Wang XF, Zhou X, Rao JH *et al.* Imputing DNA methylation by transferred learning based neural network. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 37(2): 320–329 Mar. 2022. DOI 10.1007/s11390-021-1174-6

Imputing DNA Methylation by Transferred Learning Based Neural Network

Xin-Feng Wang[1] (王新峰), Xiang Zhou[1] (周 翔), Jia-Hua Rao[1] (饶家华), Zhu-Jin Zhang[1] (张柱金), and Yue-Dong Yang[1,2,*] (杨跃东), *Member*, *CCF*

[1] *School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China*

[2] *Key Laboratory of Machine Intelligence and Advanced Computing of Ministry of Education (Sun Yat-sen University) Guangzhou 510000, China*

E-mail: {wangxf59, zhoux85, raojh6, zhangzhy58}@mail2.sysu.edu.cn; yangyd25@mail.sysu.edu.cn

Received November 23, 2020; accepted February 18, 2022.

Abstract DNA methylation is one important epigenetic type to play a vital role in many diseases including cancers. With the development of the high-throughput sequencing technology, there is much progress to disclose the relations of DNA methylation with diseases. However, the analyses of DNA methylation data are challenging due to the missing values caused by the limitations of current techniques. While many methods have been developed to impute the missing values, these methods are mostly based on the correlations between individual samples, and thus are limited for the abnormal samples in cancers. In this study, we present a novel transfer learning based neural network to impute missing DNA methylation data, namely the TDimpute-DNAmeth method. The method learns common relations between DNA methylation from pan-cancer samples, and then fine-tunes the learned relations over each specific cancer type for imputing the missing data. Tested on 16 cancer datasets, our method was shown to outperform other commonly-used methods. Further analyses indicated that DNA methylation is related to cancer survival and thus can be used as a biomarker of cancer prognosis.

Keywords neural network, transfer learning, DNA methylation, data imputation, survival analysis

1 Introduction

DNA methylation is the earliest discovered epigenetic regulation mechanism that modifies DNA sequence in a reversible way by selectively adding methyl groups to cytosine under the action of DNA methyltransferase [1]. It plays a significant role in gene expression and regulation, and participates in many cellular processes including cellular differentiation, development and tumorigenesis. The most common types of DNA methylation are 5-methylcytosine (5mC), 6-methyladenine (6mA), and 4-methylcytosine (4mC), among which 5mC is the most widely studied [2]. DNA methylation is involved in the pathogenesis of various diseases. For example, tumor cells often exhibit different DNA methylation patterns from healthy cells, and tumor-suppressor genes are likely deactivated by the hypermethylation in the promoter regions among many cancer types [3]. With the rapid improvements in sequencing technologies, there are many successful studies to disclose tumor-associated DNA methylation changes. For example, the hypermethylation of CpG islands was found to associate with low survival rates of the adrenocortical cancer (ACC) [4]. DNA methylation can be used as an independent prognostic factor to improve the prognosis [5]. Despite the huge improvements in sequencing technologies, direct use of DNA methylation was prevented since the measured data often contains a large portion of missing values [6]. It is necessary to impute the missing values in the sequenced

Regular Paper

This study was supported by the National Key Research and Development Program of China under Grant No. 2020YFB0204803, the National Natural Science Foundation of China under Grant No. 61772566, the Guangdong Key Field Research and Development Plan under Grant Nos. 2019B020228001 and 2018B010109006, the Introducing Innovative and Entrepreneurial Teams of Guangdong under Grant No. 2016ZT06D211, and the Guangzhou Science and Technology Research Plan under Grant No. 202007030010.

[*] Corresponding Author

data.

For this purpose, many approaches have been developed based on statistics or machine learning techniques. For example, KNNimpute selects the most similar samples and averages the values over these samples to fill the missing sites [7]. SVDmiss [8] and imputePCA [9], similar in principle, both calculate the covariance matrix of input data and then reconstruct the input matrix based on the eigenvectors corresponding to a few highest eigenvalues. Linear regression [10] trains a linear function to impute the missing values, while missForest [11] estimates missing values through an iterative method of random forest regression trees. However, these traditional methods do not fully utilize the available big data and show limited accuracies. With the development of deep learning techniques [12], deep neural networks (DNNs) have shown superiority in modeling complex nonlinear relationships as well as scalability and flexibility in many fields, including protein structure prediction [13–15], drug design [16–18], genome analysis [19–21], and single cell genomic analysis [22–25]. Although there are many successful applications, it remains challenging to directly implement deep learning into multi-omics analysis due to a large number of redundant features with only few samples [26]. For instance, in the Cancer Genome Atlas (TCGA) dataset, each cancer type only has hundreds of samples but tens of thousands of features [27]. Therefore, it is hard to train an accurate DNN model with millions of parameters for single cancer with the small sample size.

A promising method to solve small sample problems is transfer learning that pretrains model on a similar with big sample size and then fine-tunes the model on the target tasks. The technique has been proven to be effective in many omics data analysis tasks of cancer, such as the prediction of survival time [28], clinical outcomes [29], and common mechanisms [30, 31]. Recently, we have developed a new method (TDimpute) to impute missing RNA-seq data by combining the transfer learning technique with AutoEncoder [32]. The method was proven to effectively solve the small sample problems in the TCGA dataset but has not been proven in the imputation of other omics types.

In this study, we employ a similar architecture to impute the missing value of methylation, namely TDimpute-DNAmeth. Specifically, we first train a general imputation model suitable for all cancers based on the pan-cancer dataset, which is then transferred to the target cancer type. To the best of our knowledge, this is the first application of transfer learning to the imputation of DNA methylation data. The comprehensive tests indicate that our model outperforms other competitive approaches. More importantly, we demonstrate that the imputed data by our model significantly improves the survival predictions for multiple cancers in the independent datasets.

2 Materials and Methods

2.1 Datasets

We obtained the dataset of 33 cancer types from the Cancer Genome Atlas (TCGA) with R package TCGA-assembler [33], including DNA methylation data (JHU-USC Human Methylation 450) and clinical information. The dataset consists of 9 756 samples from 33 cancers, each with 485 577 methylation sites discretely distributed on the genome chains. The DNA methylation level is measured by the ratio of the methylated probe intensity to the total intensity including both methylated and unmethylated alleles, called β-value (ranging from 0 to 1). Even for the high-quality data generated by TCGA, there are many missing methylation sites occurring in different portions of samples (see Fig.S1 in supplementary files[①]).

Because of the extremely uneven distribution of sample sizes for 33 cancer types, we kept only cancer types with sample sizes greater than 200 and complete data of DNA methylation and clinical information, leading to 16 cancer types. Across the cancer types, we only kept 20 000 common sites with the highest variations for a balance with computational costs. As detailed in Table 1, the dataset finally contains 7 531 samples with 20 000 sites per sample, namely the TCGA dataset.

2.2 Simulated Missing Values

To assess our model under different proportions of missing data, we introduced different proportions of missing values (denoted as red "0"). The sites of missing values were selected by random, just following the previous study [10]. After the imputations, we evaluate the results through the root mean square errors of the real values and imputed values of the red "0" position.

[①] Supplementary files include Figs.S1–S8 and Table S1. https://github.com/zhoux85/TDimpute/blob/master/Imputing_DNA_methylation_by_transferred_learning_based_neural_network_Supplementary_file.pdf, Feb. 2022.

As shown in Fig.1, in the original DNA methylation matrix $\boldsymbol{D} \in \mathbb{R}^{M\times N}$ with M and N denoting the numbers of samples and DNA methylation sites respectively. N is 20 000 in all simulation experiments.

Table 1. Selected 16 Cancers and Their Sample Numbers

Cancer	Full Name	Dataset Size
BLCA	Bladder urothelial carcinoma	440
BRCA	Breast adenocarcinoma	892
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	312
COAD	Colon carcinoma	353
HNSC	Head and neck squamous cell carcinoma	580
KIRC	Kidney renal clear cell carcinoma	485
KIRP	Kidney renal papillary cell carcinoma	321
LGG	Brain lower grade glioma	534
LIHC	Liver hepatocellular carcinoma	430
LUAD	Lung adenocarcinoma	507
LUSC	Lung squamous cell carcinoma	412
PRAD	Prostate adenocarcinoma	553
SARC	Sarcoma	269
SKCM	Skin cutaneous melanoma	475
STAD	Stomach adenocarcinoma	397
THCA	Thyroid carcinoma	571

In this study, we perform experiments with four different portions of missing values: 20%, 40%, 60%, and 80%.

2.3 TDimpute-DNAmeth Method

We performed the imputation of missing DNA methylations by using a similar deep neural network architecture to our previous study [32]. To measure the ability for predicting unknown cancer types, we separated the whole dataset into the target cancer dataset and the pan-cancer dataset with the exclusion of the target cancer. Each time, the "pan-cancer" dataset with 15 cancer types was used to train a "pan-cancer" model, which was then fine-tuned on the target cancer to obtain the target model.

2.3.1 *Deep Neural Network Architecture*

As shown in Fig.2(b), the fully-connected deep neural network learns to impute missing methylation values using three layers: an input layer, a hidden layer, and an output layer. The output at layer l is:

$$\boldsymbol{X}^l = f(\boldsymbol{W}^l\boldsymbol{X}^{l-1} + \boldsymbol{b}^l),$$

where $f(\bullet)$ is the activation function (Tanh, Relu, and Sigmoid), and $\boldsymbol{W}^l$ and $\boldsymbol{b}^l$ are the weight and bias parameters at layer l, respectively. $\boldsymbol{X}^0 = \boldsymbol{D}$ is the input DNA methylation, and the output of the last layer is the imputed values. RMSE is employed as the loss function:

$$J(\omega) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y^i - y_0^i)^2},$$

where y_0^i and y^i are the actual and the predicted DNA methylation for m sites, respectively. The deep neural network can be regarded as a high-dimensional nonlinear regression function that maps from the corrupted methylation data (input) to the methylation data (output). The purpose of training is to find the appropriate parameter vectors $\boldsymbol{W}$ and $\boldsymbol{b}$.

2.3.2 *Cross-Validation and Independent Test*

After introducing the missing values, the 16 original cancer datasets were divided into two parts: pan-cancer and target cancer. The general prediction model was first trained on the pan-cancer. Then the target model was established with the same structure as the general model. Next the parameters of the encoder layer, the hidden layer, and the decoder layer in the general model were transferred to the target model. Finally,

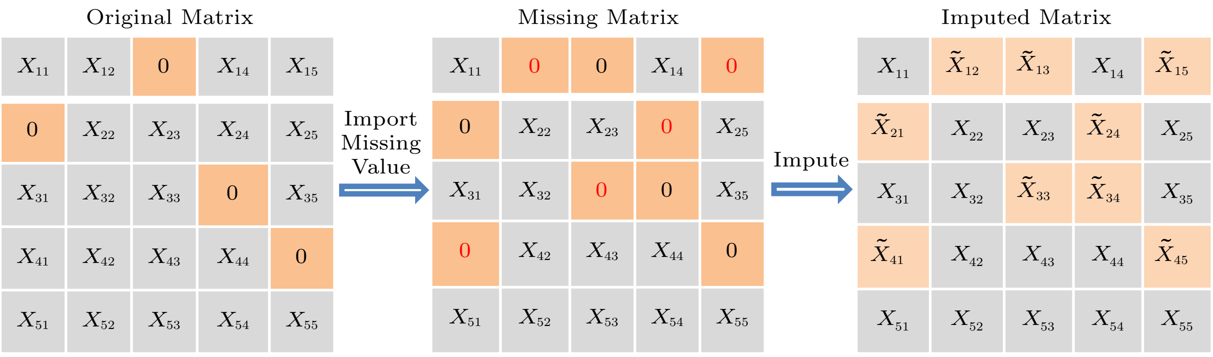

Original Matrix
X_{11} X_{12} 0 X_{14} X_{15}
0 X_{22} X_{23} X_{24} X_{25}
X_{31} X_{32} X_{33} 0 X_{35}
X_{41} X_{42} X_{43} X_{44} 0
X_{51} X_{52} X_{53} X_{54} X_{55}
Import Missing Value
Missing Matrix
X_{11} 0 0 X_{14} 0
0 X_{22} X_{23} 0 X_{25}
X_{31} X_{32} 0 0 X_{35}
0 X_{42} X_{43} X_{44} 0
X_{51} X_{52} X_{53} X_{54} X_{55}
Impute
Imputed Matrix
X_{11} $\tilde{X}_{12}$ $\tilde{X}_{13}$ X_{14} $\tilde{X}_{15}$
$\tilde{X}_{21}$ X_{22} X_{23} $\tilde{X}_{24}$ X_{25}
X_{31} X_{32} $\tilde{X}_{33}$ $\tilde{X}_{34}$ X_{35}
$\tilde{X}_{41}$ X_{42} X_{43} X_{44} $\tilde{X}_{45}$
X_{51} X_{52} X_{53} X_{54} X_{55}

Fig.1. Simulated missing value with red 0 denoting the simulated missing data and $\tilde{X}_{ij}$ for the imputed values.

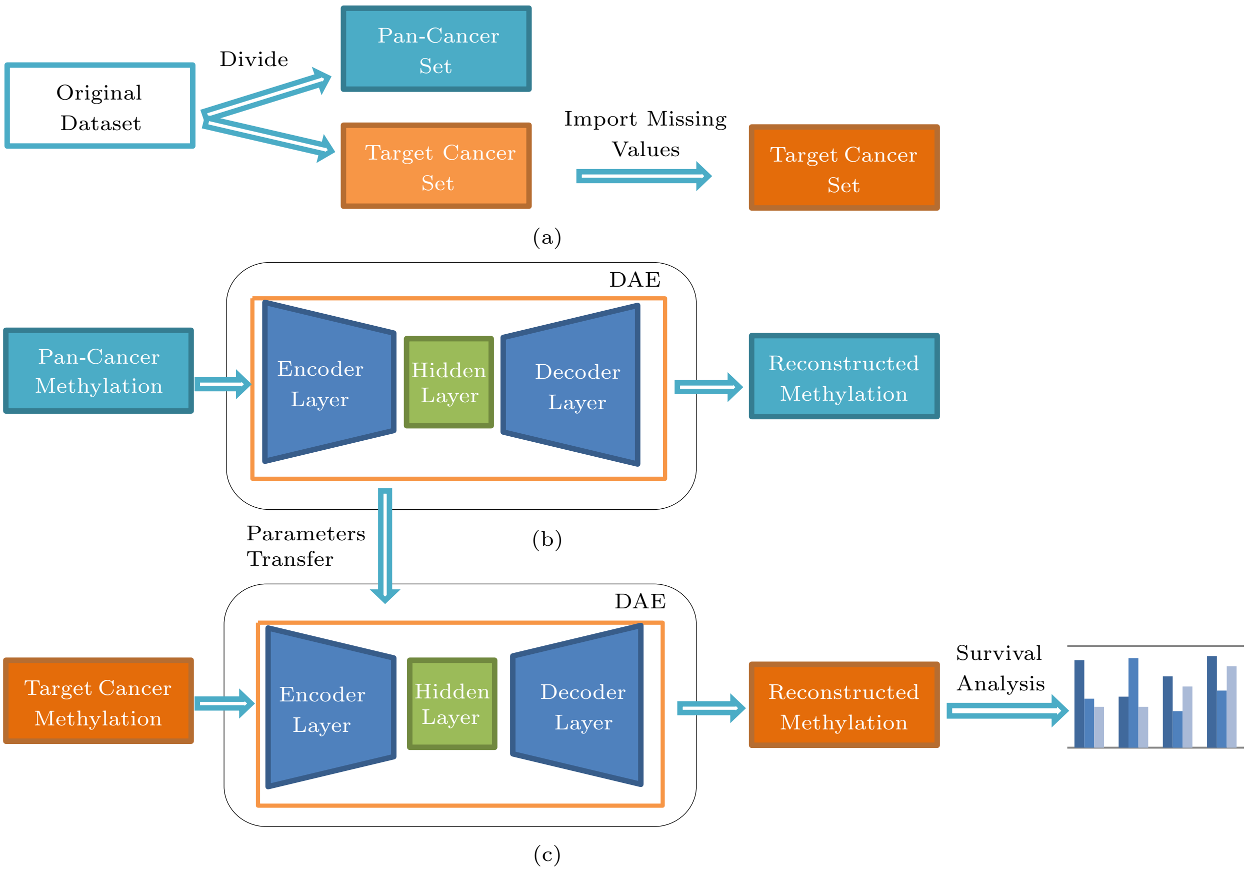


Fig.2. Step-by-step description of the processing flow. (a) The original dataset (TCGA) was divided into the pan-cancer set and the target set. In our simulated test, missing methylation values were randomly introduced to the target set. (b) The pan-cancer models were trained by the denoising autoencoder (DAE) model. And (c) the pan-cancer models were transferred to fine-tune on the target cancer. To further evaluate the usefulness of the imputed results, these imputed values were employed to perform survival analyses, and evaluated through the agreement between the predicted and the actual survival time.

target cancer data were used to fine-tune the target model, as shown in Figs.2(b) and 2(c). Specifically, the number of epochs was adjusted from 300 to 100 to avoid over-fitting in the target model.

We evaluated the performance of our model by cross-validation and independent test. For each target cancer dataset, 80% samples were randomly selected for cross-validation to select hyperparameters, and the left 20% were used for the independent test. In the cross-validation, the training dataset was randomly separated into five parts (folds), and each time four folds were employed to train a model and the left one-fold was used to test. This process was repeated for five times, and the performances of five predictions were averaged as the validation performance. With the optimized hyperparameters, a model was trained by the 80% samples and independently tested on the 20% tested samples.

2.3.3 *Parameters Setting*

The hyperparameters were optimized on the BLCA dataset to find the best performance. Table 2 shows all tested hyperparameters and the final optimal parameters.

Table 2. Tested and Selected Hyperparameters

Parameter Type	Tested Set	Selected
Hidden layer	{1, 2, 3}	1
Node size	{1 000, 2 000, 3 000, 4 000}	3 000
Epochs	{300, 600, 1 000}	300
Activation function	{Tanh, Relu, Sigmoid}	Sigmoid
Batch size	{16, 32, 64, 128}	64
Learning rate	{0.01, 1.0×10^{-3}, 1.0×10^{-4}, 1.0×10^{-5}}	1.0×10^{-4}

Fig.S2[2] shows the performances changed with two critical parameters, the hidden layer and the node size.

[2] https://github.com/zhoux85/TDimpute/blob/master/Imputing_DNA_methylation_by_transferred_learning_based_neural_network_Supplementary_file.pdf, Feb. 2022.

2.4 Survival Analysis

We evaluated the imputed values by the survival analysis using the ridge regression regularized Cox model implemented through the glmnet software package [34] in R, a model suitable for high-dimensional data fitting. The predictions were evaluated by the concordance index (C-index) and log-rank p-value. The C-index is defined as the fraction of all pairs of individuals whose predicted survival time is correctly ordered based on the Harrell's C statistics [35]. A C-index of 0.5 means a random prediction with the higher C-index values for better predictions. The log-rank p-value is the probability to separate patients into high-risk and low-risk groups better than by random.

2.5 Implementation of Other Methods

For comparison, we implemented five other methods, including the mean value, KNN [7], SVD [8], imputePCA [9] and missForest [11]. All methods were implemented with default parameters while optimizing a few key parameters. In details, we used $k = 30$ (optimized from 10, 20, 30, 50) closest neighbors for KNN, and 10 (from 5, 10, 20, 30) biggest eigenvalues for SVD. We did not include methyLImp since it needs complete samples with all values while our samples have missing values at different sites.

3 Results and Discussion

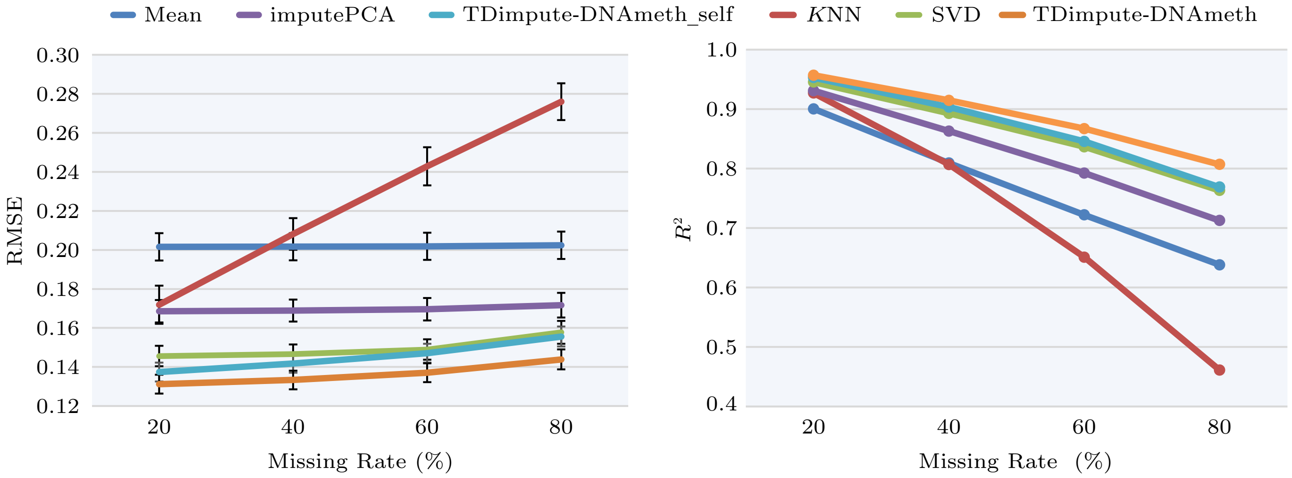
3.1 Assessment of Imputation Accuracy

To evaluate the performance of TDimpute-DNAmeth, we first performed cross-validation on 80% of the samples and tested it on 20% of the samples. Table 3 details the performance of two methods on the representative BLCA cancer dataset with the median sample size. Generally, both methods show essentially the same performances from both the 5-fold cross-validation (CV) and independent (Ind) tests at four missing rates, indicating the robustness of the methods. As expected, by using transferred learning, TDimpute-DNAmeth achieved consistently lower RMSE than TDimpute-DNAmeth_self. Similar results can be observed for other cancer types (Table S1[3]).

Table 3. RMSE of the 5-Fold Cross-Validation and Independent Test on BLCA

Missing Rate (%)	TDimpute-DNAmeth_self		TDimpute-DNAmeth	
	CV	Ind Test	CV	Ind Test
20	0.158 ± 0.005	0.158	0.148 ± 0.003	0.149
40	0.165 ± 0.005	0.165	0.152 ± 0.003	0.152
60	0.174 ± 0.005	0.173	0.157 ± 0.003	0.157
80	0.189 ± 0.005	0.188	0.168 ± 0.003	0.168

The comparison of TDimpute-DNAmeth and TDimpute-DNAmeth_self shows the impact of transfer learning on the accuracy of imputation. As shown in Fig. 3, TDimpute-DNAmeth is significantly better



(a)

(b)

Fig.3. Comparisons of six methods by the average (a) RMSE and (b) R^2 on 16 cancer datasets. The error bar shows the standard error of the mean.

[3] https://github.com/zhoux85/TDimpute/blob/master/Imputing_DNA_methylation_by_transferred_learning_based_neural_network_Supplementary_file.pdf, Feb. 2022.

than TDimpute-DNAmeth_self, with an average accuracy increase of 6.2%. Specifically, the accuracy of TDimpute-DNAmeth increased by 4.5%, 6.0%, 6.8%, and 7.6% compared with TDimpute-DNAmeth_self at the four different missing rates (20%, 40%, 60%, and 80%, respectively). Also, the advantages of the transfer learning method increased with the increase of the missing rate, since the correlation learned from pan-cancer plays an increasingly important role as the missing rate of the target cancer increases. The results directly prove that transfer learning is conducive to improving the accuracy of imputation. Fig.S3[4] shows the comparison result of TDimpute-DNAmeth and TDimpute-DNAmeth_self on 16 cancers. In addition, we compared the imputation results of the five methods at different methylation β-values, and found no significant changes (see Fig.S4[4]).

3.2 Comparison with Other Methods

We also evaluated the results of our TDimpute-DNAmeth method and existing competitive methods (Mean, KNN, imputePCA, and SVD) at different missing rates of 16 cancer datasets. As shown in Fig. 3, TDimpute-DNAmeth with transfer learning outperforms all the other methods. Specifically, the Mean and KNN methods are traditional methods with high error rates, and RMSE of KNN rapidly increases with the increasing missing rate, and is even worse than the one using mean values. This is reasonable as high missing rates mislead the selection of neighborhoods. The imputePCA achieved higher RMSE than SVD likely because imputePCA is based on fewer eigenvectors. The average RMSE by SVD is averagely 2.9% greater than TDimpute-DNAmeth_self, and 9.8% greater than TDimpute-DNAmeth. Since missForest was designed for 200 input variables (much smaller than our 20 000 variables) and ran slowly, we only ran it on the CESC dataset (Fig.S5[4]). Generally, missForest achieved similar performance to imputePCA: smaller RMSE at low missing rates but bigger RMSE at high rates. These two methods are consistently worse than SVD. Hereafter, we will not include imputePCA and missForest in our comparison. Similar trends are observed when measured by the average squared Pearson correlation coefficient (R^2) between the imputed and actual values over each sample. As shown in Fig. 3(b), though there is no obvious gap among SVD, TDimpute-DNAmeth_self, and TDimpute-DNAmeth methods, TDimpute-DNAmeth significantly outperforms the other methods at a higher missing rate. At the missing rate of 80%, the performance improved with 5.7% and 5.0%, compared with SVD and TDimpute-DNAmeth_self, respectively. Figs.S6 and S7[4] detail the RMSE and R^2 for all 16 cancer types, respectively. Besides the better performance, our method is also computationally efficient linearly increasing with the tested sample size. The TDimpute-DNAmeth method could pretrain a model and make prediction for BLCA with 440 samples within 0.2 seconds on the Titan 1080 GPU, close to the "mean" method. By comparison, the computational time of KNN and SVD strongly depends on the sample size, and is above 10 and 50 times longer in this test respectively.

3.3 Impact on Survival Analyses

To further validate the robustness and effectiveness of TDimpute-DNAmeth, we performed survival analyses on TCGA and imputed results from different imputation methods through the ridge regression regularized the Cox model. As shown in Fig.4, TDimpute-DNAmeth consistently performs the best at four missing rates. At a missing rate of 20%, the imputations by TDimpute-DNAmeth leads to a C-index of 0.649, while it is 0.65, 0.80, 0.83, and 1.8% higher than those by TDimpute-DNAmeth_self, SVD, KNN, and the imputation by the mean, respectively. The similar but lower

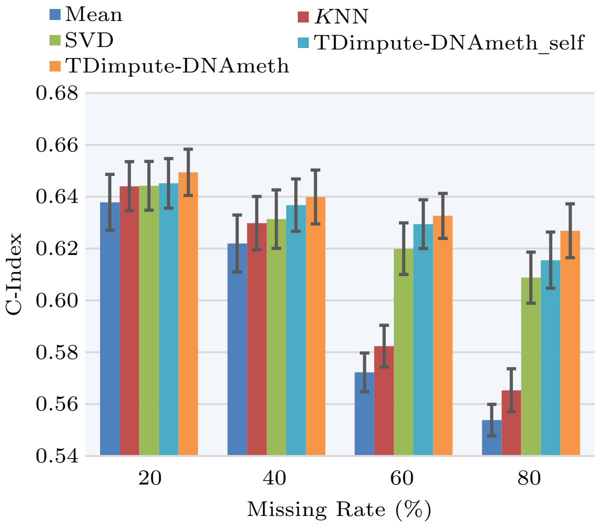


Fig.4. Comparison of methods by the average and standard error of the C-index over 16 cancer datasets through their imputed methylation values.

[4] Supplementary files include Figs.S1–S8 and Table S1. https://github.com/zhoux85/TDimpute/blob/master/Imputing_DNA_methylation_by_transferred_learning_based_neural_network_Supplementary_file.pdf, Feb. 2022.

performance by TDimpute-DNAmeth_self and SVD indicates the limit to using single cancer type and necessity to borrow information from other cancer types. As expected, the C-index values by the imputations decrease with the increase of missing rates for all methods, while the differences between imputation methods extend. At the missing rate of 80%, the imputations by TDimpute-DNAmeth achieve an average C-index of 0.627, which is 1.5 and 3.0% higher than TDimpute-DNAmeth_self and SVD, respectively. Again, the *K*NN and the Mean imputation methods cause a huge drop with C-index values of 0.554 and 0.565, respectively. Fig.S8[5] details C-index values according to actual or imputed by different methods for 16 cancer types.

3.4 Case Study on the SARC and CESC Datasets

To further indicate our method, we applied our method to the actually missing sites in the SARC and CESC dataset. To be consistent, we individually selected 20 000 sites with missing values from the datasets. As there were no known results for the actually missing values in the datasets, we assessed the imputations through their predictions of patient survival times, following our previous study of mRNA imputation[32]. By using the sites with missing rates above 20%, we performed survival analyses through the imputed values by different methods. As shown in Fig.5, using the imputed data by TDimpute-DNAmeth, the achieved C-index is 15.8% higher than the original data on SARC. The imputation by SVD causes a C-index of 0.577, 2.8% lower than TDimpute-DNAmeth on the SARC dataset. On the CESC dataset, the C-index accuracy of the TDimpute-DNAmeth is 27.2% higher than the original data and 9.5% higher than the one by SVD. Fig.6 compares the Kaplan-Meier survival curves by using the original data, and the imputed data by TDimpute-DNAmeth on SARC and CESC. By using the imputed values, the predicted high-risk patients can be better separated from low-risk patients in the plots, also indicated by the lower *P*-values.

4 Conclusions

In this study, we presented a new imputation method for DNA methylation data, namely TDimpute-DNAmeth. The transfer learning approach was applied to exploit the correlation of the DNA methylation among pan-cancer samples that effectively solved the problems of high dimensionality but a small sample size. Tested on the simulated missing DNA methylation data, TDimpute-DNAmeth was shown to outperform the existing methods in terms of both RMSE and R^2. Our results were further confirmed through survival analyses on the imputed data. More importantly, our method was also proved to outperform other methods by tests on the SARC and CESC datasets with actually missing DNA methylation values.

Our work was currently limited in the DNA methylation over the pan-cancer dataset. With the development of sequencing techniques, the method could be

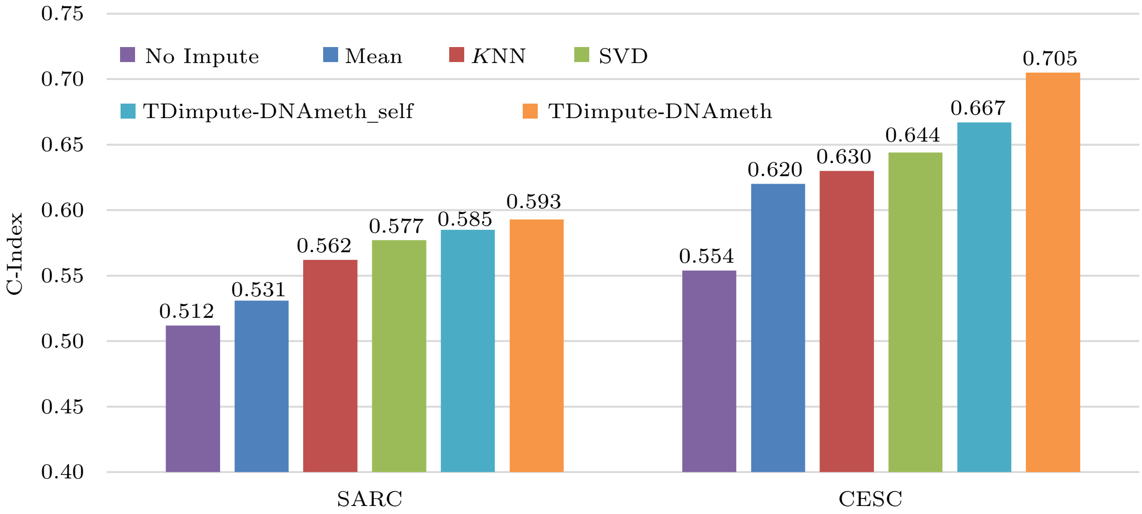
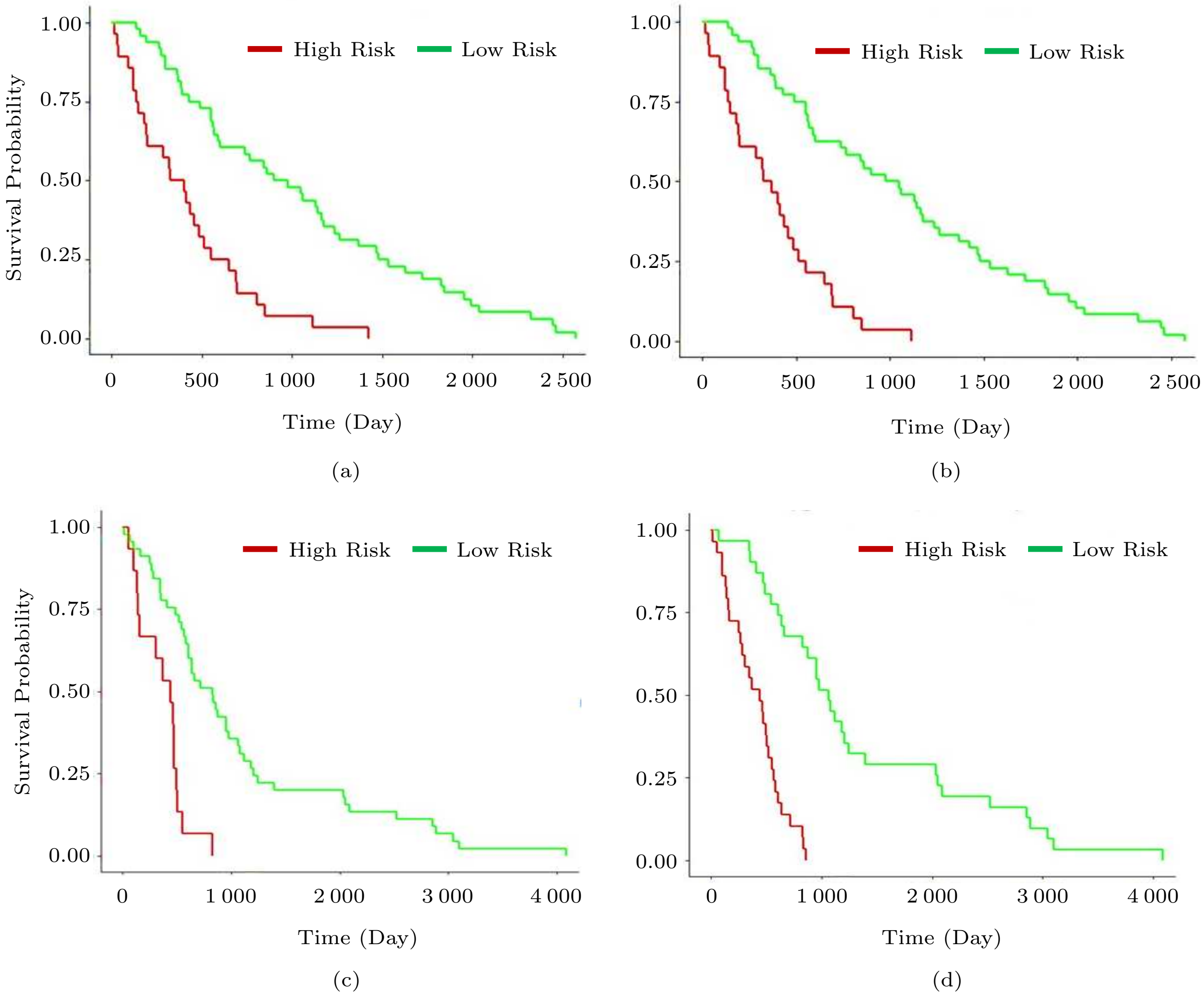


Fig.5. C-index value of survival prediction of original data (no imputation) and imputed methylation values on SARC and CESC.

[5] Supplementary files include Figs.S1–S8 and Table S1. https://github.com/zhoux85/TDimpute/blob/master/Imputing_DNA_methylation_by_transferred_learning_based_neural_network_Supplementary_file.pdf, Feb. 2022.



(a)

(b)

(c)

(d)

Fig.6. Kaplan-Meier plots of predicted high- and low-risk patients. (a) Original data (no impute, $P = 5.88 \times 10^{-7}$). (b) Predicted values by TDimpute-DNAmeth ($P = 9.34 \times 10^{-9}$) for SARC. (c) Original data (no impute, $P = 3.88 \times 10^{-6}$). (d) Predicted values by TDimpute-DNAmeth ($P = 6.72 \times 10^{-9}$) for CESC.

further applied to other omics types, other disease types, and other tasks such as age prediction and cell classification.

Acknowledgement(s) We would like to acknowledge The Cancer Genome Atlas (TCGA) to make the data publicly available.

References

[1] Francis R C. Epigenetics: The Ultimate Mystery of Inheritance. WW Norton & Company, 2011.

[2] Ye P, Luan Y, Chen K, Liu Y, Xiao C, Xie Z. MethSMRT: An integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Research*, 2016, 45(D1): D85-D89. DOI: 10.1093/nar/gkw950.

[3] Kulis M, Esteller M. DNA methylation and cancer. *Advances in Genetics*, 2010, 70(22): 27-56. DOI: 10.1016/B978-0-12-380866-0.60002-2.

[4] Gerd P. Defining driver DNA methylation changes in human cancer. *International Journal of Molecular Sciences*, 2018, 19(4): Article No. 1166. DOI: 10.3390/ijms19041166.

[5] Jouinot A, Assie G, Libe R *et al.* DNA methylation is an independent prognostic marker of survival in adrenocortical cancer. *The Journal of Clinical Endocrinology & Metabolism*, 2016, 102(3): 923-932. DOI: 10.1210/jc.2016-3205.

[6] Zhang G, Huang K C, Xu Z *et al.* Across-platform imputation of DNA methylation levels incorporating nonlocal information using penalized functional regression. *Genetic Epidemiology*, 2016, 40(4): 333-340. DOI: 10.1002/gepi.21969.

[7] Troyanskaya O, Cantor M, Sherlock G *et al.* Missing value estimation methods for DNA microarrays. *Bioinfor-*

matics, 2001, 17(6): 520-525. DOI: 10.1093/bioinformatics/17.6.520.

[8] Guttorp P, Fuentes M, Sampson P. Using transforms to analyze space-time processes. In *Statistical Methods for Spatio-Temporal Systems*, Finkenstadt B, Held L, Isham V (eds.), CRC/Chapman, 2006, pp.77-150.

[9] Josse J, Husson F. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 2012, 153(2): 77-99.

[10] Di Lena P, Sala C, Prodi A, Nardini C. Missing value estimation methods for DNA methylation data. *Bioinformatics*, 2019, 35(19): 3786-3793. DOI: 10.1093/bioinformatics/btz134.

[11] Stekhoven D J, Bühlmann P. MissForest-Non-Parametric missing value imputation for mixed-type data. *Bioinformatics*, 2012, 28(1): 112-118. DOI: 10.1093/bioinformatics/btr597.

[12] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521(7553): 436-444. DOI: 10.1038/nature14539.

[13] Heffernan R, Paliwal K, Lyons J *et al.* Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports*, 2015, 5: Article No. 11476. DOI: 10.1038/srep11476.

[14] Chen J, Zheng S, Zhao H, Yang Y. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *Journal of Cheminformatics*, 2021, 13(1): Article No. 7. DOI: 10.1186/s13321-021-00488-1.

[15] Senior A W, Evans R, Jumper J *et al.* Improved protein structure prediction using potentials from deep learning. *Nature*, 2020, 577(7792): 706-710. DOI: 10.1038/s41586-019-1923-7.

[16] Ching T, Himmelstein D S, Beaulieu-Jones B K *et al.* Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, 2018, 15(141): Article No. 20170387. DOI: 10.1098/rsif.2017.0387.

[17] Zheng S, Li Y, Chen S, Xu J, Yang Y. Predicting drug-protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, 2020, 2(2): 134-140. DOI: 10.1038/s42256-020-0152-y.

[18] Zheng S, Rao J, Zhang Z, Xu J, Yang Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, 2019, 60(1): 47-55. DOI: 10.1021/acs.jcim.9b00949.

[19] Way G P, Greene C S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac Symp Biocomput*, 2018, 23: 80-91. DOI: 10.1101/174474.

[20] Titus A J, Wilkins O M, Bobak C A, Christensen B C. Unsupervised deep learning with variational autoencoders applied to breast tumor genome-wide DNA methylation data with biologic feature extraction. `https://www.biorxiv.org/content/10.1101/433763v5`, Dec. 2021. DOI: 10.1101/433763.

[21] Lv X, Chen Z, Lu Y, Yang Y. An end-to-end Oxford Nanopore basecaller using convolution-augmented transformer. In *Proc. the 2020 IEEE International Conference on Bioinformatics and Biomedicine*, Dec. 2020, pp.337-342. DOI: 10.1109/BIBM49941.2020.9313290.

[22] Tian T, Wan J, Song Q, Wei Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 2019, 1(4): 191-198. DOI: 10.1038/s42256-019-0037-0.

[23] Lopez R, Regier J, Cole M B, Jordan M I, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 2018, 15(12): 1053-1058. DOI: 10.1038/s41592-018-0229-2.

[24] Zeng Y, Zhou X, Rao J, Lu Y, Yang Y. Accurately clustering single-cell RNA-seq data by capturing structural relations between cells through graph convolutional network. In *Proc. the 2020 IEEE International Conference on Bioinformatics and Biomedicine*, Dec. 2020, pp.519-522. DOI: 10.1109/BIBM49941.2020.9313569.

[25] Zhou X, Chai H, Zeng Y, Zhao H, Luo C H, Yang Y. scAdapt: Virtual adversarial domain adaptation network for single cell RNA-seq data classification across platforms and species. *Briefings in Bioinformatics*, 2021, 22(6): Article No. bbab281. DOI: 10.1093/bib/bbab281.

[26] Zhang Z, Zhao Y, Liao X *et al.* Deep learning in omics: A survey and guideline. *Briefings in Functional Genomics*, 2019, 18(1): 41-57. DOI: 10.1093/bfgp/ely030.

[27] The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 2020, 578(7793): 82-93. DOI: 10.1038/s41586-020-1969-6.

[28] Li Y, Wang L, Wang J, Ye J, Reddy C K. Transfer learning for survival analysis via efficient L2, 1-Norm regularized cox regression. In *Proc. the 2016 IEEE International Conference on Data Mining*, Dec. 2016, pp.231-240. DOI: 10.1109/ICDM.2016.0034.

[29] Yousefi S, Amrollahi F, Amgad M *et al.* Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports*, 2017, 7(1): Article No. 11707. DOI: 10.1038/s41598-017-11817-6.

[30] Yang X, Gao L, Zhang S. Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. *Briefings in Bioinformatics*, 2016, 18(5): 761-773. DOI: 10.1093/bib/bbw063.

[31] Hoadley K A, Yau C, Wolf D M *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 2014, 158(4): 929-944. DOI: 10.1016/j.cell.2014.06.049.

[32] Zhou X, Chai H, Zhao H, Luo C H, Yang Y. Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning-based neural network. *GigaScience*, 2020, 9(7): Article No. giaa076. DOI: 10.1093/gigascience/giaa076.

[33] Wei L, Jin Z, Yang S, Xu Y, Zhu Y, Ji Y. TCGA-assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*, 2017, 34(9): 1615-1617. DOI: 10.1093/bioinformatics/btx812.

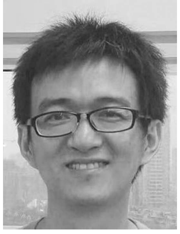
[34] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 2010, 33(1): 1-22.

[35] Van Belle V, Pelckmans K, Van Huffel S, Suykens J A. Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 2011, 53(2): 107-118. DOI: 10.1016/j.artmed.2011.06.006.

Xin-Feng Wang is a Ph.D. candidate in the School of Computer Science and Engineering at the Sun Yat-sen University, Guangzhou. His research interests include deep learning, multi-omics integration, and complex physiologic signals.

Xiang Zhou received his B.S. degree in electronic and information engineering from Northwest Agricultural and Forest University, Xianyang, in 2012, and his M.S. degree in computer science and technology from Sun Yat-sen University, Guangzhou, in 2017, where he is currently working towards his Ph.D. degree. His research interests include computational biology and deep learning methods for single cell data analysis.

Jia-Hua Rao is a Ph.D. candidate in the School of Computer Science and Engineering at the Sun Yat-sen University, Guangzhou. His research interests include deep learning, multi-omics integration, knowledge graph and computational biology.



Zhu-Jin Zhang received his Ph.D. degree in control science and technology from Huazhong University of Science and Technology, Wuhan, in 2011. He is currently a visiting scholar at Sun Yat-sen University, Guangzhou. His research interests include bioinformatics and machine learning.

Yue-Dong Yang is a professor in the School of Computer Science and Engineering and National Super Computer Center at Guangzhou, Sun Yet-sen University, Guangzhou. He received his Ph.D. degree in the computational biology from the University of Science and Technology of China (USTC), Hefei, in 2006. Dr. Yang has published more than 100 articles that have been cited more than 4 000 times, including five ESI highly cited articles. Currently his research group emphasizes on developing HPC and AI algorithms for multi-scale integration of omics data and intelligent drug design. He is also responsible for constructing the HPC platform for biomedical applications based on the Tianhe-2 supercomputer.