Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5589872, 11 pages
https://doi.org/10.1155/2021/5589872

Research Article

Attribute-Associated Neuron Modeling and Missing Value Imputation for Incomplete Data

Xiaochen Lai,[1,2] Jinchong Zhu,[1] Liyong Zhang,[3,4] Zheng Zhang,[5] and Wei Lu[3,4]

[1]*School of Software, Dalian University of Technology, Dalian 116620, China*
[2]*Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian 116620, China*
[3]*School of Control Science and Engineering, Dalian University of Technology, Dalian 116624, China*
[4]*Professional Technology Innovation Center of Distributed Control for Industrial Equipment of Liaoning Province, Dalian 116024, China*
[5]*International School of Information Science & Engineering, Dalian University of Technology, Dalian 116620, China*

Correspondence should be addressed to Liyong Zhang; zhly@dlut.edu.cn

Received 11 January 2021; Revised 9 March 2021; Accepted 9 April 2021; Published 29 April 2021

Academic Editor: Nawab Muhammad Faseeh Qureshi

The imputation of missing values is an important research content in incomplete data analysis. Based on the auto associative neural network (AANN), this paper conducts regression modeling for incomplete data and imputes missing values. Since the AANN can estimate missing values in multiple missingness patterns efficiently, we introduce incomplete records into the modeling process and propose an attribute cross fitting model (ACFM) based on AANN. ACFM reconstructs the path of data transmission between output and input neurons and optimizes the model parameters by training errors of existing data, thereby improving its own ability to fit relations between attributes of incomplete data. Besides, for the problem of incomplete model input, this paper proposes a model training scheme, which sets missing values as variables and makes missing value variables update with model parameters iteratively. The method of local learning and global approximation increases the precision of model fitting and the imputation accuracy of missing values. Finally, experiments based on several datasets verify the effectiveness of the proposed method.

1. Introduction

The interference of various factors in process of data collection, transmission and storage, etc. may cause data loss in different degrees. The incompleteness of data that leads to most of computational intelligence technologies cannot be applied directly [1]. In the cases where incomplete records cannot be simply deleted, an effective method is needed to impute missing values.

At present, researchers have proposed a variety of imputation methods. Mean imputation method imputes corresponding missing values with mean values of existing attributes [2]. The hot deck method finds the record most similar to the incomplete record in database and then imputes data with values of this record [3]. The K-nearest neighbors (KNN) imputation method takes the weighted average of K records closest to the incomplete record to impute missing values [4]. Additionally, model-based methods are usually an effective way to improve the accuracy of imputation. For example, the expectation-maximization (EM) method alternately performs the expectation step and the maximization step and iteratively updates model parameters and missing values until convergence [5]. The multiple imputation method obtains m values through one or more models and comprehensively processes the m results to impute missing values [6]. The imputation method based on linear model imputes missing values by modeling the linear relation between attributes [7]. It assumes that there is a linear correlation of the data, but the relation is complex and unknown in real data and often reflects nonlinear features.

The neural network is flexible in construction. In theory, a neural network with nonlinear activation function can approximate complex nonlinear relations [8]. The imputation

model based on neural networks can mine complex association relations within attributes of incomplete data. The imputation method based on the neural network usually uses complete records to train the network, then inputs prefilling incomplete records into the network and uses the output of network to impute missing values [9]. Sharpe and Solly [10] constructed a multilayer perceptron (MLP) for each missingness pattern, which is used to fit the regression relation between missing attributes and existing attributes. However, the number of constructed models is large, and the training is more time-consuming in the case of multiple missingness patterns. Ankaiah and Ravi [11] proposed an improved MLP imputation method, which takes each missing attribute as output and the remaining attributes as input to construct a network of single objective predictive. The number of models constructed by this method is equal to the number of missing attributes. Although the MLP imputation model can fit the regression relation between data attributes, it comes at the expense of model training time.

The auto associative neural network (AANN) is a type of network with the same number of nodes in the output layer and input layer. It is only necessary to build one model to impute incomplete data in all missingness patterns [12]. Marwala et al. [13] proposed an imputation method combining AANN and genetic algorithm (GA) and then applied it to two real datasets [14, 15]. This method takes the cost function of AANN as the fitness function of the genetic algorithm and uses the genetic algorithm to impute missing values. Based on the framework proposed by Marwala, Nelwamondo et al. adopted principal component analysis to select a reasonable number of nodes in the hidden layer [16] and reduce the dimension of data [17]. Ssali and Marwala [18] used the interval of continuous attribute divided by decision tree as data boundary, which further improved the imputation accuracy. In addition to the combination of AANN and GA, Ravi and Krishna [19] proposed four improved imputation models based on AANN, which are general regression auto associative neural network (GRAANN), particle swarm optimization based auto associative neural network (PSOAANN), auto associative wavelet neural network (AAWNN), and radial basis function auto associative neural network (RBFAANN). Among these models, GRAANN performs better than MLP and other three models in most datasets, and only needs one iteration to impute missing values. Gautam and Ravi proposed two imputation models based on AANN, which are auto associative extreme learning machine [20] and counter propagation auto associative neural network [21]. The experimental results show that the combination of local learning and global approximation can get better imputation results.

The above method only takes complete records to train the model, which avoids the problem of missing values during training. However, missing values in incomplete records will lead to incomplete model input in the imputation stage. Since the MLP imputation method constructs a specific model for each missingness pattern by taking incomplete attributes as output and complete attributes as input, it can directly input each incomplete record into the subnet of the corresponding missingness pattern. However, the AANN imputation method usually needs a prefilling method to deal with missing values during imputation. For instance, Ravi and Krishna [19] used averages to prefill missing values. Nishanth and Ravi [22] adopted K-means and K-medoids methods to prefill missing values. Gautam and Ravi [21] used the nearest neighbor method based on grey distance to prefill missing values.

The quantity of complete records is small when the missing rate in dataset is high. If only complete records are used to train the network, a large amount of information in incomplete records will be lost, and fewer records sometimes make the model unbuildable. Therefore, Silva-Ramírez et al. [23] prefilled missing values with a fixed value and then trained the network by all records. García-Laencina et al. [24–28] proposed a multitask network that uses zero to initialize missing values and allows incomplete records to participate in model training. Although the method of prefilling incomplete records with fixed values can make them participate in model training, the prefilling values have an estimation error. If the model is trained directly with prefilling data, the accuracy of the final model will be affected by the estimation error. In addition, Yoon et al. [29] proposed an imputation method base on Generative Adversarial Nets to generate data with generator. The network architecture can also try to use the inception architecture [30] in edge computing [31].

As mentioned above, the imputation method based on AANN can improve the training efficiency compared with MLP while solving multiple missingness patterns. Consequently, this paper conducts regression modeling for the attributes of incomplete data based on AANN architecture. By redesigning the data transmission structure of AANN, the representation of regression relations between data attributes is enhanced. Moreover, aiming at the problem of incomplete model input, this paper proposes a model training scheme that takes missing values as variables and makes the missing value variables update iteratively along with model parameters during model training. The improved model and training scheme make full use of the existing data in incomplete dataset and reduce the estimation error of missing value variables gradually during model training and increases the accuracy of imputation through local learning and global approximation.

The rest of this paper is organized as follows. Section 2 introduces MLP and AANN imputation models. Section 3 proposes ACFM based on AANN and a model training scheme named UMVDT. Section 4 analyses the imputation performance of ACFM and UMVDT. And the full text is summarized in Section 5.

2. MLP and AANN Imputation Models

MLP is a feed forward artificial neural network composed of input layer, output layer, and several hidden layers. When applying the MLP method to impute missing values, an MLP imputation network needs to be constructed for each missingness pattern. Figure 1 is an incomplete dataset with several missingness patterns, different positions and the number of missing values in the sample, and different deletion modes. For an incomplete data set that is missing at

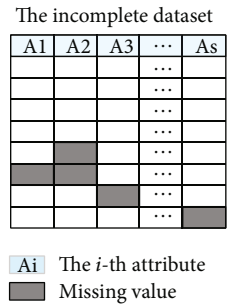

Figure 1: The incomplete dataset with multiple missing patterns.

random, the higher the missing rate, the more missing patterns. And its imputation networks are shown in Figure 2, where $x_i = [x_{i1}, x_{i2}, \cdots, x_{\text{is}}]^T$ represents the i-th record, $y_i = [y_{i1}, y_{i2}, \cdots, y_{\text{is}}]^T$ represents the network output for the i-th record, and s represents the dimension of attribute. If p_t represents indices of missing attributes in the t-th missingness pattern, the cost function of the model is

$$E_k = \frac{1}{2} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}_C} \sum_{j \in P_t} \left(f_{ij} \left(\sum_{k \notin p_t} w_{jk} \cdot x_{ik} \right) - x_{ij} \right)^2, \quad (1)$$

where X_C represents the complete records, $f_{ij}(\cdot)$ represents the nonlinear mapping of the model, and w_{jk} represents the weight of the model.

AANN requires that the number of nodes in the output layer is equal to that in the input layer. In order to prevent model overfitting, the number of nodes in the hidden layer is usually set to be less than that in the input layer. As shown in Figure 3, the imputation method based on AANN can

fill incomplete data under all missingness patterns through one structure. Generally, the model is trained by complete record subset, and the incomplete record subset is reconstructed after prefilled to impute the corresponding missing values. The cost function can be expressed as

$$\mathrm{E} = \frac{1}{2} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}_C} \sum_{j=1}^{s} \left(f_{ij} \left(\sum_{k=1}^{s} w_{jk} \cdot x_{ik} \right) - x_{ij} \right)^2. \quad (2)$$

It can be seen that each output value of AANN model is calculated by all input values. The output value is easier to learn the input value in the same position with model training, thus the quality of imputation values depends on a degree of the quality of pre-filling values in imputation stage. The output value of the MLP model is calculated by a regression network; so, AANN lacks clear regression relations to guide the model training and impute the missing value compared with the MLP model.

3. Proposed Architecture

3.1. Attribute CrossFitting Model. The AANN imputation model implements the imputation of multiple missingness patterns through one architecture, but it does not establish a clear regression relation between data attributes. In this paper, the regression relations between each attribute and rest attributes in incomplete dataset are expressed on one architecture by redesigning the cost function of the model

$$\mathrm{E} = \frac{1}{2} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} \sum_{j=1}^{s} \left(f_{ij} \left(\sum_{k=1, k \neq j}^{s} w_{jk} \cdot x_{ik} \right) - x_{ij} \right)^2, \quad (3)$$

where $\boldsymbol{x}$ represents an incomplete dataset. It can be seen from equation (3) that the j-th output value of the model is calculated from other input values except the j-th input value, which helps to establish a regression relation between each output value and remaining input values. Moreover, the output of the model is no longer dependent on the corresponding input value; thus, the effect of prefilling values is weakened during the imputation stage. In order to minimize the cost function, the network needs to fully learn the correlation between each output neuron and noncorresponding input neurons. Therefore, the cost function can effectively enhance the ability of mining internal association of attributes.

If the neural network is trained by incomplete records, the missing values need to be prefilled. However, there is an estimation error in prefilled values compared with original data. The model should limit the training error between prefilled data and its predicted data to optimize model parameters. This paper defines this error as missing value error. Hence, when training the network with an incomplete dataset, the cost function that the model needs to be optimized should be

$$\mathrm{E} = \frac{1}{2} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} \sum_{j \notin M_i} \left(f_{ij} \left(\sum_{k=1, k \neq j}^{s} w_{jk} \cdot x_{ik} \right) - x_{ij} \right)^2, \quad (4)$$

where M_i is the set of indexes for missing values in record x_i, and $j \notin M_i$ indicates that the missing value error is no longer used to optimize model parameters. The model constructed based on this cost function can fit regression relation between data attributes by one architecture, which is called attribute cross fitting model (ACFM) in this paper.

The data transmission process of output neurons of ACFM is shown in Figure 4 [32]. There is an incomplete record with two missing values $\hat{x}_{i1}$ and $\hat{x}_{i2}$ that input into ACFM. Because ACFM does not use missing value error to optimize model parameters, the output values y_{i1} and y_{i2} will not be calculated. The output value y_{i3} of ACFM is calculated by input values except x_{i3}. At the same time, the calculation of output values y_{i4} to y_{is} has similar processes. It can be seen that the calculation amount of ACFM is the same as that of AANN. In this article, the input data has been expanded by dimensionality times when it is implemented by programming. Then, perform forward calculation in a fully connected manner. Finally, the output is sliced, and the required value is taken out. Therefore, the parameter of ACFM in the experiment is the parameter of AANN multiplied by the number of attributes of the data.

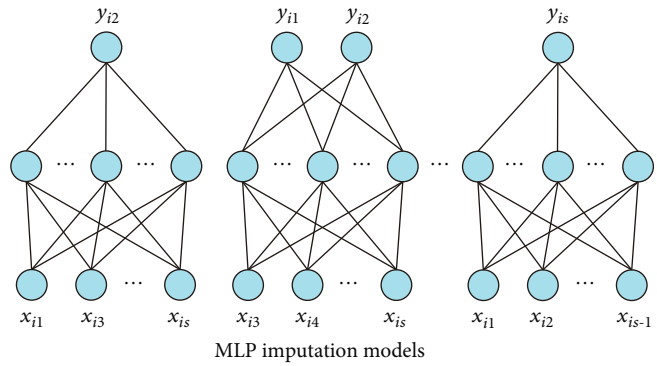

Figure 2: MLP imputation networks corresponding to each missing pattern.

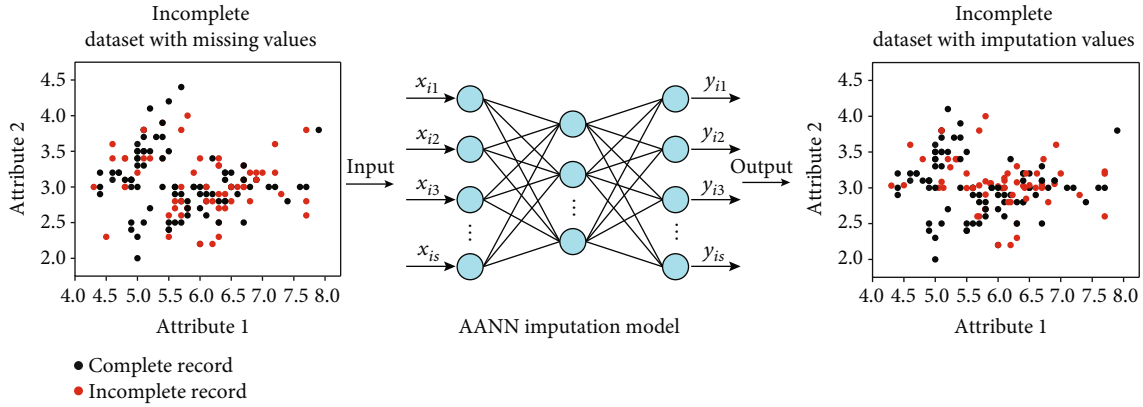

Figure 3: The diagram of AANN imputation.

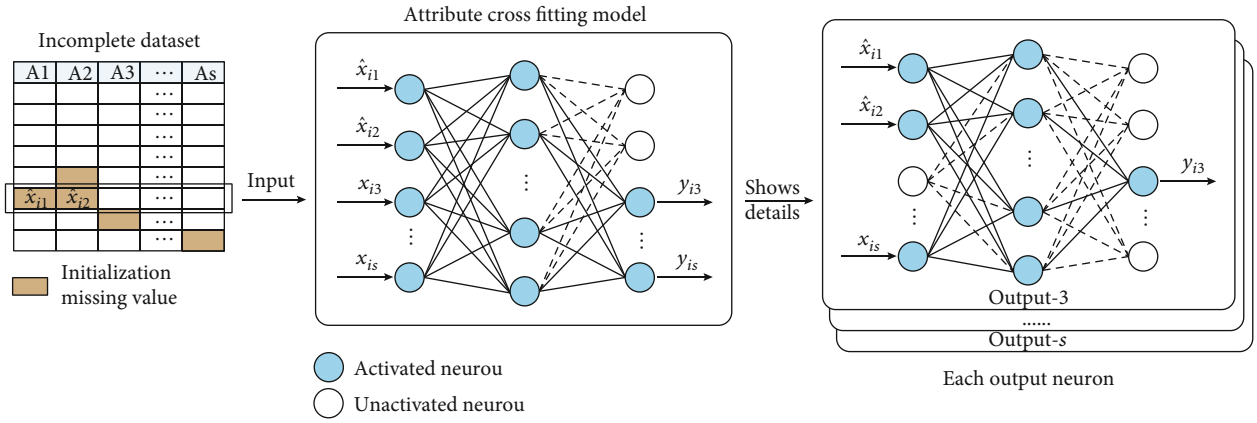

Figure 4: Schematic diagram of data transmission of ACFM, reproduced from Jinchong Zhu et al. 2020.

3.2. Updating Missing Values during Training. The prefilling missing values solve the problem of the incomplete model input, but the quality of prefilling values has an important impact on the quality of trained model when prefilling incomplete records are used to train the model directly. The prefilling values have an initial estimation error, which will reduce the accuracy of the model. Therefore, this paper proposes a model training scheme by treating missing values as variables and iteratively updating missing values during training process (UMVDT). UMVDT dynamically adjusts the values of missing and gradually reducing the estimation error of missing values, thus the missing values will meet the fitting relationship determined by existing data. As shown in Figure 5, UMVDT training scheme initializes the missing value variables in incomplete records and inputs incomplete records into ACFM for calculating the error $[e_{i1}, e_{i3},\cdots,e_{is}]$ between output and input values; then, it updates the missing value variables and the network parameters through the back propagation algorithm. The above process is repeated for all records until the model convergence. In

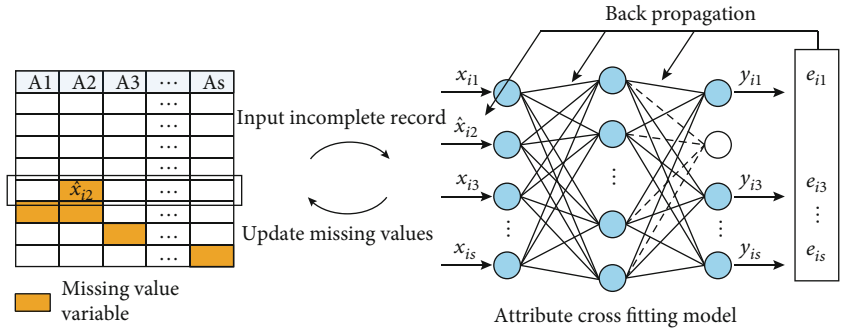

FIGURE 5: Schematic diagram of UMVDT training scheme. Reproduced from Jinchong Zhu et al. 2020.

the model based on UMVDT, the missing value variable is optimized by the regression structure within the incomplete data. The accuracy of the model will be improved with the deepening of the training, and the missing value predicted by the model will also be more accurate.

If the neurons in the input layer of ACFM are the first layer, and the output layer is $n+1$, w^l and b^l represent the weights and thresholds from layers l to $l+1$ $(1 \le l \le n)$. And each output neuron of the model is directly output after linear summation; so, it can be expressed as

$$y_{ij} = z_{ij}^{n+1} = b_j^{n+1} + \sum_{k=1}^{s_n} w_{jk}^n \cdot a_k^n, \tag{5}$$

where z_{ij}^{n+1} represents the linear summation of the j-th neuron in the layer $n+1$, s_n represents the number of neurons in layer n, and a_k^n represents the k-th output in layer n. Corresponding to each neuron j' in the output layer, the output of j-th neuron in each hidden layer can be expressed as

$$\begin{cases} \mathrm{a}_{ij}^l = g\left(z_{ij}^l\right) = g\left(b_j^{l-1} + \sum_{k=1}^{s_{l-1}} w_{jk}^{l-1} \cdot a_k^{l-1}\right), & 2 < l \le n \\ \mathrm{a}_{ij}^l = g\left(z_{ij}^l\right) = g\left(b_j^{l-1} + \sum_{k=1,k\neq j}^{'s_{l-1}} w_{jk}^{l-1} \cdot x_{ik}\right), & l = 2 \end{cases}, \tag{6}$$

where $g(\cdot)$ is the activation function. According to equation (4), the error between i -th record $\boldsymbol{x_i}$ and output $\boldsymbol{y_i}$ of the network is

$$\boldsymbol{e_i} = \frac{1}{2} \sum_{j=1,j\notin M_i}^{s_{n+1}} \left(y_{ij} - x_{ij}\right)^2. \tag{7}$$

If we define the intermediate variables δ_{ij}^{n+1} as

$$\begin{cases} \delta_{ij}^{n+1} = \dfrac{\partial e_i}{\partial z_{ij}^{n+1}} = \dfrac{\partial e_{ij}}{\partial z_{ij}^{n+1}} = \left(y_{ij} - x_{ij}\right), & \mathrm{j} \notin \mathrm{M_i} \\ \delta_{ij}^{n+1} = 0, & \mathrm{j} \in \mathrm{M_i} \end{cases} \tag{8}$$

where $j \notin M_i$ represents that the input value corresponding to the j-th predicted value is available, $j \in M_i$ represents that the input value corresponding to the j-th predicted value is missing, and thus the partial derivative is set to zero, and the corresponding model parameters are not optimized. When $2 \le l \le n$, δ_{ij}^l is

$$\delta_{ij}^l = \frac{\partial e_{ij}}{\partial z_{ij}^l} = \sum_{k=1}^{s_{l+1}} \delta_{ik}^{l+1} \cdot w_{jk}^l \cdot g'\left(z_{ij}^l\right), \tag{9}$$

and it can be concluded that the partial derivative of error e_i for the network parameter w_{jk}^l is

$$\frac{\partial \boldsymbol{e_i}}{\partial w_{jk}^l} = \frac{\partial \boldsymbol{e_i}}{\partial z_{ij}^{l+1}} \cdot \frac{\partial z_{ij}^{l+1}}{\partial w_{jk}^l} = \delta_{ij}^{l+1} \cdot a_k^l. \tag{10}$$

Similarly, the partial derivative of error e_i for the network parameter b_j^l is

$$\frac{\partial \boldsymbol{e_i}}{\partial b_j^l} = \frac{\partial \boldsymbol{e_i}}{\partial z_{ij}^{l+1}} = \delta_{ij}^{l+1}. \tag{11}$$

Assuming that the learning rate is η, and when the gradient descent method is used to optimize the model, the updating rule of the model parameters is

$$\begin{cases} w_{jk}^l = w_{jk}^l - \eta \cdot \dfrac{\partial \boldsymbol{e_i}}{\partial w_{jk}^l} \\ b_j^l = b_j^l - \eta \cdot \dfrac{\partial \boldsymbol{e_i}}{\partial b_j^l} \end{cases}. \tag{12}$$

Missing value variables are updated with the model parameters during model training. It can be deduced from equation (9) that the partial derivative of error e_i for the missing value variable $\hat{x}_{ik}$ $(k \in M_i)$ is

$$\frac{\partial \boldsymbol{e_i}}{\partial \hat{x}_{ik}} = \sum_{j=1}^{s_2} \delta_{ij}^2 \cdot w_{jk}^1, \tag{13}$$

INPUT: complete dataset D, missing rate, ACFM, learning rate η, maximum rounds T.
OUTPUT: the imputation error of D at specified missing rate.
Generate an incomplete dataset D' according to specified missing rate.
Initialize missing values as variables, model weights, and thresholds.
Set t =0, precision =1.
while t<T and precision<0.001 do.
  $t = t + 1$.
  for x in D':
    Input x into model and get output y.
    Calculate the error for updating the model parameters and missing value variables respectively.
  end for
  Reconstruct model output and predict missing values.
  Calculate the imputation error and precision.
end while
Output the imputation error.

Algorithm 1: The imputation based on ACFM and UMVDT.

and the updating rule of missing value variable $\widehat{x}_{ik}$ is

$$\widehat{x}_{ik} = \widehat{x}_{ik} - \eta \cdot \frac{\partial \boldsymbol{e_i}}{\partial \widehat{x}_{ik}}. \tag{14}$$

In summary, the imputation algorithm based on the ACFM model and UMVDT training scheme is described as follows:

Table 1: Description of datasets.

Datasets	Records	Attributes	Datasets	Records	Attributes
Blood	748	4	Iris	150	4
Buddymove	249	6	Seeds	210	7
Ecoli	336	7	Stock	252	12
Glass	214	10	Wine	178	13
Concrete	1030	9	Abalone	4177	7

4. Experiment

4.1. Datasets. In order to verify the imputation performance of proposed method, ten complete datasets obtained from the UCI database are used in our experiment, and the description of datasets is shown in Table 1. Among them, Stock is often used for clustering tasks, Concrete is often used for regression tasks, and the remaining data sets can be used for classification tasks. Most of these data are numeric, and some of them are nonnumeric in the ID column, which was deleted in the experiment. Additional information can refer to data sets UCI official website. For the sake of forming incomplete datasets, partial data are deleted randomly according to specified deletion rates which are set as 5%, 10%, 15%, 20%, 25%, and 30% and ensure that each incomplete record has at least one attribute value, which can be used for normal training.

4.2. Experimental Design. Six imputation methods based on MLP, AANN, and ACFM are realized. The method based on AANN and ACFM realizes the training by traditional training scheme and UMVDT training scheme. Traditional training scheme only uses the mean value to prefill missing value, and does not update missing values. To verify the effect of missing value error on imputation accuracy of the model, this paper uses equation (3) with missing value error and equation (4) without missing value error as the cost function, respectively. The specific methods are described as follows:

(1) The imputation method based on the MLP model and traditional training scheme (MLP-I): taking missing attributes as output and other attributes as input, multiple networks of single objective predictive are established based on MLP. These models are trained with complete records during the training stage. In the imputation stage, the incomplete records are prefilling with the mean method, and missing values are imputed with the reconstructed model output

(2) The imputation method based on the AANN model and traditional training scheme (AANN-I): the imputation process is same as MLP-I, but the architecture is AANN

(3) The imputation method based on the ACFM model where missing value error is used to optimize model parameters (ACFM-MEI): equation (3) is used as the cost function of ACFM. The incomplete records are prefilling with the mean method, and then all records are used to train the model. Finally, the reconstructed model output is used to impute missing values

(4) The imputation method based on the ACFM model where missing value error is not used to optimize model parameters (ACFM-I): equation (4) is used as the cost function of ACFM, and the process is same as ACFM-MEI

(5) The imputation method based on the AANN model and UMVDT training scheme (AANN-UMVDT): the mean value is used to initialize missing value variables. After that, the method uses all data to train the

Table 2: The MAPE values of first five datasets.

Datasets	Methods	Missing rates					
		5%	10%	15%	20%	25%	30%
Blood	MLP-I	1.113	1.122	0.872	0.68	0.861	0.985
	AANN-I	0.983	1.087	1.274	1.114	1.188	1.212
	AANN-UMVDT	0.941	0.968	1.005	0.974	1.026	1.124
	ACFM-MEI	0.513	0.575	0.683	0.728	0.787	0.854
	ACFM-I	0.488	0.537	0.62	0.671	0.728	0.764
	ACFM-UMVDT	0.449	0.51	0.599	0.633	0.754	0.8
Buddymove	MLP-I	0.24	0.151	0.166	0.199	0.213	0.237
	AANN-I	0.202	0.234	0.243	0.272	0.261	0.292
	AANN-UMVDT	0.14	0.147	0.159	0.163	0.167	0.177
	ACFM-MEI	0.122	0.149	0.172	0.199	0.2	0.222
	ACFM-I	0.122	0.142	0.164	0.184	0.186	0.197
	ACFM-UMVDT	0.105	0.115	0.131	0.15	0.15	0.176
Ecoli	MLP-I	0.472	0.231	0.233	0.236	0.274	0.285
	AANN-I	0.192	0.273	0.281	0.259	0.275	0.304
	AANN-UMVDT	0.185	0.247	0.261	0.238	0.257	0.276
	ACFM-MEI	0.159	0.226	0.253	0.237	0.259	0.283
	ACFM-I	0.157	0.22	0.241	0.226	0.247	0.269
	ACFM-UMVDT	0.155	0.212	0.248	0.232	0.241	0.269
Glass	MLP-I	0.287	0.351	0.298	0.343	0.423	0.45
	AANN-I	0.376	0.405	0.407	0.363	0.417	0.439
	AANN-UMVDT	0.407	0.429	0.403	0.356	0.414	0.419
	ACFM-MEI	0.311	0.342	0.344	0.303	0.357	0.354
	ACFM-I	0.29	0.346	0.338	0.314	0.35	0.35
	ACFM-UMVDT	0.286	0.327	0.346	0.314	0.353	0.345
Iris	MLP-I	0.157	0.15	0.237	0.234	0.272	0.335
	AANN-I	0.298	0.358	0.376	0.386	0.401	0.455
	AANN-UMVDT	0.15	0.158	0.19	0.188	0.189	0.234
	ACFM-MEI	0.17	0.186	0.25	0.279	0.297	0.367
	ACFM-I	0.153	0.139	0.219	0.217	0.237	0.272
	ACFM-UMVDT	0.139	0.128	0.167	0.173	0.186	0.234

model, dynamically updates missing value variables during model training, and reconstructs the model output to impute missing values

(6) The imputation method based on the ACFM model and UMVDT training scheme (ACFM-UMVDT): the process is same as AANN-UMVDT, but the architecture is ACFM

All models are optimized based on the gradient descent method with momentum. The learning rate is set as 0.2, and momentum is set as 0.9. All methods were repeated ten times at each missing rate, and the average value of ten imputation errors was taken as experimental results. Imputation error is evaluated by mean absolute percentage error (MAPE):

$$\text{MAPE} = \frac{1}{\sum_{i=1}^{n} |M_i|} \sum_{\boldsymbol{x_i} \in X_I} \sum_{j \in M_i} \left| \frac{y_{ij} - x_{ij}}{x_{ij}} \right|, \tag{15}$$

where X_I represents incomplete records subset, and $\sum_{i=1}^{n} |M_i|$ represents the number of missing values.

4.3. Experimental Results. The experimental results are shown in Tables 2 and 3.

4.4. Experimental Discussion. The impact of architecture on imputation results: by observing the MAPE values of ACFM-I, AANN-I, and MLP-I in Tables 2 and 3, we can see that the results of ACFM-I are slightly worse than those of MLP-I in four cases, which is Ecoli at the missing rate of 15%, Glass at the missing rates of 5% and 15%, and Stock at the missing rate of 5%. In addition to the above, all the results of ACFM-I are better than MLP-I and AANN-I. Besides, there are forty-three results of MLP-I superior to the AANN-I among sixty imputation results. This result shows that MLP can more accurately characterize the regression relation within dataset than AANN, thereby obtaining higher imputation accuracy. ACFM increases the ability to

Table 3: The MAPE values of last five datasets.

Datasets	Methods	Missing rates					
		5%	10%	15%	20%	25%	30%
Seeds	MLP-I	0.071	0.093	0.104	0.114	0.096	0.151
	AANN-I	0.083	0.096	0.095	0.109	0.097	0.122
	AANN-UMVDT	0.067	0.077	0.072	0.084	0.076	0.085
	ACFM-MEI	0.07	0.08	0.08	0.097	0.088	0.099
	ACFM-I	0.067	0.077	0.076	0.09	0.083	0.09
	ACFM-UMVDT	0.062	0.068	0.067	0.081	0.076	0.088
Stock	MLP-I	0.109	0.145	0.171	0.227	0.279	0.341
	AANN-I	0.181	0.203	0.22	0.236	0.268	0.32
	AANN-UMVDT	0.177	0.185	0.185	0.194	0.19	0.201
	ACFM-MEI	0.123	0.15	0.159	0.175	0.176	0.186
	ACFM-I	0.113	0.144	0.154	0.168	0.172	0.176
	ACFM-UMVDT	0.102	0.126	0.14	0.17	0.181	0.186
Wine	MLP-I	0.183	0.215	0.282	0.324	0.372	0.488
	AANN-I	0.211	0.25	0.246	0.294	0.352	0.392
	AANN-UMVDT	0.199	0.214	0.205	0.208	0.215	0.233
	ACFM-MEI	0.172	0.197	0.198	0.204	0.211	0.225
	ACFM-I	0.176	0.185	0.189	0.196	0.201	0.21
	ACFM-UMVDT	0.175	0.186	0.194	0.199	0.206	0.215
Concrete	MLP-I	0.452	0.402	0.401	0.45	0.481	0.57
	AANN-I	0.465	0.405	0.478	0.519	0.524	0.569
	AANN-UMVDT	0.501	0.429	0.466	0.498	0.493	0.511
	ACFM-MEI	0.3	0.288	0.32	0.372	0.372	0.403
	ACFM-I	0.31	0.298	0.331	0.383	0.361	0.386
	ACFM-UMVDT	0.336	0.342	0.4	0.436	0.431	0.504
Abalone	MLP-I	0.155	0.219	0.352	0.499	0.451	0.631
	AANN-I	0.567	0.547	0.605	0.633	0.632	0.535
	AANN-UMVDT	0.133	0.114	0.154	0.189	0.191	0.337
	ACFM-MEI	0.184	0.212	0.248	0.336	0.381	0.467
	ACFM-I	0.145	0.163	0.196	0.223	0.243	0.246
	ACFM-UMVDT	0.119	0.137	0.125	0.168	0.171	0.193

fit regression relations by modifying the cost function compared to AANN. Meanwhile, compared with MLP, ACFM fits multiple regression relations through one architecture, which increases the generalization ability of ACFM on the premise of improving the imputation accuracy.

The impact of missing value error on imputation results: it can be observed from Tables 2 and 3 that ACFM-I performs slightly worse than ACFM-MEI under the missing rates of 10% and 20% in the Glass dataset, 5% in the Wine dataset, and four kinds of missing rates on the Concrete dataset. In addition, the imputation result of ACFM-I is better than ACFM-MEI in imputation results. It shows that the optimization of model parameters by missing value error affects the accuracy of modeling and thus leads to the poor performance of imputation results.

Taking the Iris dataset as an example, the imputation results of ACFM-MEI and ACFM-I at missing rates of 5%-30% are shown in Figure 6. With the increase of missing rate, the gap between imputation results of ACFM-MEI and ACFM-I becomes larger. If we continue to use the missing value error to optimize the model parameters when there are more and more missing values in the dataset, the deviation of model will also increase. Therefore, in this paper, equation (4) is used as the cost function of ACFM in this paper; that is, only the errors of existing data are used to optimize the model parameters, which have certain reasonableness and correctness.

Comparison between UMVDT and traditional training scheme: except for the results of Glass and Concrete datasets at missing rates of 5% and 10%, the results of AANN-UMVDT are better than those of AANN-I. Among the 60 imputation results, the results of ACFM-UMVDT are better than those of ACFM-I accounted for 66.7%, wherein the Concrete data set of data values vary greatly. There are many zero values and large values, and many samples have the same value in attributes. UMVDT will change the missing values during the training process, which may cause the imputation results of many samples with the same value to

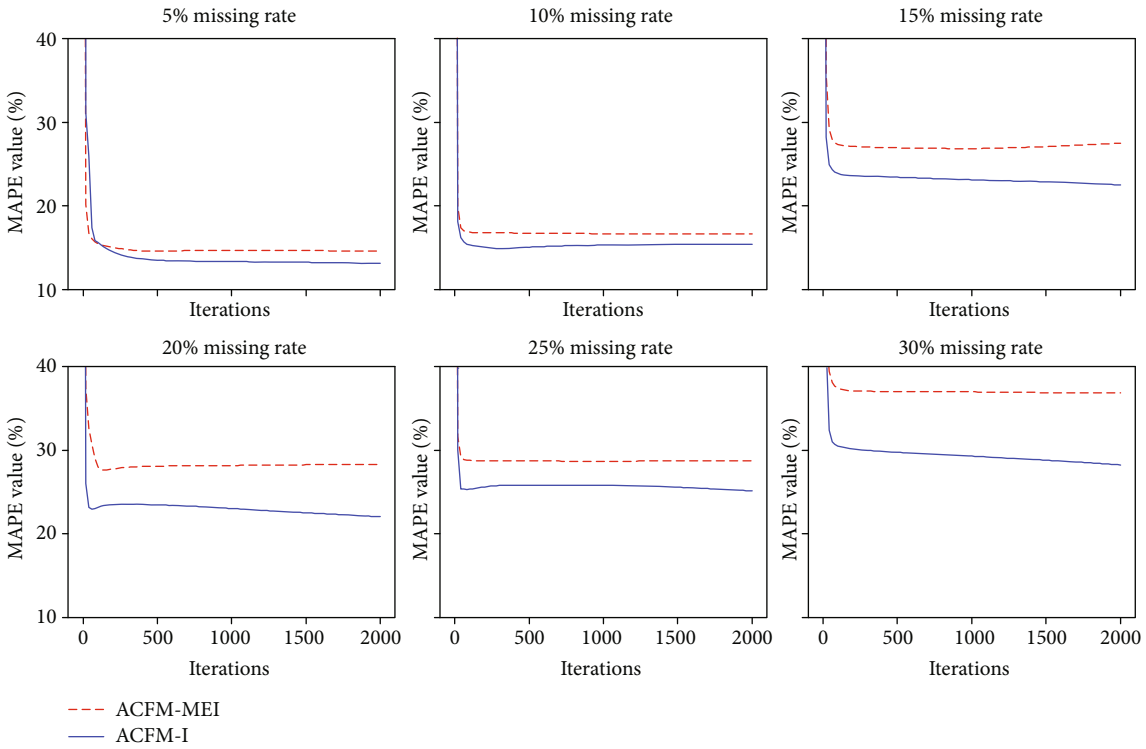

Figure 6: MAPE values of ACFM-I and ACFM-MEI on the Iris dataset.

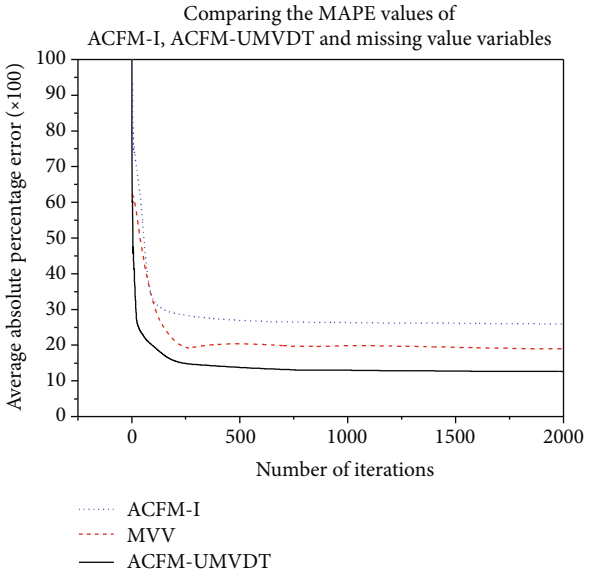

Figure 7: The imputation results of ACFM-I and ACFM-UMVDT and updating results of missing value variables on the Iris dataset, reproduced from Jinchong Zhu et al. 2020.

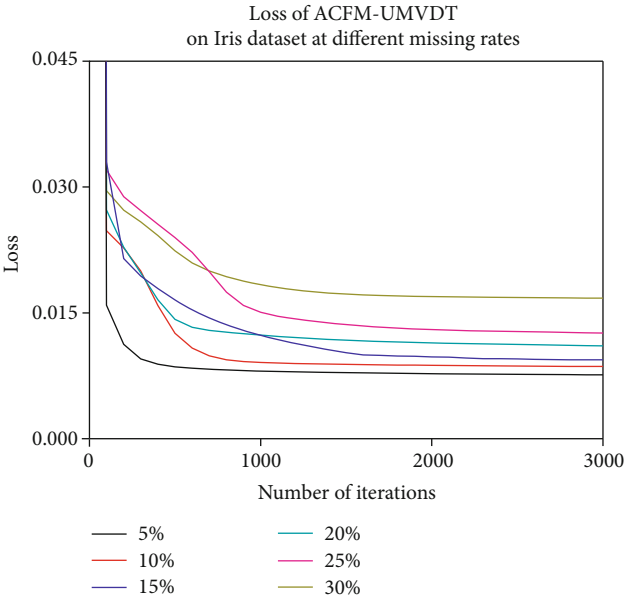

Figure 8: Convergence curve of fitting error of ACFM-UMVDT on the Iris dataset.

be unstable. The above results show that UMVDT training scheme has higher imputation performance than traditional one. UMVDT training scheme makes full use of the whole existing data in incomplete records and takes missing values as variables to make them gradually match the fitting relationship. The missing value variables and model parameters are updated alternately; so, the imputation effect can be improved significantly.

When the missing rate of the Iris dataset is 15%, the imputation of ACFM-I and ACFM-UMVDT and the variation of the missing value variables (MVV) of ACFM-UMVDT in each round are shown in Figure 7. It can be found that the missing values tend to be stable soon after a short period of fluctuation, and the imputation results of ACFM-UMVDT also tend to be stable with the increase of iteration rounds. The missing value is updated iteratively.

Not only the MAPE values calculated by missing value variables are more accurate than those of original model but also the imputation accuracy can be further improved by the model which is trained by the data updated iteratively.

The convergence of the proposed method: we take the Iris dataset as an example to verify the convergence of the proposed method. Figure 8 shows the fitting error of ACFM-UMVDT at various missing rates. It can be observed that all curves of fitting errors decrease in different degrees at beginning and become stabilized gradually. It is because the UMVDT training scheme constantly updates missing value variables and changes missing values in incomplete records. Missing value variables and model parameters converge continuously in the alternate updating process. The curves in Figure 8 show that the imputation method proposed in this paper has ideal convergence.

5. Conclusions

To solve the problem of imputation of missing values, this paper conducts attribute association modeling for incomplete data based on AANN. By modifying the cost function of AANN, this paper represents the regression relation between each attribute and the rest attributes of incomplete data on one architecture and redesign ACFM for enhanced to fit the association relation between data attributes. And we only use the training errors of existing data to optimize the model to reduce the inaccurate error between missing values and its predicted values in incomplete data to optimize the model. In addition, for the problem of incomplete model input, this paper proposes UMVDT training scheme, which sets missing values as variables and updates the model parameters and missing value variables alternately. UMVDT gradually optimizes the missing value variables through the regression structure of the model and further reduces the negative impact of the uncertainty of missing values during model input on the model. Experimental results show that the ACFM model can obtain more accurate imputation results compared with MLP and AANN models, and UMVDT further improves the accuracy of imputation on AANN and ACFM models by gradually iterating the missing value variables compared with traditional training scheme.

Data Availability

All datasets in this study can be downloaded from http://archive.ics.uci.edu/ml/datasets.php. And all experimental results are included in this published article.

Conflicts of Interest

We declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of China (62076050, 62073056) and National Key R&D Program of China (2018YFB1700200).

References

[1] F. V. Nelwamondo, D. Golding, and T. Marwala, "A dynamic programming approach to missing data estimation using neural networks," *Information Sciences*, vol. 237, pp. 49–58, 2013.

[2] S. M. C. M. Nor, S. M. Shaharudin, S. Ismail, N. H. Zainuddin, and M. L. Tan, "A comparative study of different imputation methods for daily rainfall data in east-coast peninsular Malaysia," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 2, pp. 635–643, 2020.

[3] V. R. Elgin Christo, H. Khanna Nehemiah, B. Minu, and A. Kannan, "Correlation-Based Ensemble Feature Selection Using Bioinspired Algorithms and Classification Using Backpropagation Neural Network," *Computational and Mathematical Methods in Medicine*, vol. 2019, Article ID 7398307, 17 pages, 2019.

[4] I. B. Aydilek and A. Arslan, "A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks," *International Journal of Innovative Computing, Information and Control*, vol. 7, no. 8, pp. 4705–4717, 2012.

[5] Z. Ghahramani and M. Jordan, "Supervised learning from incomplete data via an EM approach," in *Advances in Neural Information Processing Systems*, J. Cowan, G. Tesauro, and J. Alspector, Eds., vol. 6, pp. 120–127, Morgan-Kaufmann, 1994.

[6] R. Suphanchaimat, S. Limwattananon, and W. Putthasri, "Multiple imputation technique: handling missing data in real world health care research," *Southeast Asian Journal of Tropical Medicine and Public Health*, vol. 48, no. 3, pp. 694–703, 2017.

[7] K. Yang, J. Li, and C. Wang, "Missing values estimation in microarray data with partial least squares regression," in *International Conference on Computational Science*, pp. 662–669, Springer, Berlin, Heidelberg, May 2006.

[8] A. Tealab, "Time series forecasting using artificial neural networks methodologies: a systematic review," *Future Computing and Informatics Journal*, vol. 3, no. 2, pp. 334–340, 2018.

[9] I. A. Gheyas and L. S. Smith, "A neural network-based framework for the reconstruction of incomplete data sets," *Neurocomputing*, vol. 73, no. 16-18, pp. 3039–3065, 2010.

[10] P. K. Sharpe and R. J. Solly, "Dealing with missing values in neural network-based diagnostic systems," *Neural Computing & Applications*, vol. 3, no. 2, pp. 73–77, 1995.

[11] N. Ankaiah and V. Ravi, "A novel soft computing hybrid for data imputation," in *Proceedings of the International Conference on Data Mining (DMIN) (P. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)*, Las Vegas, USA, 2011.

[12] L. Gondara and K. Wang, "Mida: multiple imputation using denoising autoencoders," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 260–272, Springer, Cham, 2018.

[13] M. Abdella and T. Marwala, "The use of genetic algorithms and neural networks to approximate missing data in database," in *IEEE 3rd international conference on computational cybernetics, 2005. ICCC 2005*, pp. 207–212, Mauritius, April 2005.

[14] B. L. Betechuoh, T. Marwala, and T. Tettey, "Autoencoder networks for HIV classification," *Current Science*, vol. 91, no. 11, pp. 1467–1473, 2006.

[15] T. Marwala and S. Chakraverty, "Fault classification in structures with incomplete measured data using autoassociative

neural networks and genetic algorithm," *Current Science*, vol. 90, no. 4, pp. 542–548, 2006.

[16] F. J. Mistry, F. V. Nelwamondo, and T. Marwala, "Missing data estimation using principle component analysis and autoassociative neural networks," *Journal of Systemics, Cybernatics and Informatics*, vol. 7, no. 3, pp. 72–79, 2009.

[17] A. K. Mohamed, F. V. Nelwamondo, and T. Marwala, "Estimating missing data using neural network techniques, principal component analysis and genetic algorithms," in *Proceedings of the Eighteenth Annual Symposium of the Pattern Recognition Association of South Africa*, Pietermaritzburg, South Africa, 2007.

[18] G. Ssali and T. Marwala, "Estimation of missing data using computational intelligence and decision trees," 2007, https://arxiv.org/abs/0709.1640.

[19] V. Ravi and M. Krishna, "A new online data imputation method based on general regression auto associative neural network," *Neurocomputing*, vol. 138, pp. 106–113, 2014.

[20] C. Gautam and V. Ravi, "Data imputation via evolutionary computation, clustering and a neural network," *Neurocomputing*, vol. 156, pp. 134–142, 2015.

[21] C. Gautam and V. Ravi, "Counter propagation auto-associative neural network based data imputation," *Information Sciences*, vol. 325, pp. 288–299, 2015.

[22] K. J. Nishanth and V. Ravi, "A computational intelligence based online data imputation method: an application for banking," *Journal of Information Processing Systems*, vol. 9, no. 4, pp. 633–650, 2013.

[23] E. L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, and M. D. Cubiles-de-la-Vega, "Missing value imputation on missing completely at random data using multilayer perceptrons," *Neural Networks*, vol. 24, no. 1, pp. 121–129, 2011.

[24] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Classifying patterns with missing values using multi-task learning perceptrons," *Expert Systems with Applications*, vol. 40, no. 4, pp. 1333–1341, 2013.

[25] P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.

[26] J. M. Jerez, I. Molina, P. J. García-Laencina et al., "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 105–115, 2010.

[27] P. J. García-Laencina, J. Serrano, A. R. Figueiras-Vidal, and J. L. Sancho-Gómez, "Multi-task neural networks for dealing with missing inputs," in *International work-conference on the interplay between natural and artificial computation*, pp. 282–291, Springer, Berlin, Heidelberg, June 2007.

[28] P. J. García-Laencina, J. Sancho-Gomez, and A. R. Figueiras-Vidal, "Pattern classification with missing values using multi-task learning," in *The 2006 IEEE international joint conference on neural network proceedings*, pp. 3594–3601, Vancouver, BC, Canada, July 2006.

[29] J. Yoon, J. Jordon, and M. Schaar, "Gain: missing data imputation using generative adversarial nets," in *International conference on machine learning*, vol. 80, pp. 5689–5698, Stockholm, Sweden, July 2018, PMLR.

[30] X. Kong, K. Wang, S. Wang et al., "Real-time mask identification for COVID-19: an edge computing-based deep learning framework," *IEEE Internet of Things Journal*, 2021.

[31] X. Kong, S. Tong, H. Gao et al., "Mobile edge cooperation optimization for wearable internet of things: a network representation-based framework," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 5050–5058, 2021.

[32] J. Zhu, L. Zhang, X. Lai, and G. Zhang, "Imputation of incomplete data based on attribute cross fitting model and iterative missing value variables," in *International symposium on neural networks*, pp. 167–175, Springer, Cham, December 2020.