisfies the "sign-flip" property:

$$k_j([\mathbf{X}, \widetilde{\mathrm{X}}]_{\mathrm{swap}(S)}, \mathbf{y}) = \begin{cases} k_j([\mathbf{X}, \widetilde{\mathrm{X}}], \mathbf{y}), j \notin S \\ -k_j([\mathbf{X}, \widetilde{\mathrm{X}}], \mathbf{y}), j \in S \end{cases} \tag{1}$$

where S denotes any subset of $\{1, ..., d\}$. A threshold that does not violate the target FDR is required to select significant features using the constructed knockoff statistics. The set of important features is selected as $\widehat{S} = \{j : K_j \geq t\}$ with $t = T$ or $t = T_+$, where T is the knockoff threshold, and T_+ is the knockoff+ threshold. For target FDR level q, the knockoff thresholds are defined as follows:

$$T = \min\left\{t > 0 : \frac{\left|\{j : K_j \leq -t\}\right|}{\max\left\{\left|j : K_j \geq t\right|, 1\right\}} \leq q\right\} \tag{2}$$

$$T_+ = \min\left\{t > 0 : \frac{1 + \left|\{j : K_j \leq -t\}\right|}{\max\left\{\left|j : K_j \geq t\right|, 1\right\}} \leq q\right\} \tag{3}$$

FDR control was achieved if knockoff statistics of the null features were symmetrically distributed.

Proposed model

Here, we designed a novel DNN architecture DeepPIG for constructing knockoff statistics to improve the detection power and maintain the FDR control property of knockoff filter methods. We integrated the architectures of STG and DeepPINK[12,13]. The original and knockoff variables were combined and fed into the feature selection unit, followed by the hidden layers (Fig 1a). The final output layer produced the response variable y, which is the target output of the network. A detailed view of the feature selection unit and the training strategy are

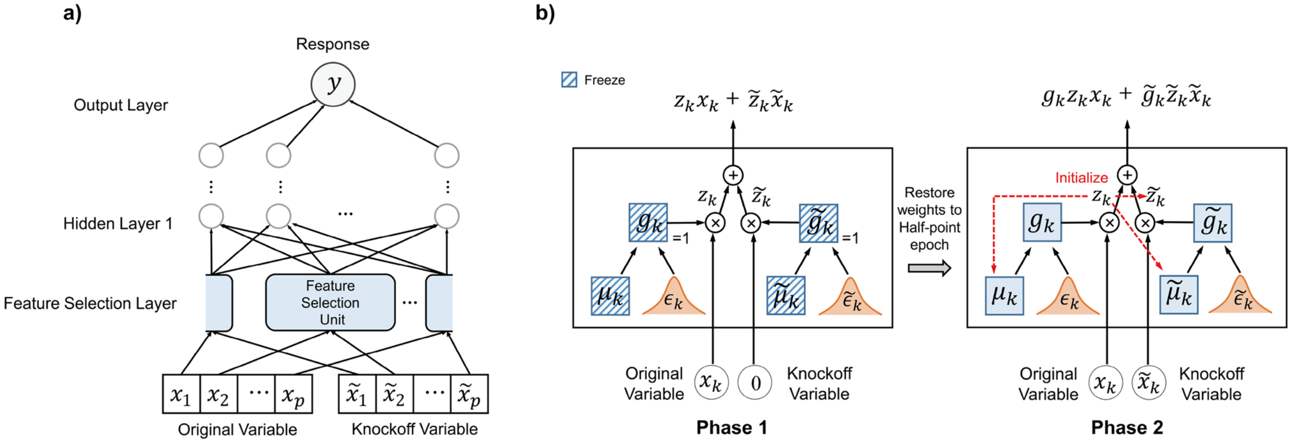

Figure 1. The architecture of DeepPIG and training strategy. (**a**) DeepPIG utilized the feature selection layer incorporating original and knockoff variables to select important features related to responses. (**b**) Detailed view of feature selection unit and their training scheme.

illustrated in Fig. 1b. In the feature selection unit, the stochastic gates are attached to each variable and utilized in a linear combination. STG was employed to utilize the $\ell 0$-norm of features or determine the number of selected features during DNN training. Since exact $\ell 0$ regularization can be computationally expensive and intractable for high dimensions, the stochastic gate component was designed to relax the Bernoulli distribution for the $\ell 0$ -norm with a continuous probabilistic distribution. Stochastic gates were attached to each input feature, where the trainable parameter, μ_j, and the random noise, ϵ_j, regulate the probability of the jth gate being active; these are called relaxed Bernoulli variables and are given as follows:

$$g_j = \max(0, \min(0, \mu_j + \epsilon_j)), \epsilon_j \sim N(0, \sigma^2) \tag{4}$$

where N denotes the normal distribution with fixed variance σ. The relaxed Bernoulli variables were clipped, mean-shifted, and random Gaussian. Given a loss L, the stochastic gate model is trained by minimizing the empirical risk:

$$\min_{\theta,\mu} \widehat{\mathbb{E}}_{X,Y} \mathbb{E}_G [L(f_\theta(\mathbf{X} \odot \mathbf{G}, Y) + \lambda \|\mathbf{G}\|_0)], \tag{5}$$

where f_θ is a model parameterized by θ, $\mathbf{G}$ is a random vector with D independent variables g_j for $j \in [D]$, Y is a response vector and $\odot$ denotes element-wise multiplication. STG considers features to be important if their gate probability values are high, such as 1. Although STG achieved noteworthy performance in finding important features in various experiments, it did not explicitly consider FDR control, which led to the selection of too many features and failure to control FDR in several experimental settings.

Next, we paired the original knockoff variable that successfully passed through the stochastic gate with its corresponding knockoff variable by the plug-in filter layer. The output of this layer is a linear combination of the weighted input variables:

$$\text{output}_{filter} = g_k z_k x_k + \tilde{g}_k \tilde{z}_k \tilde{x}_k \tag{6}$$

Through this design, the filter weights connected to the original and knockoff features compete with each other during training. The filter weights corresponding to the original features were much larger if the original features were significant for the response vectors, thereby providing evidence for the selection of important features.

To train DeepPIG, we applied two strategies: (1) a pre-training effect by masking knockoff variables into null vectors in the early stage and (2) a training-stopping criterion using the paired t-test results of the knockoff statistics K. See Algorithm 1 for the pseudocode of the training scheme. In the first phase, we only fed the original features to the model, replacing the knockoff variables with vectors in which all the elements were set to zero. In addition, gating probabilities μ and $\tilde{\mu}$ were frozen with the open state (Algorithm 1 lines 2–5). When the validation loss stabilized, the model weights were restored to a point corresponding to half of the epoch of the stopping point, and the second phase began (Algorithm 1 line 7). For example, if the validation loss stabilized at epoch 10, all model weights were restored to epoch 5. After resetting, the numeric values of the filter weights z, connected to the original variables, were copied to the filter weights $\tilde{z}$ which were connected to knockoff variables (Algorithm 1 line 8). Simultaneously, the gating probabilities of both variables (μ and $\tilde{\mu}$) were set to 0–1 scaled absolute values of the filter weights of the original variables z. This procedure allows the model to roughly identify probable features.

Next, the training was continued with both the original and knockoff variables to drop insignificant features until the model encountered the stopping criterion. The stopping criterion considers the knockoff statistics, which we have revised as follows:

$$K_j = Z_j - \widetilde{Z}_j, j = 1, ..., d, \tag{7}$$

where $Z_j = \mu_j(z_j w_j)^2$, $\widetilde{Z}_j = \widetilde{\mu}_j(\widetilde{z}_j w_j)^2$ and $w \in R^d$ are products of fully connected layer weights in the reshaped dimension. Because the importances of Z_j and $\widetilde{Z}_j$ are paired and assuming that the distribution of Z_j is greater than that of its counterpart, we applied a paired *t*-test between Z_j and $\widetilde{Z}_j$, where the alternative hypothesis is that the mean of the original importance Z_j is greater than that of the knockoff importance $\widetilde{Z}_j$. Outliers within two standard deviations of the mean were excluded from this test (Algorithm 1 lines 14–17). This stopping strategy was intended for FDR control because it depends on the assumption that the original and knockoff importance scores of the null features are symmetrically distributed. As the training progressed, the knockoff importance $\widetilde{Z}_j$ increased, and the difference between Z_j decreased. The training was stopped when 1) the p-value of the paired t-test was no longer significant ($p > 0.05$) and 2) the validation loss stabilized (Algorithm 1 lines 18–20). Finally, feature selection was conducted with a knockoff filter procedure using knockoff statistics.

Input: Original data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, knockoff data matrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times d}$, response $y \in \mathbb{R}^n$, model f_θ with parameters $\Theta = \{W_{FC}, z, \widetilde{z}, \mu, \widetilde{\mu}\}$, where fully connected layer weights W_{FC}, filter layer weights $z, \widetilde{z}$, and gate probabilities $\mu, \widetilde{\mu}$ corresponding to original and knockoff variables, respectively. Learning rate γ, and parameter history B

Output: Trained model f_Θ

```
1:  Initialize the model parameter $\Theta$.
2:  while validation loss decreasing do
3:      Compute the loss $\widehat{L}(f_\Theta([\mathbf{X},\mathbf{0}]), y)$
4:      Update $z \Leftarrow z - \gamma\nabla_\theta\widehat{L}$ and $W_{FC} \Leftarrow W_{FC} - \gamma\nabla_\theta\widehat{L}$
5:      Append $\Theta$ to $B$
6:  end while
7:  Returning $\Theta \Leftarrow B_{half-point}$
8:  $\widetilde{z} \Leftarrow z$; and $\mu = \widetilde{\mu} \Leftarrow$ min-max$(|z|)$
9:  while validation loss decreasing do
10:     Sample $\varepsilon, \widetilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$
11:     Compute the gate probability $\mathbf{s} = \max(0, \min(1, \mu + \varepsilon))$ and $\widetilde{\mathbf{s}} = \max(0, \min(1, \widetilde{\mu} + \widetilde{\varepsilon}))$
12:     Compute the loss $\widehat{L}(f_\Theta([\mathbf{X} \odot \mathbf{s}, \widetilde{\mathbf{X}} \odot \widetilde{\mathbf{x}}]), y)$
13:     Update $\Theta \Leftarrow \Theta - \gamma\nabla_\theta\widehat{L}$
14:     $Z = \mu(z \odot W_{FC})^2$ and $\widetilde{Z} = \widetilde{\mu}(\widetilde{z} \odot W_{FC})^2$
15:     $Z_{non-outlier} = \{Z \mid Z \leq \text{mean}(Z) + 2 \times std(Z)\}$
16:     $\widetilde{Z}_{non-outlier} = \left\{\widetilde{Z} \mid \widetilde{Z} \text{ paired with } Z_{non-outlier}\right\}$
17:     $p$ =p-value of paired t-test$(Z_{non-outlier}, \widetilde{Z}_{non-outlier})$
18:     if not validation loss decreasing and $p > 0.05$ then
19:         break
20:     end if
21: end while
```

Algorithm 1. Pseudocode for the DeepPIG training scheme

Results

Simulation studies

Synthetic data

Mirroring the simulation studies by DeepLINK, we designed the simulation experiment settings as follows: linear factor model and logistic factor model (Eqs. 8, 9).

$$x_i = \Lambda f_i + \epsilon_i \tag{8}$$

$$x_{ij} = \frac{c_j}{1 + \exp([1, \mathbf{f}_i^T]\boldsymbol{\lambda}_j)} + \epsilon_{ij}, j = 1, ..., d \tag{9}$$

Here, $f_i = (f_i^1, f_i^2, f_i^3)^T$ is the latent factor vector, Λ and $\boldsymbol{\lambda}_j$ are the factor loading parameters of the desirable dimensions, c_j's are constants, and ϵ denotes random noise. All the parameters were drawn independently from the standard normal distribution $N(0, 1)$. The response vector $y = (y_1, ..., y_n)^T$ is assumed to depend on x_i via the following linear and nonlinear link functions (Eqs. 10, 11).

$$y_{\text{linear}} = \mathbf{x}^T\boldsymbol{\beta} \tag{10}$$

$$y_{\text{nonlinear}} = \sin(\mathbf{x}^T\boldsymbol{\beta})\exp(\mathbf{x}^T\boldsymbol{\beta}) \tag{11}$$

where the coefficient vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_d)^T$. The locations of the true features were randomly selected, and the corresponding β_j was set to the amplitude of nonzero A, either positive or negative, with equal probability. The remaining features were considered to be null, and their corresponding β_j was set to zero.

Here, we set the sample size n to 1000, feature dimension d to 500, and true feature size s to 10. The values of the amplitude A varied from three to 25. For all settings, we conducted experiments 100 times with the target FDR q as 0.2.

Simulation results

The model's feature selection performances on the synthetic datasets were determined using metrics such as power and FDR. Power is defined as the expectation of the true discovery proportion (TDP).

$$\text{Power} := \mathbb{E}[\text{TDP}] \text{ with TDP} := \frac{|S \bigcap S_0|}{|S_0|} \tag{12}$$

where S denotes the subset of selected features and S_0 denotes the subset of true features. Power can be interpreted as recall as well. In contrast, FDR is formally defined as the expectation of the false discovery proportion (FDP).

$$\text{FDR} := \mathbb{E}[\text{FDP}] \text{ with FDP} := \frac{|S \bigcap S_0^c|}{\max\{|S|, 1\}} \tag{13}$$

FDR can be interpreted as 1-precision.

DeepPIG showed better detection power than DeepLINK while controlling the FDR to less than 0.2 in various settings, as illustrated in Fig. 2. When the link function was linear, the signal of the true features increased as the amplitude increased. By contrast, the signal and amplitude are no longer monotonic to the nonlinear link function, as shown in Fig. 2b, d. Notably, DeepPIG showed higher detection power even when the signal amplitudes were weak. Statistically, we conducted the paired t-test to compare the obtained powers by DeepPIG and DeepPINK when the amplitudes were three to eight combined. The p-values were 1.61 e−08, 1.12 e−14, 5.56 e−08 and 0.695 for the settings described in Fig. 2a–d. For the last case, the p-value was 1.20e-08 when amplitudes were 20 to 25 combined. The full scales of the results are depicted in Fig. S1. Further, DeepPIG exhibited higher F1 scores than baseline models when signals were weak, as illustrated in Fig. S2. Based on these observations, it is expected that DeepPIG will be effective in identifying inconspicuous features. STG failed to control the FDR when the response vector was generated using a nonlinear link function, and the amplitudes were large.

We compared the performance of our model with other methods such as SHapley Additive exPlanations (SHAP)[16], a representative method in the explainable AI area. We utilized the SHAP method on basic DNN models. Furthermore, we applied linear regression and recursive feature elimination (RFE) on Lasso regression for conventional feature selection approaches. We selected the top 10 features based on their shapley values or coefficients, respectively. We observed that these methods were effective in scenarios where the significant features were conspicuous but failed to control FDR when the signals were weak (Fig. 2 and Fig. S3). When the link function was linear, it was difficult to identify important features when the amplitudes were low because the amplitudes and signal were monotonic. SHAP, linear regression, and RFE showed relatively high FDRs when the amplitudes of the features were low. Conversely, when the link function was nonlinear, these methods also exhibited high FDRs for relatively high amplitudes, as the relationship between the amplitudes and signal was not monotonic.

We conducted additional experiments for hyperparameter analysis for the simulation study results. We experimented with the restore epoch that was set to 30% and 90% in addition to 50% to demonstrate the effect of weight transfer timing. The later the weight transfer occurred, DeepPIG tended to select more features. DeepPIG failed to control FDR when the restore epoch was set too late, meaning the knockoff variables were not trained enough. We observed that the optimal performance was achieved when the restore epoch was set to 50% (Fig. S4a).

Next, various ranges of regularization coefficients were assessed to determine their effects on the model parameter. A higher regularization coefficient typically resulted in a controlled FDR with decreased power, leading to selection of fewer features (Fig. S4b). To verify the robustness of our findings, we conducted the experiments again using a synthetic dataset with a different system that had different random seeds and achieved equivalent results (Fig. S5).

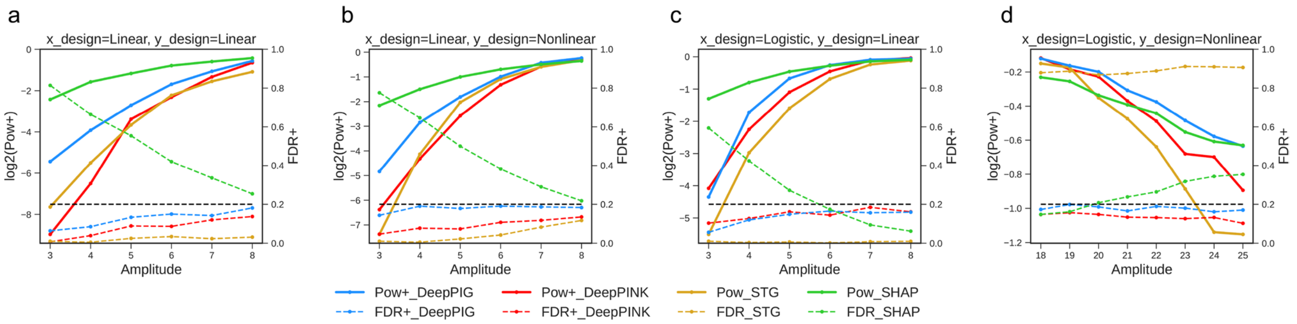

Figure 2. Simulation study results. (**a–d**) DeepPIG, DeepPINK, STG, and SHAP were applied to the synthetic dataset for feature selection. Empirical power and FDR of DeepPIG and DeepPINK were obtained using the knockoff+ threshold. Powers are illustrated as solid lines and FDRs as dashed lines. The preselected FDR target is 0.2, as shown in black dashed lines.

Real data analysis

Transcriptomic markers of cancer prognosis

Predicting cancer prognosis is challenging in the field of cancer therapy. In this study, we used DeepPIG to identify cancer prognostic genes. Transcriptomic profiles and prognostic information of patients with kidney, liver, and pancreatic cancer were collected from The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) databases[17–20]. For each cancer type, we categorized the patients as long-term survivors (LTS) or short-term survivors (non-LTS) based on their survival duration and death event occurrences. Patients were labeled as LTS if their survival times were larger than a specific threshold regardless of their death event occurrence, whereas non-LTS were labeled if their survival times were less than the threshold and the event had occurred. The properties of each dataset are listed in Table 1.

The distance correlation was applied to all three datasets as a screening step before they were fed into the feature selection models[21]. The prediction performance is reportedly poor when these datasets are employed without screening steps, suggesting that the feature space must be reduced before training the models[15]. Using ICGC datasets for screening steps, prognostic genes were ranked by the distance correlation between the gene expression level and LTS status. TCGA datasets were filtered using screened genes and employed in feature selection models.

We applied DeepPIG and DeepPINK 100 times to select the features and compared the number of selected features. For DeepPIG, we set the minimum epoch for weight transfer to increase the sensitivity. For each repetition, we randomly split each dataset into 80% training and 20% testing. To ensure fairness, each repetition used identical knockoff variables and a training-test split for DeepPIG and DeepPINK. Further, repetitions that could not select any features were denoted as "empty repetitions" and excluded from the 100 repetitions when deriving prediction performance. After selection, their prediction abilities were measured using independent vanilla DNNs, and the area under the curve (AUC) and classification errors were determined as the performance metrics. For further analysis, we used all the screened features and the same number of randomly selected features as those selected by DeepPIG for the same repetition. We also computed the precision, recall, and F1 scores and have reported them in Table S7.

We observed that DeepPIG outperformed DeepPINK in terms of the performance metrics and the number of selected features in all three datasets, as summarized in Table 1. Notably, DeepPIG has fewer empty repetitions and greater robustness in feature selection. Despite searching for appropriate hyperparameters, such as the learning rate and regularization coefficients, DeepPINK showed empty repetitions with high chances. The top 10 most frequently selected genes and their selection ratios, i,e., the number of selection times when it was not an empty repetition, are reported in Table 2. DeepPIG showed higher selection ratios for top-ranked genes than DeepPINK, suggesting the robustness of DeepPIG during repetitions. The top 100 genes and their selection ratios are listed in Tables S1–S3.

Finally, we investigated the biological roles and associations of the top-ranked genes with cancers. COL11A1 is essential for bone development and collagen fiber assembly and acts as a prognostic marker in many solid cancers, including renal carcinoma[22–24]. Hepatocyte growth factor (HGF) is a pleiotropic factor that is crucial for tubular repair, regeneration after acute renal injury, renal development, and the maintenance of normal adult kidney structure and function[25,26]. Increased PLOD2 expression is often found in advanced tumors and is correlated with a poor prognosis in patients with hepatocellular carcinoma[27]. It was reported that the overexpression of Stanniocalcin 2 (STC2) was correlated with tumor growth, invasion, metastasis, and prognosis associated with many types of cancers, including liver cancer[28,29]. C15orf48 is highly expressed in pancreatic cancer and is significantly associated with the prognosis of pancreatic adenocarcinoma[30]. Furthermore, the role of RRAD in the occurrence of ferroptosis in pancreatic cancer has been previously reported[31].

To further investigate the significance of the selected prognostic genes, we conducted a univariate Cox proportional hazards analysis on the corresponding TCGA cohorts, including all patients whose prognostic

Tissue	Dataset[a]		# of screened features	Mean # of selected features		Empty repetitions		Mean ± SD test AUC (mean ± SD test classification error)[b]			
	Main	Screening		DeepPIG	DeepPINK	DeepPIG	DeepPINK	DeepPIG	DeepPINK	All screened features	Random features[c]
Kidney	TCGA-KIRC, 222 LTS (48), 82 nonLTS (24)	RECA-EU, 43 LTS (60) 18 nonLTS (18)	400	9.00	1.92	0 / 100	56 / 100	0.651 ± 0.087 (0.259 ± 0.044)	0.587 ± 0.087 (0.266 ± 0.030)	0.596 ± 0.076 (0.308 ± 0.046)	0.600 ± 0.089 (0.272 ± 0.032)
Liver	TCGA-LIHC, 91 LTS (36), 104 nonLTS (36)	LIRI-JP, 19 LTS (48), 17 nonLTS (12)	200	8.27	2.86	9 / 100	43 / 100	0.631 ± 0.094 (0.404 ± 0.080)	0.596 ± 0.109 (0.419 ± 0.085)	0.661 ± 0.084 (0.374 ± 0.076)	0.570 ± 0.104 (0.455 ± 0.079)
Pancreas	TCGA-PAAD, 66 LTS (18), 66 nonLTS (18)	PAAD-CA, 19 LTS (48), 43 nonLTS (12)	100	5.54	1.26	16 / 100	73 / 100	0.573 ± 0.106 (0.455 ± 0.090)	0.537 ± 0.103 (0.479 ± 0.092)	0.628 ± 0.091 (0.427 ± 0.092)	0.563 ± 0.111 (0.461 ± 0.091)

Table 1. Long-term survivor classification results for cancer datasets. [a] Dataset name, the number of long-term survivors (LTS) and short-term survivors (nonLTS), and survival time threshold in the month in parentheses. [b] Empty repetitions were excluded. [c] The same number of features as those of DeepPIG was randomly selected.

	Kidney				Liver				Pancreas			
Rank	DeepPIG		DeepPINK		DeepPIG		DeepPINK		DeepPIG		DeepPINK	
1	COL11A1	0.88	IFI44	0.36	PLOD2	0.87	PLOD2	0.33	C15orf48	0.8	C15orf48	0.48
2	HGF	0.88	HGF	0.27	STC2	0.67	ADAM9	0.25	RRAD	0.7	RRAD	0.26
3	IFI44	0.78	COL11A1	0.23	TMX1	0.6	GCLM	0.23	PSMB8	0.52	PSMB8	0.26
4	LYPD6B	0.7	NTM	0.23	ADAM9	0.49	STC2	0.23	GPBAR1	0.45	USP22	0.19
5	KCNE5	0.67	KCNE5	0.18	PARD3	0.41	PIK3IP1	0.19	MAP1LC3B	0.43	RCOR1	0.19
6	BCAT1	0.63	NKAIN4	0.14	IFI6	0.34	MRPL3	0.16	PPP1R10	0.39	FAM19A5	0.19
7	PGC	0.5	LYPD6B	0.14	MERTK	0.31	ZWINT	0.16	TGFBR3	0.27	PPP1R10	0.19
8	C16orf89	0.38	PGC	0.14	GCLM	0.3	PARD3	0.14	UGT2B15	0.19	HIST1H2AC	0.15
9	HSPB7	0.25	BCAT1	0.11	IGFBP3	0.29	GTPBP4	0.12	SH3BP4	0.19	SOX9	0.15
10	APOD	0.24	IRS4	0.11	C5	0.26	MEX3D	0.12	UGT2B17	0.18	GPBAR1	0.11

Table 2. Top 10 frequently selected prognostic genes with selection ratio. (Ratio of times selected to non-empty repetitions).

information was available. We observed that the top-ranked genes frequently exhibited significant p-values, showing their association with survival in cancer patients, as summarized in Table 3

Microbiome and single-cell datasets

We compared the performances of DeepPIG and DeepPINK using the same datasets and procedures described in the DeepLINK study[15]. The datasets used were the human microbiome dataset from a colorectal cancer (CRC) study[32,33], the human single-cell dataset from a glioblastoma study[34], and the murine single-cell dataset from a lipopolysaccharide (LPS)-stimulated transcriptomic effect study[35]. Human microbiome datasets were utilized to identify important microbial species related to colorectal cancer by classifying 184 individuals (91 patients with colorectal cancer and 93 healthy controls). The human single-cell dataset contained 632 cells (580 tumor cells and 52 surrounding peripheries) from patients with glioblastoma, and these single-cell gene expression data were employed to investigate the differential gene expression between both cell types. The murine single-cell dataset was collected to investigate the effect of LPS-stimulated nuclear factor-κB (NF- κB) on gene expression. Classification of 580 cells (202 unstimulated cells and 368 LPS-stimulated cells) based on their condition revealed significantly differentially expressed genes under both conditions.

Similar to the previous section, we applied screening steps and applied DeepPIG and DeepPINK 100 times to select the features. As shown in Table 4, DeepPIG selected more features than DeepPINK and exhibited better test errors. The frequently selected features are listed in Tables S4–S6.

The frequently selected features are similar to those in the analysis in the DeepLINK study. However, some genes were selected more frequently by DeepPIG, suggesting its sensitive detection capability. For instance, for the human microbiome dataset analysis, *Parvimonas micra* was selected 94 times out of 100 repetitions by DeepPIG, whereas it was selected 52 times by DeepPINK. Several recent studies have reported the biological effects of *P. micra* on CRC. *Parvimonas micra* promotes the development of CRC and can be considered as a predictor of poor outcomes in patients with CRC[36,37]. It also influences proliferation, wound healing, and inflammation in CRC cell lines[38]. For human single-cell dataset analysis, B2M and C1R were selected 47 and 42 times, respectively, using DeepPIG, compared to 6 and 3 times using DeepPINK. Some studies have reported that B2M has a significant relationship with the tumor-immune microenvironment and plays a critical role in tumor progression,

Kidney			Liver			Pancreas		
Gene	Hazard ratio	CoxPH p-value	Gene	Hazard ratio	CoxPH p-value	Gene	Hazard ratio	CoxPH p-value
COL11A1	1.018	**0.0001235**	PLOD2	1.027	**5.43E–06**	C15orf48	1.019	**0.000525**
HGF	1.017	**0.00024**	STC2	1.025	**0.0002739**	RRAD	1.007	**0.000857**
IFI44	1.026	**6.39E–12**	TMX1	1.067	**0.0056695**	PSMB8	1.028	**2.94E-05**
LYPD6B	1.017	0.3040778	ADAM9	1.088	**3.83E–05**	GPBAR1	0.919	**0.032268**
KCNE5	1.016	**0.0399017**	PARD3	1.054	**6.63E–05**	MAP1LC3B	1.003	0.865728
BCAT1	1.067	**3.66E–08**	IFI6	1	0.6973015	PPP1R10	0.932	**0.000291**
PGC	1.019	0.0828003	MERTK	1.011	0.214183	TGFBR3	0.994	0.847321
C16orf89	0.997	0.6961147	GCLM	1.017	**1.43E–05**	UGT2B15	1.004	0.323858
HSPB7	0.994	0.7014953	IGFBP3	1.001	0.5672436	SH3BP4	1.016	0.289861
APOD	1.016	**0.0052317**	C5	0.996	**0.0007997**	UGT2B17	0.96	0.351064

Table 3. Survival analysis of top 10 frequently selected prognostic genes. p-values under 0.05 are presented in bold.

Dataset	# of screened features	Mean # of selected features		Empty repetitions		Mean ± SD test classification error[a]			
		DeepPIG	DeepPINK	DeepPIG	DeepPINK	DeepPIG	DeepPINK	All screened features	Random features[b]
Human microbiome	30	13.69	10.66	0/100	0/100	0.275 ± 0.075	0.287 ± 0.075	0.298 ± 0.072	0.342 ± 0.083
Human single-cell	200	28.93	17.36	0/100	11/100	0.048 ± 0.026	0.053 ± 0.032	0.035 ± 0.022	0.068 ± 0.029
Murine single-cell	200	50.96	34.48	0/100	0/100	0.011 ± 0.013	0.015 ± 0.017	0.009 ± 0.012	0.038 ± 0.031

Table 4. Performance comparison with microbiome and single-cell datasets. [a] Empty repetitions were excluded. [b] The same number of features as those of DeepPIG was randomly selected.

patient prognosis, and immunotherapy of gliomas[39–41]. Furthermore, the expression level of C1R was associated with immune cell infiltration and prognosis of glioblastoma[42].

Discussion

Various explainable AI (XAI) models such as SHAP[16] are used for measuring feature importance and understanding which feature contribute significantly to the output of the neural network models. Although the knockoff models and XAI techniques both employ the weights of parameters to compute the feature importance, their main aspects are somewhat different. XAI techniques are mainly applied to interpret the results of trained models that are considered "black box" models that users can not catch how the model comes to the specific results. On the contrary, the knockoff framework is designed to select significant features among a large number of features while keeping the FDR not violated.

One advantage of the knockoff framework over conventional feature selection approaches or SHAP is its ability to determine the threshold using knockoff variables. In other methods, it is necessary to specify the number of features to be selected. We examined the performance of existing methods with the top 10 features, as we were aware that 10 significant features existed within the synthetic datasets. The knockoff framework is useful for real dataset analysis, especially when it is unclear how many features should be selected.

The motivation for designing DeepPIG is that the knockoff filter method often fails to select a single feature in real data analysis. We focused on constructing a sensitive feature-selection model for low-amplitude signal cases. Cancer prognosis prediction is an example of this, as genes related to survival are uncommon. DeepPIG exhibited better detection power when the amplitude of the features was weak, as demonstrated by synthetic and real data.

Some components of the proposed training strategy were determined empirically. For example, in selecting the specific epoch for weight transfer, the late transfer failed in FDR control. If DeepPIG was "overcooked" with the original variable only, the knockoff variables had an insufficient influence on the model and could not overcome the score differences between the original and knockoff variables. In this context, the "half point epoch" criterion was decided empirically based on the simulation study. Although an epoch earlier than half is acceptable, we recommend staying within half of the epoch where the validation loss is stabilized.

Additionally, the 0.05 p-value of the paired *t*-test for the stopping criterion was adjusted. This is because the knockoff statistics within non-outlier regions do not necessarily indicate that they are actual null features. Since a p-value of 0.05 is considered the general criterion in statistical fields, we stuck with it as a criterion for our study. We utilized paired t-tests as a "gadget" to verify how the knockoff importance scores catch up with that of the original. An alternative method for determining the time for weight transfer and testing the symmetry of the original and knockoff scores should be explored in future studies.

Conclusion

In this study, we present DeepPIG, a DNN architecture, and a training scheme for feature selection. We integrated the key structures of DeepPINK and STG to improve detection power while keeping FDR under a preselected level. Using synthetic data, we achieved a higher power, especially when the amplitudes of the features were weak. We applied DeepPIG to renal carcinoma, hepatocellular carcinoma, and pancreatic carcinoma datasets to classify patients with cancer as LTS or non-LTS. DeepPIG robustly selects several prognostic genes at high frequencies. Furthermore, we compared DeepPIG with the baseline model DeepPINK using the human microbiome, human single-cell, and murine single-cell datasets, which were employed in the baseline model study. It was observed that DeepPIG selected a greater number of features and had superior prediction capacities. In conclusion, DeepPIG is a robust feature selection approach even when the signals of features are weak.

Data availability

Simulation data, the human microbiome dataset from a colorectal cancer study, the human single-cell dataset from a glioblastoma study, the murine single-cell dataset from a lipopolysaccharide (LPS)-stimulated transcriptomic effect study, and the cancer prognosis dataset are publicly available at https://github.com/DMCB-GIST/DeepPIG.

Received: 28 March 2024; Accepted: 26 June 2024
Published online: 06 July 2024

References

1. Adadi, A. & Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018).

2. Saleem, R. *et al.* Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing* **513**, 165–180 (2022).
3. Roelofs, R. *et al.* A meta-analysis of overfitting in machine learning. *Adv. Neural Inf. Process. Syst.* **32** (2019).
4. Ying, X. An overview of overfitting and its solutions. *J. Phys. Conf. Ser.* **1168**, 022022 (IOP Publishing, 2019).
5. Jović, A. *et al.* A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 1200–1205 (2015).
6. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
7. Benjamini, Y. Discovering the false discovery rate. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72**, 405–416 (2010).
8. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995).
9. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 1165–1188 (2001).
10. Ghorbani, A. *et al.* Interpretation of neural networks is fragile. *Proc. AAAI Conf. Artif. Intell.* **33**, 3681–3688 (2019).
11. Candes, E. *et al.* Panning for gold:'Model-x' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **80**, 551–577 (2018).
12. Yamada, Y. *et al.* Feature selection using stochastic gates. In *International Conference on Machine Learning*. 10648–10659 (PMLR, 2020).
13. Lu, Y. *et al.* Deeppink: Reproducible feature selection in deep neural networks. *Adv. Neural Inf. Process. Syst.* **31** (2018).
14. Romano, Y. *et al.* Deep knockoffs. *J. Am. Stat. Assoc.* **115**, 1861–1872 (2020).
15. Zhu, Z. *et al.* Deeplink: Deep learning inference using knockoffs with applications to genomics. *Proc. Natl. Acad. Sci.* **118**, e2104683118 (2021).
16. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17. 4768–4777 (Curran Associates Inc., 2017).
17. Creighton, C. J. *et al.* Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).
18. Raphael, B. J. *et al.* Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell* **32**, 185–203.e13 (2017).
19. Wheeler, D. A. *et al.* Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327 (2017).
20. Zhang, J. *et al.* The international cancer genome consortium data portal. *Nat. Biotechnol.* **37**, 367–369 (2019).
21. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **35**, 2769–2794 (2007).
22. Li, C. *et al.* Identification of potential core genes in metastatic renal cell carcinoma using bioinformatics analysis. *Am. J. Transl. Res.* **11**, 6812 (2019).
23. Wu, Y.-H. & Chou, C.-Y. Collagen xi alpha 1 chain, a novel therapeutic target for cancer treatment. *Front. Oncol.* **12**, 925165 (2022).
24. Nallanthighal, S. *et al.* Collagen type xi alpha 1 (col11a1): A novel biomarker and a key player in cancer. *Cancers* **13**, 935 (2021).
25. Liu, Y. Hepatocyte growth factor in kidney fibrosis: Therapeutic potential and mechanisms of action. *Am. J. Physiol.-Renal Physiol.* **287**, F7–F16 (2004).
26. Liu, Y. Hepatocyte growth factor and the kidney. *Curr. Opin. Nephrol. Hypertens.* **11**, 23–30 (2002).
27. Li, K. *et al.* Dysregulation of plod2 promotes tumor metastasis and invasion in hepatocellular carcinoma. *J. Clin. Transl. Hepatol.* **11**, 1094 (2023).
28. Bu, Q. *et al.* Stc2 is a potential biomarker of hepatocellular carcinoma with its expression being upregulated in nrf1α-deficient cells, but downregulated in nrf2-deficient cells. *Int. J. Biol. Macromol.* **253**, 127575 (2023).
29. Qie, S. & Sang, N. Stanniocalcin 2 (stc2): A universal tumour biomarker and a potential therapeutical target. *J. Exp. Clin. Cancer Res.* **41**, 1–19 (2022).
30. Li, C. *et al.* The prognostic and immune significance of c15orf48 in pan-cancer and its relationship with proliferation and apoptosis of thyroid carcinoma. *Front. Immunol.* **14**, 1131870 (2023).
31. Lu, Z. *et al.* Setd8 inhibits ferroptosis in pancreatic cancer by inhibiting the expression of rrad. *Cancer Cell Int.* **23**, 1–16 (2023).
32. Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
33. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
34. Darmanis, S. *et al.* Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.* **21**, 1399–1410 (2017).
35. Lane, K. *et al.* Measuring signaling and RNA-seq in the same cell links gene expression to dynamic patterns of nf-κb activation. *Cell Syst.* **4**, 458–469.e5 (2017).
36. Chang, Y. *et al.* *Parvimonas micra* activates the RAS/ERK/c-FOS pathway by upregulating MIR-218-5p to promote colorectal cancer progression. *J. Exp. Clin. Cancer Res.* **42**, 13 (2023).
37. Zhao, L. *et al.* *Parvimonas micra* promotes colorectal tumorigenesis and is associated with prognosis of colorectal cancer patients. *Oncogene* **41**, 4200–4210 (2022).
38. Hatta, M. *et al.* *Parvimonas micra* infection enhances proliferation, wound healing, and inflammation of a colorectal cancer cell line. *Biosci. Rep.* **43** (2023).
39. Tang, F. *et al.* Impact of beta-2 microglobulin expression on the survival of glioma patients via modulating the tumor immune microenvironment. *CNS Neurosci. Ther.* **27**, 951–962 (2021).
40. Zhang, H. *et al.* B2m overexpression correlates with malignancy and immune signatures in human gliomas. *Sci. Rep.* **11**, 5045 (2021).
41. Li, D. *et al.* β2-microglobulin maintains glioblastoma stem cells and induces m2-like polarization of tumor-associated macrophages. *Cancer Res.* **82**, 3321–3334 (2022).
42. Wang, X. *et al.* C1r, ccl2, and tnfrsf1a genes in coronavirus disease-COVID-19 pathway serve as novel molecular biomarkers of GBM prognosis and immune infiltration. *Dis. Markers* **2022**, 8602068 (2022).

Acknowledgements

This work was funded by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) [No. 2019-0-00567, Development of Intelligent SW systems for uncovering genetic variation and developing personalized medicine for cancer patients with unknown molecular genetic mechanisms; No. 2019-0-01842, Artificial Intelligence Graduate School Program (GIST)] and a GIST Research Project grant funded by the GIST in 2023.

Author contributions

E.O. and H.L. conceptualized and conducted the experiments, analyzed the results, wrote and reviewed the manuscript. H.L. supervised the project and funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-66061-6.

Correspondence and requests for materials should be addressed to H.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.