Future Generation Computer Systems 135 (2022) 20–29

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

DWT-CV: Dense weight transfer-based cross validation strategy for model selection in biomedical data analysis

Jianhong Cheng [a,b], Hulin Kuang [a,*], Qichang Zhao [a], Yahui Wang [a], Lei Xu [a], Jin Liu [a], Jianxin Wang [a]

[a] Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China
[b] Institute of Guizhou Aerospace Measuring and Testing Technology, Guiyang 550009, China

ARTICLE INFO

Article history:
Received 14 January 2022
Received in revised form 30 March 2022
Accepted 23 April 2022
Available online 30 April 2022

Keywords:
Model selection
Cross validation
Weight transfer
Biomedical data

ABSTRACT

Model selection for deep learning algorithms is an extremely important step in the process of extracting knowledge from limited data, especially in biomedical data. The common approach is to adopt cross-validation techniques to randomly divide a small subset of the training set as the validation data for parameter tuning and model selection. However, this method may choose a sub-optimal model due to insufficient data utilization, and the process, such as *k*-fold cross-validation, is cumbersome and time-consuming. In this study, we propose a dense weight transfer-based cross validation (DWT-CV) strategy for biomedical data analysis and use this strategy to improve the generalization of deep learning algorithms with reduced training time using weight transfer learning. DWT-CV utilizes a dense weight aggregation and weight transfer mechanism to make the model more general and converge faster during the cross validation. The effectiveness of the proposed strategy is evaluated on multiple experiments with three different domains including biomedical image classification, drug–target affinity prediction, and medical image segmentation. Extensive experimental results demonstrate that our proposed DWT-CV strategy can make several deep learning benchmark methods perform better on multiple biomedical datasets, which implies that it may be an alternative to the traditional cross validation criterion for model selection.

1. Introduction

Biomedical data analysis, such as biomedical image classification and image segmentation, is an integral part of artificial intelligence applications in biomedicine and precision medicine [1–3]. These applications usually require models with extremely high generalization performance to be deployed on various application platforms. A good generalized model usually requires a large and rich annotated data set. However, some scenarios, such as biomedical data, require high cost to obtain annotation data, making it challenging to train a general model in real life, especially for limited annotation data.

Cross-validation (CV) [4,5] is a model-agnostic technique commonly used in machine learning or deep learning for model selection and performance evaluation. Deep learning-based methods, especially convolutional neural networks (CNNs), have recently achieved impressive success in many intelligent tasks such as biomedical image classifications [6,7], image segmentations [8], and drug–target predictions [9,10]. While CNNs have become the most famous benchmark in deep learning and surpassed human performance in certain domains [3,11–18], their training process is usually based on cross-validation and iterative optimization manner to fine-tune hyperparameters and select models. However, CV may potentially introduce dataset shift problems [5], which is a harmful factor that is often not overlooked and may lead to inaccurate performance estimation. Previous studies have proposed some strategies to alleviate these problems. For example, stratified cross-validation (SCV) was proposed to avoid dataset shift by maintaining equal class distributions in all partitions [19]. Distribution-balanced SCV attempted to minimize the covariate shift by keeping the data distribution between the training fold and the test fold as similar as possible [20]. Importance-weighted CV was also designed for unbiased classification under covariate shift [21]. Furthermore, some scholars have also studied the impact of partition-induced covariate shift on different *k*-fold cross validation (*k*-FCV) schemes [22]. The *k*-FCV mentioned above needs to be repeatedly trained and selected from scratch, which is cumbersome and time-consuming. For simplification, some scholars divide a small subset from training dataset as the fixed validation data for parameter turning and

* Corresponding author.
E-mail addresses: jianhong_cheng@csu.edu.cn (J. Cheng), hulinkuang@csu.edu.cn (H. Kuang), zhaoqichang@csu.edu.cn (Q. Zhao), yahuiwang@csu.edu.cn (Y. Wang), louris@csu.edu.cn (L. Xu), liujin06@csu.edu.cn (J. Liu), jxwang@mail.csu.edu.cn (J. Wang).

https://doi.org/10.1016/j.future.2022.04.025
0167-739X/

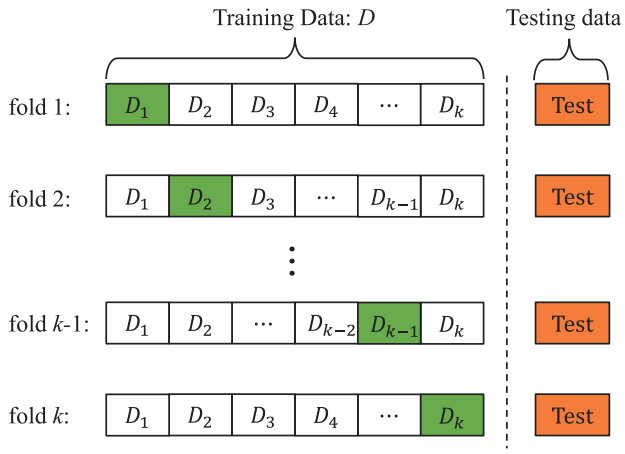

Fig. 1. The traditional k fold cross-validation, namely k-FCV. The part of the white box is used as the training set, and the green part is used as the validation set. D_k denotes the kth portion of the data D after being divided into equal parts.

model selection [23,24]. This method is very convenient, but the performance of deep learning models may be affected by data bias, especially for the limited data samples.

In the absence of an independent testing dataset, k-FCV is usually used to estimate the performance of prediction models. For example, Deepak et al. used 5-FCV on the magnetic resonance imaging (MRI) dataset to evaluate their proposed classification system [25]. Yang et al. proposed a CNN model based on transfer learning for glioma grading and evaluated the performance of the model using 5-FCV [26]. Iesmantas et al. employed a convolutional capsule network (CapsNet) trained with cross-validation for classification of breast cancer histology images [27]. Esteva et al. utilized a CNN to classify three-class disease partition and nine-class disease partition and demonstrated that the overall accuracy of the CNN trained with 9-FCV was better than that of two dermatologists [28]. 3-FCV and 4-FCV were performed and used to evaluate their works [29]. Another case is that k-FCV is used to prevent overfitting and select optimal model when the independent testing dataset is present. There are numerous outstanding works trained with k-FCV to obtain the optimal parameters of their models. A multiple instance learning-based deep learning system was presented and used for diagnosing tissue types [30], a total of 44,732 whole slide images from 15,187 patients were evaluated, and 15% of which were used for tuning the hyperparameters. Our previous study proposed an RAAU-Net for brain tumor segmentation and trained the network with 5-FCV on the training data for tuning the hyperparameters [31]. These networks optimize the learning rate and prevent overfitting by determining whether the validation loss decreases within different fixed epochs. These strategies can reduce time-consuming and avoid unnecessary epochs of iterative training compared with the scheme fixed training epochs. However, there are still some challenges of insufficient data utilization and high cost of training time when cross-validation experiments are performed.

In this study, we only focus on the situation where k-FCV is used for model selection on given training data and evaluated on the independent testing data. As shown in Fig. 1, the original training set is randomly partitioned into k subsets. Of these k subsets, a single subset is used as the validation data for tuning parameters of the model, and the remaining $k-1$ subsets are retained as the training data. The cross-validation is then repeated k times, with each selected model is evaluated on the independent testing data. These k results from k folds are then averaged to generate the performance estimation. From this process of CV, we can easily find that there are the following drawbacks: (1) part of the data is divided into validation data to select the model, and this part of the validation data is not really involved in the training model, which leads to under-utilization of the data; (2) each fold requires training from scratch, which brings a long training period; and (3) this process will obtain k models with different performances, which are difficult to choose in the practical application. To remedy this, we improve the process of k-FCV and propose a dense weight transfer-based cross validation (DWT-CV) for selecting a more general model with less time-consuming. The proposed DWT-CV makes use of a dense weight aggregation and weight transfer mechanism to make the model more general and converge faster during the cross validation. The contributions of this work include:

(1) We propose a successive weight transfer-based cross validation for model selection to improve the performance of deep learning model with reduced training time.
(2) To alleviate data partitioning bias, we further propose a dense weight transfer-based cross validation for model selection to make the deep learning model more generalized by using a dense weight aggregation mechanism.
(3) Experimentally validating the effectiveness of the proposed training strategy to multiple benchmark methods on twelve publicly available datasets for biomedical data analysis.

Preliminary results have been published in ISBRA 2021 conference [32]. Compared with the previous strategy, we further improve it by using a dense weight aggregation mechanism to alleviate data partition bias. Substantial experiments are conducted in this study by extensively validating the effectiveness of the proposed method on twelve various domains datasets. Furthermore, the experimental results are analyzed in detail, more benchmark methods are compared to illustrate the superiority, and some insightful discussions are presented.

The remainder of this study is organized as follow. Section 2 introduces the methodology of this study. Section 3 presents the experiment datasets, benchmark methods, and the results analysis. Section 4 reports the discussion about our study, and Section 5 provides a conclusion.

2. Method

2.1. k-fold cross-validation for model selection

The k-FCV is commonly used as an appropriate model selection criterion to estimate the generalization performance for machine learning algorithms. In the presence of independent testing data, k-FCV is employed as a model selection strategy. Specifically, the data are randomly partitioned to generate k disjoint subsets of approximately equal size. Meanwhile, to reduce the impact of dataset shift, the k-fold stratified cross-validation is employed to make the number of samples of each class in each partition equal. As shown in Fig. 1, in the kth fold, the kth subset is used as the validation data to select a model trained on the remaining $k-1$ subsets under the early stopping criterion. The results of k models evaluated on the independent testing data are averaged to generate the performance estimation. The method follows the steps detailed in Algorithm 1.

2.2. Weight transfer-based cross-validation for model selection

To alleviate the drawbacks of the traditional k-FCV strategy, we first attempt to propose a weight transfer-based cross-validation (WT-CV) strategy for the model selection. Specifically, weight transfer learning is introduced into the cross-validation

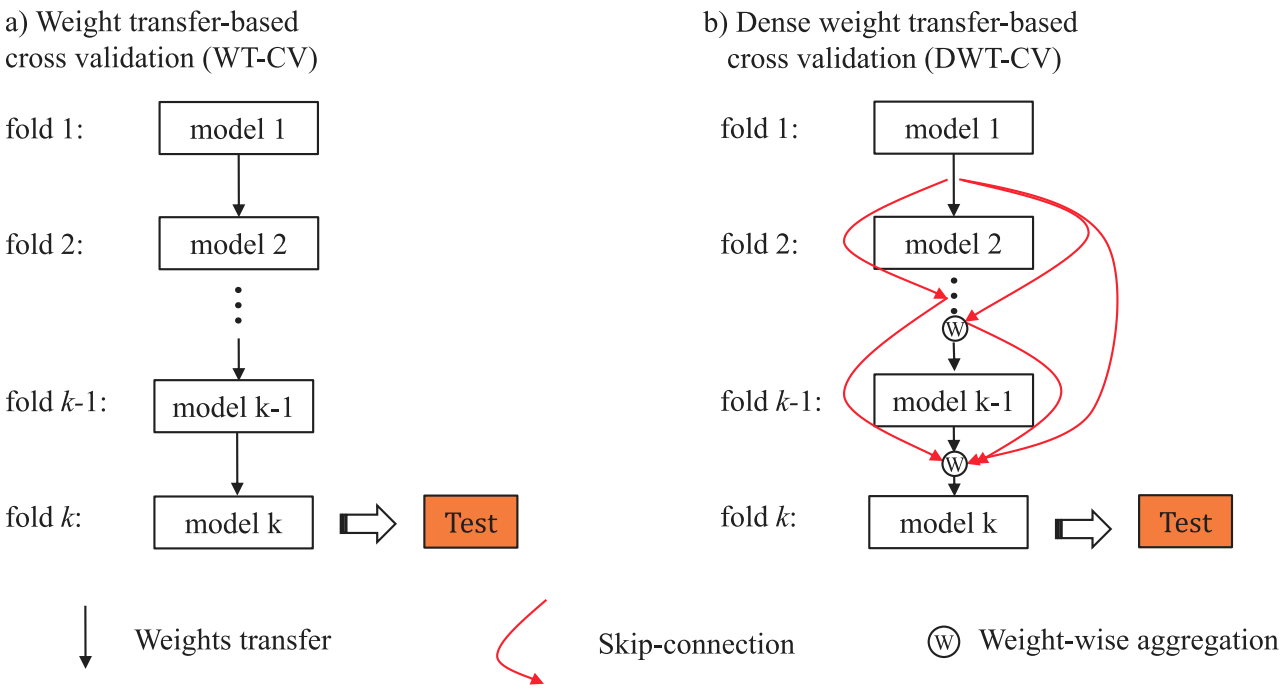

Fig. 2. The proposed WT-CV and DWT-CV strategies. Weight-wise aggregation is to aggregate the weights of all previous $k-1$ models by using weight-wise average as the initial weights of the current model.

procedure for selecting more general models with reduced training time. As shown in Fig. 2(a), we randomly split the available training data into k disjoint parts of approximately equal. As mentioned in the k-FCV, k-1 parts of which are used as the training set and the remaining one are used as the validation set. For each fold, the model is selected by fine-tuning under the early stopping criterion, and the training weight of the previous fold is used as the initial weight of the next fold in turn, until the last fold. Then, we use an independent testing data to evaluate the performance of the final trained model. This method follows the steps detailed in Algorithm 2.

Algorithm 1 k-FCV for model selection and performance estimation.

Require: Training data:D; Independent testing data:D_t; Criterion;
Deep learning model: E;
// Training and selecting model using k-FCV
Perform data partitioning with k folds.
for $i \leftarrow 1$ to k folds **do**
  Assign the training set: $D_t \leftarrow D \backslash D_i$
  Assign the validation set: $D_v \leftarrow D_i$
  Initialize deep learning model parameters: θ_i
  for epoch $e = 1, 2, \ldots$ **do**
    Update parameters θ_i on D_t; //back propagation
    if Criterion on D_v **then**
      Save the model parameters θ_i;
      break;
    end if
  end for
end for
// performance estimation
for $i \leftarrow 1$ to k **do**
  $R[i] \leftarrow E(\theta_i, D_t)$; // testing results
end for
return models:$\{E(\theta_i) \mid i = 1, 2, \ldots, k\}$; $Average(R)$.

2.3. Dense weight transfer-based cross-validation for model selection

Although the WT-CV algorithm is convenient and time-saving for model selection compared to the traditional k-FCV, it may be subject to data partitioning bias as the number of fold iterations increases [32]. To alleviate this problem, inspired by soft voting in ensemble learning, we design a DWT-CV strategy for the model selection by fully utilizing the learned weights of each fold.

Algorithm 2 WT-CV for model selection and performance estimation.

Require: Training data:D; Independent testing data:D_t; Criterion;
Deep learning model: E;
// Training and selecting model using k-FCV
Perform data partitioning with k folds.
for $i \leftarrow 1$ to k folds **do**
  Assign the training set: $D_t \leftarrow D \backslash D_i$
  Assign the validation set: $D_v \leftarrow D_i$
  if i==1 **then**
    Initialize deep learning model parameters: θ_i
  else if i>1 **then**
    Load the model parameters:θ_{i-1}
  end if
  for epoch $e = 1, 2, \ldots$ **do**
    Update parameters θ_i on D_t; //back propagation
    if Criterion on D_v **then**
      Save the model parameters θ_i;
      break;
    end if
  end for
end for
// performance estimation
$R \leftarrow E(\theta_k, D_t)$; // testing results
return model:$E(\theta_k)$; R.

Algorithm 3 DWT-CV for model selection and performance estimation.

Require: Training data:D; Independent testing data:D_t; Criterion; Deep learning model: E; Measurement:M;
// Training and selecting model using k-FCV
Perform data partitioning with k folds.
for $i \leftarrow 1$ to k folds **do**
  Assign the training set: $D_t \leftarrow D \backslash D_i$;
  Assign the validation set: $D_v \leftarrow D_i$;
  if i==1 **then**
    Initialize deep learning model parameters: θ_i;
  else if i>1 **then**
    // Weights aggregation using average operation
    $\theta' \leftarrow average(\theta_1, ..., \theta_{k-1})$;
    Load the model parameters: θ';
  end if
  for epoch $e = 1, 2, ...$ **do**
    Update parameters θ_i on D_t; //back propagation
    if Criterion on D_v **then**
      Save the model parameters θ_i;
      break;
    end if
  end for
end for
// performance estimation
$R \leftarrow E(\theta_k, D_t)$; // testing results
return model:$E(\theta_k)$; R.

Specifically, DWT-CV makes use of a dense weight aggregation and weight transfer mechanism to make the model more general and converge faster during the cross validation. The dense weight aggregation is to incorporate model weights from previous folds by performing a weight-wise average. As shown in Fig. 2(b), for the kth fold, the weights from the all previous $k-1$ models are aggregated by weighted average as the initial weights of the current model and the current kth subset is used as validation data to select a model. Then, the final selected model is used to evaluate the performance of the independent testing data. This method follows the steps detailed in Algorithm 3.

3. Experiments and results

In order to evaluate the effectiveness of our proposed training strategy, we conduct extensive experiments on 12 public biomedical data benchmarks, including biomedical image classification, drug–target relationship prediction, and medical image segmentation.

3.1. Experimental datasets

The summary of datasets used in this study is shown in Table 1. We first use eight biomedical image classification datasets from MedMNIST to validate the effectiveness of the proposed method [23,24]. Furthermore, we also validate the proposed training strategy on two drug–target relationship (DTR) datasets [33,34] and two stroke lesion segmentation datasets [35–37].

3.1.1. MedMNIST

As mentioned before, we have first employed 8 biomedical image classification datasets to validate the proposed method in the following experiments. For the biomedical image classification, we employ three 2D and five 3D public standardized datasets from the MedMNIST database,[1] which is a large-scale lightweight public benchmark dataset for 2D and 3D biomedical image classification [23,24]. The datasets used in this study cover multiple data modalities (e.g., X-ray, Ultrasound, CT, MRA) and diverse classification tasks (Binary and Multi-class). The usage of these datasets describes as follows:

PneumoniaMNIST consists of 5856 pediatric chest X-ray images, which are used for binary classification of pneumonia and normal. These gray images are center-cropped and resized into $1 \times 28 \times 28$.

BreastMNIST consists of 780 breast ultrasound images used for binary classification of benign and malignant. These ultrasound images are resized into $1 \times 28 \times 28$.

DermaMNIST is a large dataset of multi-source dermatoscopic images of common pigmented skin lesions, which is composed of 10,015 dermatoscopic images used for distinguishing 7 different categories. These images are resized into $3 \times 28 \times 28$.

OrganMNIST3D consists of 1743 abdominal CT images, which is used to perform multi-class classification of 11 body organs. Each image is processed into $28 \times 28 \times 28$ using 3D bounding boxes.

NoduleMNIST3D consists of 1849 lung nodule chest CT images used to classify two types of malignancy levels. Each data is center-cropped the spatially normalized images into $28 \times 28 \times 28$.

AdrenalMNIST3D consists of 1584 left and right adrenal glands CT images, which are used for distinguishing normal adrenal gland from adrenal mass. These CT images are also center-cropped and resized into $28 \times 28 \times 28$.

FractureMNIST3D consists of 1370 chest CT images of rib fractures and is used for classifying three types of rib fractures (*i.e.*, buckle, nondisplaced, and displaced). For each annotated fracture area, the images are center-cropped into $28 \times 28 \times 28$.

VesselMNIST3D consists of 1909 brain vessels collected by reconstructing MRI images, which are used for binary classification of the aneurysm and healthy vessel segments. For the data preprocessing, the non-watertight mesh is fixed with PyMeshFix [38] and then voxelized into $28 \times 28 \times 28$.

The number of training samples and testing samples of these datasets is shown in Table 1.

3.1.2. Drug–target relationship datasets

Davis[2] is widely used as a benchmark data for drug–target affinities (DTA) prediction [33]. The Davis dataset contains selectivity assays of the kinase protein family and the relevant inhibitor and the binding affinity values are measured in dissociation constant (Kd). The Kd values are usually transformed into logspace (pKd) to reduce the large variance, which is ranging from 5.0 to 10.8. Finally, 24,044 samples are used as the training data, and 6011 samples are used as the testing data.

DrugBank[3] is the most commonly used dataset for drug–target interactions (DTI) prediction [34]. We use version 4.5.0 (released on 2016-04-20) as the dataset. After brushing, there are 1412 drugs, 1520 proteins, and 6262 positive samples in the dataset. We randomly select negative samples from the unlabeled compound–protein pairs which have an equal number as positive samples. Finally, 10,019 samples are used as the training data, and 2504 samples are used as the testing data.

[1] https://medmnist.com/.
[2] https://github.com/hkmztrk/DeepDTA/tree/master/data/davis.
[3] https://github.com/Bjoux2/DeepDTIs_DBN/tree/master/data.

Table 1
Summary of the datasets used in this study.

Task type	Dataset	Data modality	Task (Classes/Labels)	Training samples	Testing samples
Biomedical image classification	PneumoniaMNIST	Chest X-ray	Binary-class (2)	5232	624
	BreastMNIST	Breast ultrasound	Binary-class (2)	624	156
	DermaMNIST	Dermatoscope	Multi-class (7)	8010	2005
	OrganMNIST3D	Abdominal CT	Multi-class (11)	1133	610
	NoduleMNIST3D	Chest CT	Binary-class (2)	1323	526
	AdrenalMNIST3D	Abdominal CT	Binary-class (2)	1286	298
	FractureMNIST3D	Chest CT	Multi-class (3)	1130	240
	VesselMNIST3D	Brain MRA	Binary-class (2)	1527	382
Drug–target prediction	Davis	Drug–target affinity	Regression (2)	24,044	6011
	DrugBank	Drug–target interaction	Binary-class (2)	10,019	2504
Biomedical image segmentation	ISLES2018	Multi-modal brain CT	Binary-class (2)	94	62
	AISD	Brain CT	Multi-class (2)	317	80

3.1.3. Stroke lesion segmentation datasets

ISLES2018[4] is the ischemic stroke lesion segmentation challenge 2018 dataset [35,39], which provides 94 brain multi-modal CT images with segmentation labeled images as the training data, and 62 unlabeled CT images as the testing data. The multi-modal images including a CT scan and four perfusion maps (CBF, CBV, MTT and Tmax) are used. The slice of each image is ranging from 2 to 22 with a thickness of 5 mm and a resolution of 256×256. In our study, the size of each modality is cropped and padded into $5 \times 208 \times 176$. These five modalities are concatenated along the channel dimension as network inputs.

AISD dataset[5] contains 397 non-contrast-enhanced CT scans with segmentation ground truths [36]. To evaluate our methods, we randomly divided the dataset into 80% of data as the training dataset and the remaining 20% of data as the testing dataset. The thickness of CT is 5 mm and resolution is 512×512. In our study, the size of each image is cropped and padded into $16 \times 320 \times 320$.

3.2. Benchmark methods

For the biomedical image classification, we adopt the ResNet-'X' [40] variants such as ResNet-18, ResNet-34, and ResNet-50 as backbones used to valid the proposed method. These networks mainly consist of multiple convolutional blocks, multiple residual blocks, a global average pooling (GAP) layer, and a fully connected layer. The convolutional block consists of a convolution layer, a batch normalization layer, and a ReLU layer. Each residual block is composed of multiple convolutional layers, a batch normalization layer, and a shortcut connection. For the 3D biomedical data, we extend 2D ResNet-'X' into 3D architecture for the classification task.

For the drug–target relationship predictions, we adopt DeepDTA as the benchmark method [9], which comprises two CNN blocks to learn high-level representations from drug SMILES strings and protein sequences, to predict binding affinity. The benchmark DeepConV-DTI [10] is used to predict drug–target interaction, which utilizes DNN to process ECFP fingerprints of drugs and applies multi-scale 1D-CNN to process protein sequences.

For the medical image segmentation, we employ a novel framework, namely CoTr [41], as the benchmark method for stroke lesion segmentation. This framework is composed of convolutional neural networks and transformers to extract feature representations while capturing long-range dependencies, which performs well on medical image segmentation. For each of these benchmark methods, three strategies k-FCV, WT-CV, and DWT-CV are performed in this study.

[4] https://www.smir.ch/ISLES/Start2018.

[5] https://github.com/griffinliang/aisd.

Table 2
Parameter settings of four benchmark networks. Adam and SGD are adaptive moment estimation and stochastic gradient descent, respectively.

Network	Optimizer	Learning rate	Early stopping
ResNet-'X' [40]	Adam	1e−3	10
DeepConV-DTI [10]	Adam	1e−3	20
DeepDTA [9]	Adam	1e−3	5
CoTr [41]	SGD	1e−2	50

3.3. Experimental settings

Our whole experimental procedure is implemented on an NVIDIA TESLA V100 GPU, and the development of all networks is based on the Pytorch library. During the training, the cross-entropy loss is used as the loss function for the biomedical data classification task. Mean square error (MSE) loss is used as the loss function for the regression task. The sum of dice loss and cross-entropy loss is used as the loss function for the medical image segmentation task [42]. Other parameter settings of three networks are presented in Table 2. For example, Adam is used as an optimizer with an initial learning rate of 1e−3 for ResNet-'X'. The criterion for model selection is based on the validation loss, that is, the model will be saved when the validation loss decreases. To avoid overfitting, the early stopping strategy is adopted, that is, the training will be terminated if the validation loss does not decrease within 10 epochs. These default values are set by experiments conducted by our used datasets.

For the measurements, area under the ROC curve (AUC) and accuracy (ACC) are used as the measurements of biomedical image classification. AUC is a threshold-free measurement to assess the continuous prediction values, while ACC can be used to evaluate the discrete prediction labels according to the threshold or Argmax. As we know, AUC is less sensitive to class imbalance than ACC. In the biomedical image classification, the class imbalance is not severe, thus ACC can be also served as a good measurement [23]. In addition, MSE and Dice similarity score (Dice) are used as the main measurements for regression and segmentation tasks. For the biomedical image classification, the results are obtained by the average of three tails. For the drug–target prediction, the results are obtained by the average of three trails. For the stroke lesion segmentation, the results are obtained by one trial. To make a tradeoff between the effectiveness and training time, we set $k = 5$ in our all experiments.

3.4. Assessment on biomedical image classification

3.4.1. The effectiveness of the proposed method

The overall performance of the methods on different training strategies are presented in Table 3. The comparison results show that the benchmark networks trained with DWT-CV perform well on most datasets overall. Compared with the traditional k-FCV

Table 3
Performance of our proposed method for three backbone networks on 8 datasets for biomedical image classification.

Dataset	Training strategy	ResNet-18			ResNet-34			ResNet-50		
		AUC (%)	ACC (%)	Time (min)	AUC (%)	ACC (%)	Time (min)	AUC (%)	ACC (%)	Time (min)
PneumoniaMNIST	*k*-FCV,	94.92	85.23	2.67	95.59	84.55	6.00	94.57	83.73	7.97
	WT-CV	94.99	85.47	2.56	94.83	84.67	5.15	93.58	83.65	7.23
	DWT-CV	**95.69**	**88.67**	2.54	**95.20**	**85.04**	5.71	**94.96**	**85.74**	7.54
	Official	93.40	81.10	0.56	95.88	83.12	1.18	95.08	83.76	2.15
BreastMNIST	*k*-FCV	85.15	82.31	0.60	82.37	80.04	1.20	84.75	80.64	1.32
	WT-CV	**89.28**	84.40	0.57	88.29	82.05	1.00	87.22	83.54	1.13
	DWT-CV	89.10	**85.68**	0.56	88.52	**84.40**	1.06	**89.77**	**84.83**	1.20
	Official	86.67	82.69	0.14	**88**.89	83.76	0.37	85.33	80.13	0.34
DermaMNIST	*k*-FCV	91.45	73.93	5.47	91.32	73.90	11.74	90.68	73.38	14.83
	WT-CV	91.11	74.06	3.29	90.82	73.72	7.41	90.85	73.55	10.45
	DWT-CV	**91.47**	**74.48**	3.35	**91.67**	**74.75**	7.45	**90.86**	**74.00**	10.84
	Official	91.36	74.31	0.95	91.58	74.73	2.69	90.86	73.81	3.48
OrganMNIST3D	*k*-FCV	99.24	88.28	61.02	99.49	89.82	87.28	99.27	89.30	69.74
	WT-CV	**99.55**	90.93	30.14	99.34	89.29	41.78	99.34	90.22	46.80
	DWT-CV	99.44	**91.20**	25.54	**99.49**	**90.05**	49.10	**99.44**	**90.22**	52.32
	Official	99.47	90.49	10.38	99.10	86.94	10.65	99.33	89.89	15.99
NoduleMNIST3D	*k*-FCV	89.27	90.87	30.52	90.94	91.29	60.02	90.12	87.32	49.13
	WT-CV	90.69	90.05	21.49	93.60	90.49	42.10	88.56	91.19	37.42
	DWT-CV	**91.17**	**91.25**	22.11	**94.11**	**91.51**	49.18	89.23	**91.76**	38.81
	Official	90.92	91.11	7.36	89.27	90.46	12.75	**91.01**	90.68	11.49
AdrenalMNIST3D	*k*-FCV	83.75	80.52	21.75	86.75	81.33	48.68	86.33	80.34	38.12
	WT-CV	80.70	79.64	20.12	80.45	79.08	40.16	86.35	80.20	34.58
	DWT-CV	**86.63**	**81.20**	19.72	**86.88**	**81.54**	41.55	85.03	**82.44**	37.47
	Official	86.06	81.09	5.12	86.32	81.32	8.26	**86.41**	80.20	8.33
FractureMNIST3D	*k*-FCV	71.78	52.47	19.32	72.88	55.14	41.40	72.24	50.94	37.26
	WT-CV	68.31	50.84	16.61	72.59	53.20	29.65	65.93	47.64	29.16
	DWT-CV	**72.32**	**53.06**	15.68	**75.11**	**56.53**	30.30	69.20	**51.94**	29.66
	Official	67.23	46.25	3.90	72.71	52.64	9.58	**72.76**	50.83	8.66
VesselMNIST3D	*k*-FCV	86.94	90.23	32.87	91.69	91.83	75.47	86.65	90.05	59.33
	WT-CV	86.21	89.70	25.04	93.22	92.89	59.73	89.72	89.70	47.69
	DWT-CV	**90.45**	**91.18**	25.34	**94.13**	**93.11**	55.91	**90.09**	**90.11**	45.68
	Official	86.65	90.58	7.43	92.02	91.10	15.06	89.42	89.97	15.86

strategy, the network trained using DWT-CV not only performs well, but also requires less time for training. For example, the benchmark ResNet-18 trained with DWT-CV achieves the performance with an AUC of 91.17% and an accuracy of 91.25% on NoduleMNIST3D dataset, which exceeds the performance of the network trained with *k*-FCV and saves 27.55% training time. For the WT-CV strategy, the benchmark networks trained with WT-CV performs well on some datasets such as BreastMNIST and DermaMINIT, but is not always stable on the other datasets due to data partition bias. Extensive experimental results illustrate that our proposed DWT-CV strategy can alleviate this problem to some extent and make the model more generalized.

3.4.2. Comparison with the fixed data partition

To illustrate the superiority, we further compare our proposed DWT-CV strategy with the fixed data partition strategy for model selection. The official data partition strategy is to train the model on the training data and select the model on the fixed validation data. The official data distribution of training and validation data can be viewed from the website (https://medmnist.com/). We retrain the benchmark models using the official provided data partition. The results are shown in the last row of each dataset row in Table 3. The comparison results show that the official strategy spends less training time on model selection, but the performance of the network using the strategy is inferior to that using our proposed DWT-CV strategy in general. Therefore, the DWT-CV training strategy is a better method to select the model with stronger prediction performance.

3.4.3. Comparison with state-of-the-art methods

To further demonstrate the superiority, we compare with benchmark methods on various datasets. The benchmark methods such as ResNet-18 and ResNet-50 trained with the official training strategy are directly used for Ref. [23]. Furthermore, we also compare two AutoML methods including Auto-sklearn [43] and Auto-keras [44]. They are representative of the open source AutoML tools used for statistical machine learning and deep learning, respectively. The comparison results of these benchmark methods are derived from the literature [23] and are directly listed in Table 4. The last three columns in Table 4 present the results of our benchmark methods trained with the proposed DWT-CV strategy. The comparison results show that our benchmark method such as ResNet-18 has better prediction performance than the official ResNet-18 on the PneumoniaMNIST, AdrenalMNIST3D, FractureMNIST3D, and VesselMNIST3D datasets. It is worth mentioning that our adopted benchmark ResNet-34 achieves better performance compared with other benchmark methods on most datasets except for BreastMNIST and OrganMNIST3D. Although the AutoML methods achieve better accuracy on DermaMINIT and NoduleMNIST3D, our benchmarks are also competitive in terms of AUC.

3.5. Assessment on drug–target prediction

To further verify the effectiveness of this proposed training strategy in biomedical data prediction, we also evaluate the proposed strategy on DTA and DTI prediction. We report the average MSE and AUC of the benchmark DeepDTA and DeepConV-DTI

Table 4
Comparison performance with benchmark methods on various datasets for biomedical image classification.

Dataset	Metrics	ResNet-18 (official) [23]	ResNet-50 (official) [23]	Auto-sklearn [43]	Auto-keras [44]	**ResNet-18 (Our study)**	**ResNet-34 (Our study)**	**ResNet-50 (Our study)**
PneumoniaMNIST	AUC (%)	94.40	94.80	94.20	94.70	**95.69**	95.20	94.96
	ACC (%)	85.40	85.40	85.50	87.80	**88.67**	85.04	85.74
BreastMNIST	AUC (%)	**90.10**	85.70	83.60	87.10	89.10	88.52	89.77
	ACC (%)	**86.30**	81.20	80.30	83.10	85.68	84.40	84.83
DermaMNIST	AUC (%)	91.70	91.30	90.20	91.50	91.47	**91.67**	90.86
	ACC (%)	73.50	73.50	71.90	**74.90**	74.48	74.75	74.00
OrganMNIST3D	AUC (%)	**99.60**	99.40	97.70	97.90	99.44	99.49	99.44
	ACC (%)	90.70	88.30	81.40	80.40	**91.20**	90.05	90.22
NoduleMNIST3D	AUC (%)	91.50	90.20	87.20	84.70	91.17	**94.11**	89.23
	ACC (%)	90.80	91.00	**92.60**	90.20	91.25	91.51	91.76
AdrenalMNIST3D	AUC (%)	82.70	82.80	82.80	80.40	86.63	**86.88**	85.03
	ACC (%)	72.10	74.50	80.20	70.50	81.20	81.54	**82.44**
FractureMNIST3D	AUC (%)	71.20	72.50	62.80	64.20	72.32	**75.11**	69.20
	ACC (%)	50.80	49.40	45.30	45.80	53.06	**56.53**	51.94
VesselMNIST3D	AUC (%)	87.40	90.70	91.00	77.30	90.45	**94.13**	90.09
	ACC (%)	87.70	91.80	91.50	89.40	91.18	**93.11**	90.11

Table 5
Performance on drug–target relationship datasets.

Dataset	Training strategy	DeepDTA		DeepConV-DTI	
		MSE (%)	Time (h)	AUC (%)	Time (h)
Davis	*k*-FCV	24.27	1.37	–	–
	WT-CV	23.07	0.49	–	–
	DWT-CV	**22.76**	0.51	–	–
DrugBank	*k*-FCV	–	–	87.61	1.76
	WT-CV	–	–	89.36	0.73
	DWT-CV	–	–	**90.17**	0.43

Table 6
Performance on stroke lesion segmentation datasets.

Dataset	Training strategy	Dice	Time (h)
ISLES2018	*k*-FCV	0.42	12.21
	WT-CV	0.42	9.05
	DWT-CV	**0.43**	9.33
AISD	*k*-FCV	0.43	20.72
	WT-CV	0.49	14.71
	DWT-CV	**0.50**	16.90

methods repeated for five independent experiments, and the results are shown in Table 5. The DeepDTA trained with the DWT-CV strategy improves 1.51% MSE and saves 62.77% training time on the Davis dataset compared with the method trained with the traditional *k*-FCV. On the DrugBank dataset, the DeepConv-DTI trained with the DWT-CV strategy improves 2.56% AUC and saves 75.57% training time compared with the method trained with the traditional *k*-FCV. Therefore, the proposed training strategy can play the same role in computational drug discovery.

3.6. Assessment on stroke lesion segmentation

The effectiveness of the proposed strategy also is evaluated on the stroke lesion segmentation with two datasets. The comparison results are shown in Table 6. From the these results, we can see that the CoTr model trained with the DWT-CV strategy achieves a better segmentation performance on the ISLES2018 dataset, with a Dice of 0.43 and training time of 9.33 h, compared to the method trained using traditional *k*-FCV criterion. Similarly, the CoTr trained with DWT-CV strategy achieves a Dice of 0.50 on the AISD dataset, which improves by 7% compared to the method trained with the traditional *k*-FCV criterion. Therefore, our proposed DWT-CV strategy can also get the performance gain in medical image segmentation.

4. Discussion

Model selection is very important in deep learning, especially in the application of biomedical data with a limited number of samples. Although *k*-FCV is widely adopted as a model selection criterion, the model performance generalizes poorly on new unseen data sets, especially in the case of the limited available data samples. In this study, we propose a dense weight transfer-based cross-validation strategy for model selection of deep learning algorithms. Compared with the traditional *k*-FCV, the proposed DWT-CV strategy can not only improve the generalization of the model but also reduce training time through weight transfer learning and making full use of each partitioned data.

In this study, we find that the number of folds *k* in the cross-validation data partition has some impact on model selection and its performance. If *k* is too large or small, it may result in model selection bias on the partitioned validation data. To investigate this effect, we validate the proposed DWT-CV using the benchmark ResNet-18 on 8 biomedical image datasets with different *k*, where *k* is set to 3, 5, and 10. To comprehensively compare the performance of the DWT-CV with various *k*, we also report the average AUC, average ACC, and average training time of each *k* over all datasets. The comparison results are shown in Table 7. From these results, we can see that the performance of the model performs better in general when *k* is equal to 5. The *k* is too large or small, the overall performance will be affected. In addition, the greater the *k*, the more training time will be spent. Therefore, to make a tradeoff between the performance and training time, we set $k = 5$ for all datasets used in this study.

We have demonstrated that the effectiveness of the proposed DWT-CV strategy on multiple benchmark methods with various biomedical image datasets. Compared with the WT-CV strategy, the DWT-CV strategy makes the deep learning model more generalized on the unseen independent testing data. Although the models trained with WT-CV strategy also perform better than those trained with the traditional *k*-FCV on some datasets such as BreastMNIST, they are not stable on other datasets. This is because the WT-CV is susceptible to data partitioning bias. However, this bias is mitigated by the proposed DWT-CV using a dense weight aggregation mechanism. Furthermore, we also validate the effectiveness of DWT-CV strategy on the drug–target relationship datasets (See Table 5) and stroke lesion segmentation datasets (See Table 6).

Table 7
Performance of the ResNet for three backbone networks on 8 datasets with different k.

Dataset	$k = 3$			$k = 5$			$k = 10$		
	AUC (%)	ACC (%)	Time (min)	AUC (%)	ACC (%)	Time (min)	AUC (%)	ACC (%)	Time (min)
PneumoniaMNIST	95.51	85.90	1.23	**95.69**	**88.67**	2.54	92.92	83.06	4.96
BreastMNIST	85.55	83.76	0.31	**89.10**	**85.68**	0.56	88.05	83.12	1.21
DermaMNIST	91.22	74.26	2.09	**91.47**	**74.48**	3.35	90.85	74.11	6.13
OrganMNIST3D	99.29	89.46	14.80	99.44	91.20	25.54	**99.49**	**91.32**	56.39
NoduleMNIST3D	**94.68**	**91.82**	12.63	91.17	91.25	22.11	91.85	91.59	47.35
AdrenalMNIST3D	83.86	80.42	10.65	**86.63**	**81.20**	19.72	80.60	80.98	42.75
FractureMNIST3D	70.15	51.94	8.79	**72.32**	**53.06**	15.68	68.14	50.14	34.04
VesselMNIST3D	81.61	89.01	15.07	90.45	**91.18**	25.34	**91.49**	91.10	54.87
Average	87.73	80.82	8.20	**89.53**	**82.09**	14.35	87.92	80.68	30.96

Although our proposed DWT-CV strategy can make the model more generalized compared to the traditional cross-validation approach, there are some limitations in our study. First, the data used in this study are a public standard dataset, which the data shift problem may not be very serious. Future work will focus on validating the effectiveness of the proposed method on multi-institutional datasets. Second, the performance of the deep learning algorithm trained with the proposed strategy is still subject to the number of data partition folds. Knowledge distillation technology is applied to achieve significant performance gain in federated learning [45]. The next work will be devoted to the research of knowledge distillation to improve the robustness of deep learning models.

5. Conclusion

In conclusion, we propose a dense weight transfer-based cross validation strategy for model selection in biomedical data prediction. The proposed strategy can not only make the deep learning model more generalized but also reduce the training time compared with the traditional cross validation criterion. Comprehensive experiments are conducted on twelve datasets with different domains by using multiple benchmark methods to verify the effectiveness. It is believed that our proposed method can be easily generalized to various intelligent tasks, which is an alternative to the traditional cross-validation criterion for model selection.

CRediT authorship contribution statement

Jianhong Cheng: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Hulin Kuang:** Conceptualization, Writing – review & edition, Supervision, Project administration. **Qichang Zhao:** Methodology, Software, Validation, Formal analysis, Data curation. **Yahui Wang:** Methodology, Software, Validation, Formal analysis, Data curation. **Lei Xu:** Software, Validation, Formal analysis, Data curation. **Jin Liu:** Conceptualization, Writing – review & edition, Supervision. **Jianxin Wang:** Conceptualization, Writing – review & edition, Supervision, Formal analysis, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China (No. 2021YFF1201200), the National Natural Science Foundation of China under Grant (No. 62102454, No. 62172444), the Hunan Provincial Science and Technology Innovation Leading Plan, China (No. 2020GK2019), the Science and Technology Plan of Department of Natural Resources of Hunan Province, China (2021-17), and the High Performance Computing Center of Central South University.

References

[1] H. Luo, M. Li, M. Yang, F.-X. Wu, Y. Li, J. Wang, Biomedical data and computational models for drug repositioning: a comprehensive review, Brief. Bioinform. 22 (2) (2021) 1604–1619.
[2] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.
[3] J. Cheng, M. Gao, J. Liu, H. Yue, H. Kuang, J. Liu, J. Wang, Multimodal disentangled variational autoencoder with game theoretic interpretability for glioma grading, IEEE J. Biomed. Health Inf. 26 (2) (2022) 673–684.
[4] D. Allan, The relationship between variable selection and prediction, Technometrics 16 (1974) 125–127.
[5] G.C. Cawley, N.L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (2010) 2079–2107.
[6] J. Zhang, Y. Xie, Q. Wu, Y. Xia, Medical image classification using synergic deep learning, Med. Image Anal. 54 (2019) 10–19.
[7] J. Cheng, J. Liu, H. Kuang, J. Wang, A fully automated multimodal MRI-based multi-task learning for glioma segmentation and IDH genotyping, IEEE Trans. Med. Imaging (2022) 1, http://dx.doi.org/10.1109/TMI.2022.3142321.
[8] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, I.B. Ayed, HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation, IEEE Trans. Med. Imaging 38 (5) (2018) 1116–1126.
[9] H. Öztürk, A. Özgür, E. Ozkirimli, DeepDTA: deep drug–target binding affinity prediction, Bioinformatics 34 (17) (2018) i821–i829.
[10] I. Lee, J. Keum, H. Nam, DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences, PLoS Comput. Biol. 15 (6) (2019) e1007129.
[11] L. Du, J. Zhang, F. Liu, H. Wang, L. Guo, J. Han, et al., Identifying associations among genomic, proteomic and imaging biomarkers via adaptive sparse multi-view canonical correlation analysis, Medical Image Analysis 70 (2021) 102003.
[12] J. Cheng, et al., Prediction of glioma grade using intratumoral and peritumoral radiomic features from multiparametric MRI images, IEEE/ACM Trans. Comput. Biol. Bioinform. (2020) http://dx.doi.org/10.1109/TCBB.2020.3033538.
[13] L. Du, F. Liu, K. Liu, X. Yao, S.L. Risacher, J. Han, A.J. Saykin, L. Shen, Associating multi-modal brain imaging phenotypes and genetic risk factors via a dirty multi-task learning method, IEEE transactions on medical imaging 39 (11) (2020) 3416–3428.
[14] J. Liu, D. Zeng, R. Guo, M. Lu, F.-X. Wu, J. Wang, Mmhge: detecting mild cognitive impairment based on multi-atlas multi-view hybrid graph convolutional networks and ensemble learning, Cluster Comput. 24 (1) (2021) 103–113.
[15] J. Cheng, W. Zhao, J. Liu, X. Xie, S. Wu, L. Liu, H. Yue, J. Li, J. Wang, J. Liu, Automated diagnosis of covid-19 using deep supervised autoencoder with multi-view features from CT images, IEEE/ACM Trans. Comput. Biol. Bioinform. (2021) http://dx.doi.org/10.1109/TCBB.2021.3102584.

[16] Y. Wang, J. Liu, Y. Xiang, J. Wang, Q. Chen, J. Chong, MAGE: automatic diagnosis of autism spectrum disorders using multi-atlas graph convolutional networks and ensemble learning, Neurocomputing 469 (2022) 346–353.
[17] J. Cheng, J. Sollee, C. Hsieh, H. Yue, N. Vandal, J. Shanahan, J.W. Choi, T.M.L. Tran, K. Halsey, F. Iheanacho, et al., COVID-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest X-rays and clinical data, Eur. Radiol. (2022) http://dx.doi.org/10.1007/s00330-022-08588-8.
[18] J. Cheng, J. Liu, L. Liu, Y. Pan, J. Wang, Multi-level glioma segmentation using 3D U-Net combined attention mechanism with atrous convolution, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 1031–1036.
[19] J.R. Cano, F. Herrera, M. Lozano, Evolutionary stratified training set selection for extracting classification rules with trade off precision-interpretability, Data Knowl. Eng. 60 (1) (2007) 90–108.
[20] X. Zeng, T.R. Martinez, Distribution-balanced stratified cross-validation for accuracy estimation, J. Exp. Theor. Artif. Intell. 12 (1) (2000) 1–12.
[21] M. Sugiyama, M. Krauledat, K.-R. Müller, Covariate shift adaptation by importance weighted cross validation, J. Mach. Learn. Res. 8 (5) (2007).
[22] J.G. Moreno-Torres, J.A. Sáez, F. Herrera, Study on the impact of partition-induced dataset shift on k-fold cross-validation, IEEE Trans. Neural Netw. Learn. Syst. 23 (8) (2012) 1304–1312.
[23] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification, 2021, arXiv preprint arXiv:2110.14795.
[24] J. Yang, R. Shi, B. Ni, MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis, in: IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 191–195.
[25] S. Deepak, P. Ameer, Brain tumor classification using deep CNN features via transfer learning, Comput. Biol. Med. 111 (2019) 103345.
[26] Y. Yang, L.-F. Yan, X. Zhang, Y. Han, H.-Y. Nan, Y.-C. Hu, B. Hu, S.-L. Yan, J. Zhang, D.-L. Cheng, et al., Glioma grading on conventional MR images: a deep learning study with transfer learning, Front. Neurosci. 12 (2018) 804.
[27] T. Iesmantas, R. Alzbutas, Convolutional capsule network for classification of breast cancer histology images, in: International Conference Image Analysis and Recognition, Springer, 2018, pp. 853–860.
[28] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.
[29] Y. Yuan, Z. Bar-Joseph, Deep learning for inferring gene relationships from single-cell expression data, Proc. Natl. Acad. Sci. 116 (52) (2019) 27151–27158.
[30] G. Campanella, M.G. Hanna, L. Geneslaw, A. Miraflor, V.W.K. Silva, K.J. Busam, E. Brogi, V.E. Reuter, D.S. Klimstra, T.J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature Med. 25 (8) (2019) 1301–1309.
[31] J. Cheng, J. Liu, L. Liu, Y. Pan, J. Wang, Multi-level glioma segmentation using 3D U-Net combined attention mechanism with atrous convolution, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 1031–1036.
[32] J. Cheng, Q. Zhao, L. Xu, J. Liu, A new deep learning training scheme: Application to biomedical data, in: Bioinformatics Research and Applications, Springer International Publishing, Cham, 2021, pp. 451–459.
[33] M.I. Davis, J.P. Hunt, S. Herrgard, P. Ciceri, L.M. Wodicka, G. Pallares, M.D. Hocker, D.K. Treiber, P.P. Zarrinkar, Comprehensive analysis of kinase inhibitor selectivity, Nature Biotechnol. 29 (11) (2011) 1046–1051.
[34] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 34 (suppl_1) (2006) D668–D672.
[35] O. Maier, B.H. Menze, J. von der Gablentz, L. Häni, M.P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, et al., ISLES 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI, Med. Image Anal. 35 (2017) 250–269.
[36] K. Liang, K. Han, X. Li, X. Cheng, Y. Li, Y. Wang, Y. Yu, Symmetry-enhanced attention network for acute ischemic infarct segmentation with non-contrast CT images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 432–441.
[37] H. Kuang, B.K. Menon, S.I. Sohn, W. Qiu, EIS-Net: Segmenting early infarct and scoring ASPECTS simultaneously on non-contrast CT of patients with acute ischemic stroke, Med. Image Anal. 70 (2021) 101984.
[38] M. Attene, A lightweight approach to repairing digitized polygon meshes, Vis. Comput. 26 (11) (2010) 1393–1406.
[39] M. Kistler, S. Bonaretti, M. Pfahrer, R. Niklaus, P. Büchler, The virtual skeleton database: an open access repository for biomedical research and collaboration, J. Med. Internet Res. 15 (11) (2013) e245.
[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[41] Y. Xie, J. Zhang, C. Shen, Y. Xia, CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention, 2021, pp. 171–180.
[42] J. Zhang, Y. Xie, P. Zhang, H. Chen, Y. Xia, C. Shen, Light-weight hybrid convolutional network for liver tumor segmentation, in: IJCAI, Vol. 19, 2019, pp. 4271–4277.
[43] M. Feurer, A. Klein, K. Eggensperger, J.T. Springenberg, M. Blum, F. Hutter, Auto-sklearn: efficient and robust automated machine learning, in: Automated Machine Learning, Springer, Cham, 2019, pp. 113–134.
[44] H. Jin, Q. Song, X. Hu, Auto-keras: An efficient neural architecture search system, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1946–1956.
[45] J. Zhang, S. Guo, X. Ma, H. Wang, W. Xu, F. Wu, Parameterized knowledge transfer for personalized federated learning, Adv. Neural Inf. Process. Syst. 34 (2021) 1–15.

Jianhong Cheng received the B.S. degree from Liaoning Technical University, Fuxin, China, in 2014, the M.S. degree in software engineering from Central South University, Changsha, China, in 2017. He is currently a Ph.D. candidate in School of Computer Science and Engineering, Central South University, Changsha, China. His research interests include machine learning, deep learning and medical image analysis.

Hulin Kuang received the Ph.D. degree from the City University of Hong Kong in 2016 and was a Postdoctoral Fellow with the Department of Clinical Neurosciences, University of Calgary, Canada from 2017 to 2020. Now he is a distinguished associate professor with School of Computer Science and Engineering, Central South University, China. His research interests include medical image analysis, deep learning, and intelligent transportation systems.

Qichang Zhao is a Ph.D. candidate in the Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering at Central South University, Hunan, China. His current research interests include machine learning, deep learning, and bioinformatics.

Yahui Wang received the B.S. degree in College of Mathematics and Applied Mathematics, Hunan University of Science and Technology in 2020. He is currently a M.S. candidate in School of Computer Technology, Central South University, Changsha, Hunan, P.R. China. His research interests include deep learning and medical image analysis.

Lei Xu received the B.S degree in Digital Media Technology from the Information Engineering School of Nanchang University, Nanchang, China, in 2019. He is currently working toward the Master's degree in computer technology from the School of Computer Science and Engineering, Central South University, Changsha, China. His current research interests include medical image analysis, bioinformatics, adventitious respiratory sound classification and deep learning.

Jin Liu received the PhD degree in computer science from Central South University, China, in 2017. He is currently an associate professor at the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include medical image analysis, machine learning and bioinformatics.

Jianxin Wang received his B.S. and M.S. degree in Computer Science and Application from Central South University of Technology, P. R. China, and his PhD degree in Computer Science and Technology from Central South University. Currently, he is the dean and a professor in School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. He is also a leader in Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics and computer network. He has published more than 200 papers in various International journals and refereed conferences. He is a senior member of the IEEE. He has been on numerous program committees and NSFC review panels, and served as editors for several journals such as IEEE/ACM Trans. Computational Biology and Bioinformatics (TCBB), International Journal of Bioinformatics Research and Applications, Current Bioinformatics, Current Protein & Peptide Science, Protein & Peptide Letters.