Research Article

Missing Data Interpolation of Alzheimer's Disease Based on Column-by-Column Mixed Mode

Shi-di Miao ¹, Si-qi Li ¹, Xu-yang Zheng,¹ Rui-tao Wang ², Jing Li ³, Si-si Ding,⁴
and Jun-feng Ma ¹

¹School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150000, China

²Department of Internal Medicine, Harbin Medical University Cancer Hospital, Harbin Medical University, Harbin 150000, China

³Department of Geriatrics, The Second Affiliated Hospital, Harbin Medical University, Harbin 150000, China

⁴Information Consulting Division, Harbin Institute of Technology, Harbin 150000, China

Correspondence should be addressed to Jing Li; 15331886892@189.cn

Received 14 May 2021; Revised 22 August 2021; Accepted 3 September 2021; Published 24 September 2021

Academic Editor: Daniele Salvati

Copyright © 2021 Shi-di Miao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Research on clinical data sets of Alzheimer's disease can predict and develop early intervention treatment. Missing data is a common problem in medical research. Failure to deal with more missing data will reduce the efficiency of the test, resulting in information loss and result bias. To address these issues, this paper designs and implements the missing data interpolation method of mixed interpolation according to columns by combining the four methods of mean interpolation, regression interpolation, support vector machine (SVM) interpolation, and multiple interpolation. By comparing the effects of the mixed interpolation method with the above four interpolation methods and giving the comparison results, the experiment shows that the results of the mixed interpolation method under different data missing rates have better performance in terms of root mean square error (RMSE), mean absolute error (MSE), and error rate, which proves the effectiveness of the interpolation mechanism. The characteristics of different variables might lead to different interpolation strategy choices, and column-by-column mixed interpolation can dynamically select the best method according to the difference of features. To a certain extent, it selects the best method suitable for each feature and improves the interpolation effect of the data set as a whole, which is beneficial to the clinical study of Alzheimer's disease. In addition, in the processing of missing data, a combination of deletion method and interpolation method is adopted with reference to expert knowledge. Compared with the direct interpolation method, the data set obtained by this method is more accurate.

1. Introduction

Alzheimer's disease (AD) is a high-incidence and irreversible progressive central nervous degenerative disease, mainly occurring in the elderly over 65 years old, with insidious onset and unknown cause [1]. The disease early major clinical manifestations of memory impairment, cognitive dysfunction, and mental damage, with the passage of time, will gradually deteriorate, as well as the final performance for mental health and memory loss, language, and action, such as capacity [2]. This seriously influences old people's physical and mental health and life, work, and social life, causing a heavy burden to family and society. According to relevant research reports, the

number of AD patients in the world is expected to reach 132 million by 2050, with an average of one in 85 people suffering from AD [3]. China is the country with the largest population in the world, and the number of the elderly is close to 300 million. The prevalence of the elderly over 65 years old is 5%, and the prevalence of the elderly over 85 years old is 20%, so the country is facing a severe challenge of AD [4]. Therefore, the research on the auxiliary diagnosis and disease prediction of AD is increasingly important.

With the application of machine learning in the medical field and the further development of related research, the research on computer-aided diagnosis models has attracted wide attention and achieved remarkable results [5–7].

However, due to the complexity of clinical data, there are relatively few achievements that can be applied in clinical practice and can provide effective assistance for clinical medical experts in disease diagnosis [8]. The current research on Alzheimer's disease mostly adopts data sets with complete and high accuracy, without paying attention to the complexity and incompleteness of clinical data. Therefore, considering the actual clinical data, it is of great significance to establish the auxiliary diagnosis system from the data preprocessing.

Missing value refers to the loss of data content caused by human negligence, machine failure, and lack of data sources in the process of data collection and collation, resulting in the incomplete data set, as shown in Figure 1. Data missing exists in all walks of life, such as clinical and medical data collection, report data statistics, and experimental data recording [9], while the direct use of data with missing data for data mining will affect the modeling results or even lead to errors. In addition, most of the current data mining algorithms cannot be directly applied to data sets containing missing data and have strong sensitivity to the proportion of missing data in data sets. Missing data processing is a very important step to achieve the integrity of data needed in data mining, so as to improve the quality of data and meet the needs of mining.

Reasonable processing of the missing data in the data set can effectively improve the data quality and improve the accuracy of subsequent modeling. At present, there are three main processing methods for missing data: deletion method, ignorance method, and interpolation method.

Taking the AD data set as the research goal, this paper uses different methods to interpolate different features, selects the most appropriate method, and combines this method with the direct deletion method to deal with the missing data. This method makes up for the shortcomings of single processing method and single interpolation method, makes the complete data obtained with higher accuracy, and preserves the integrity of the data to the maximum extent.

The paper is structured as follows. Section 2 describes the research status of missing data processing. Section 3 introduces the relevant technologies of missing data interpolation and the process of column-by-column mixed interpolation. Section 4 carries out missing value processing for data sets with different miss rates and analyzes the interpolation results according to RMSE and MSE. Section 5 gives the conclusion and the works in the future.

2. Literature Review

With the popularization and application of data mining in various industries, data preprocessing technology has been widely concerned by scholars from all walks of life at home and abroad, and its related technologies have also been rapidly developed.

There are two main processing methods for missing data: deleting data containing missing data and interpolating missing data [10]. The deletion method is to delete the instance sample data that contains missing data in the data set to obtain the remaining complete data set for subsequent

analysis. This method is simple and feasible, but its advantages and disadvantages are quite obvious. When the proportion of missing data is small, especially when a data sample contains multiple missing data, the overall impact of deleting data containing missing data is small. However, it may also lead to sample imbalance and loss of important data information. With the increase of the proportion of missing data, after the deletion of missing data, the remaining data will be difficult to reflect the true information, especially in the case of nonrandom missing data [11]. Therefore, the current research on missing data mainly focuses on interpolation.

Data mining is the most commonly used method to fill in the missing data in the row data set. The filling idea of this kind of method is to mine useful relevant information from the current data set, establish a model according to the mined information, and then predict and estimate the final value of missing data through this model [12]. The study in [13] improved the SNI algorithm on the basis of KNN algorithm, and its accuracy in filling experiments of mixed data sets was better than that of KNN algorithm. At the same time, the authors also proposed a new filling method NIIA, which is an iterative assignment scheme. The missing data is continuously assigned iteratively, and all the missing data are successively filled until convergence. Although this method has a high final filling accuracy, its speed is very slow, and it is difficult to apply to large data sets.

In literature [14], six different methods of mean, k -nearest neighbor (KNN), fuzzy k -means (FKM), singular value decomposition (SVD), Bayesian principal component analysis (BPCA), and chain equation multiple normalization (Mice) were compared under the hypothesis of complete random deletion, and Bayesian principal component analysis and fuzzy k -means interpolation were considered to have better effects. The study in [15], based on longitudinal data with "no person data" methods, "baseline data + other covariates" methods, "before data only" methods, and "before and after data" methods, estimated a missing value, using the root mean square error, the average absolute deviation, and deviation and relative variance as the evaluation standards for comparing four interpolation methods. The study in [16] used Naive Bayes and multiple interpolation to process the lost data of Alzheimer's disease cost analysis, adjusted the estimation rules by taking advantage of the missing information, and improved the use of single estimation technology, so as to obtain a more accurate cost estimate. The study in [17] proposed a data-driven missing value estimation method, which estimated the missing value for each feature using five methods, including global mean/mode, age-based mean/model, previous observation carried forward, previous and next observation combined, and the k -nearest neighbors. The best estimation method was selected according to the features, and interpolation results were evaluated under independent classifier and dependent classifier. The above literature mainly selects the appropriate interpolation method through the comparison of multiple interpolation methods. The above missing value interpolation methods are all the applications of the original methods. Some methods are only applicable to longitudinal data and

FIGURE 1: Comparison of data set. (a) Complete data set collection. (b) Missing data set collection.

are not applicable to the missing value interpolation of mixed data.

Based on literature [18], a principal component analysis method, the balance in the building variant dimension when all the influence of continuous variables and classification for the missing value prediction is based on the similarity between the individual and the relationship between the variables, the method in classification of variable assignment and highly linear relationship between continuous variables showed better performance. In literature [19], a new method based on correlation maximization is used to estimate missing data. Firstly, a base set is selected from the complete examples, and then the base set and other complete examples are used to generate data segments with strong correlation. Finally, a linear model is applied to the discovered data segments to calculate each missing value. The study in [20] inputs numbers and classifies missing values by making educated guesses based on records similar to those with missing values. When identifying a group of similar records and making guesses based on this group, a fuzzy clustering method and a new fuzzy expectation maximization algorithm are applied. Literature [21] proposed four different methods combining case selection and missing value estimation and compared them in data classification. The study in [22] studied a layered missing value estimation method based on the related k neighbor and chose the nearest neighbor when considering the relation, according to the number of missing values in each record in turn input incomplete records, in the process of reduction, full records by computing the correlation coefficient, which will update relevant record and reduction of the complete merger coefficient. With the improvement of data utilization, attributes are more accurately associated, making it more appropriate to select the nearest neighbor. The study in [23] proposed a missing data interpolation scheme combining evolutionary computing technology for the first time, and this scheme showed different effects for combining different algorithms. The above literatures all adopt a single missing value estimation method to interpolate the missing data, and the above missing data interpolation algorithms and schemes all regard all features of the data set as a whole,

without considering the differences among features in the interpolation process.

In recent years, the missing value interpolation technology has been applied in various fields. In literature [24], based on the long-term monitoring data of the steel structure of the Hangzhou Olympic Center Stadium, the correlation between the stress change of the measuring points is studied, and an interpolation method of the missing stress data is proposed. Data of daytime and nighttime are fitted separately for interpolation. The study in [25] introduces the soft sensor for industrial process monitoring, control, and optimization. Before it is applied to engineering systems, signal (energy) transformation and data sampling are usually required, and the sampled data need to be processed. The study in [26] developed a key performance indicator oriented fault detection Toolbox (DB-Kit) based on MATLAB toolbox data and provided the evaluation results of defects in the data set for the prediction and diagnosis system of key performance indicators. In literature [27], a novel ST-correlated proximate missing data imputation model was proposed to analyze the missing values of IoT data. The above literatures all deal with the missing values in different scenarios. The study in [28] proposed a novel latent representation learning method for multimodality based AD diagnosis; they not only use samples with complete multimodality data to learn a common latent representation, but also use samples with incomplete multimodality data to learn independent modality-specific latent representations. The study in [29] proposed an Auto-Encoder-based Multi-View missing data Completion framework (AEMVC) to learn common representations for AD diagnosis. This method firstly maps the original complete view to a latent space using an autoencoder network framework. Then, the latent representations measuring statistical dependence learned from the complete view are used to complement the kernel matrix of the incomplete view in the kernel space. The study in [30] proposes a complete unified “two-level” learning model for multisource data and extends it to incomplete data, which avoids estimation of missing data and provides superior performance. The study in [31] proposes a view-aligned hypergraph learning (VAHL) method to explicitly simulate

the consistency between views. Raw data is divided into several views according to possible pattern combinations, and the hypergraph building process based on sparse representation is carried out in each view, and each view corresponds to a specific pattern or a combination of several patterns. The study in [32] presented a Complete Multimodality Latent Space (CMLS) learning model for complete multimodality data and also an Incomplete Multimodality Latent Space (IMLS) learning model for incomplete multimodality data. Literature [33] proposed a high-order Laplacian regularized low-rank representation method for dementia diagnosis using block-wise missing multimodal data. The study in [34] proposed a multi-hypergraph learning method for dealing with incomplete multimodality data. Specifically, we first construct multiple hypergraphs to represent the high-order relationships among subjects by dividing them into several groups according to the availability of their data modalities. The study in [35] proposed a spatially constrained Fisher representation framework for brain disease diagnosis based on incomplete multimodal neural images, using a mixed antagonistic network to estimate missing PET images according to corresponding MRI scans. The above literature used different methods to deal with the missing values of the clinical data of Alzheimer's disease. However, the single methods can cause a variety of problems, such as reducing the difference of data changes, making partial interpolation results extreme value and the convergence rate slow, and if the partial missing values in the prediction model are unrelated of other attribute values, the prediction results are meaningless.

The study in [36] proposed a new regression calculation method, GP-KNN, which is a hybrid method. It adopts two concepts of genetic programming reduction (GPI) and k -nearest neighbor (KNN). GP-KNN considers both feature and case correlation. On the basis of artificial neural network, the study in [37] proposed a single value assignment method of multilayer perceptrons trained by different learning rules, and a multivalued assignment method based on the combination of multilayer perceptrons and k -nearest neighbor. Among them, the k -nearest neighbor algorithm is an interpolation method based on distance calculation. The accuracy of this algorithm is very high, but its shortcomings are also obvious. When missing data accounts for a large proportion of the total data, its filling accuracy will be greatly reduced. In this study, a new column-by-column mixed interpolation method based on data-driven is proposed. First of all, different interpolation methods are used to estimate the missing value of each feature of the data set, and the interpolation methods of each feature are sorted according to the average error, and then the interpolation method with the minimum error is selected as the most appropriate interpolation method.

3. Missing Data Interpolation

The existing interpolation methods for missing data are mainly divided into single interpolation and multiple interpolation, in which the single difference interpolation includes mean interpolation, maximum expectation

interpolation, regression interpolation, and decision tree interpolation. There are three main methods of multiple interpolation: regression prediction method, propensity score method, and MCMC. Since the deletion mode of Alzheimer's disease data is arbitrary deletion, this paper adopts MCMC multiple interpolation method and three common single interpolation methods of high accuracy mean interpolation, multiple linear regression interpolation, and support vector machine interpolation to form a new interpolation method.

3.1. Related Technologies

3.1.1. Mean Interpolation. Mean interpolation is to replace the missing data with the mean value of the valid values of the feature attribute, where the missing data is located. The mean interpolation method is calculated as follows:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n \lambda_i y_i. \quad (1)$$

Here, n_i is the number of samples of data, λ_i is whether the i sample has a value in this attribute, 1 represents the existence, and 0 represents the nonexistence.

The interpolation results of some characteristic attributes obtained by this method are consistent, which will lead to sample distortion, so layered mean interpolation is commonly used. Hierarchical interpolation refers to stratifying the data set according to the attributes of variables before interpolation and then interpolating the mean value of the samples of each layer:

$$\bar{Y} = \frac{1}{n} \sum_{h=1}^H \left[\sum_{i=1}^{n_h} [\lambda_i y_i + (1 - \lambda_i) \bar{y}_{1h}] \right] = \frac{1}{n} \sum_{h=1}^H n_h \bar{y}_{1h}, \quad (2)$$

wherein H is the number of layers of data, n_h represents the number of samples contained in the h layer, and $\bar{y}_{1h}$ represents the mean value of samples containing real values for an attribute in the h layer.

3.1.2. Regression Interpolation. The interpolation of missing data by regression method takes the feature attribute of missing data as the dependent variable and other feature attributes in the data set as the independent variable, establishes a regression model by using the relationship between independent variables and dependent variables, and uses the model to predict the missing data so as to complete the interpolation of missing data. Its interpolation value is calculated as follows:

$$y_k = \alpha_0 + \sum_{i=1}^m \alpha_i x_{ki}, \quad (3)$$

where y_k is the prediction result of the k missing data, α_i is the regression coefficient of the model, and x_i is the auxiliary variable. According to the formula, when the auxiliary variable x_i is the same, the prediction interpolation result is the same as the mean interpolation; that is, the predicted value of all the missing data is the same. Therefore, during

regression interpolation, it is necessary to construct a random residual term and add the predicted results to form the final interpolation value, so as to avoid sample distortion. Take equation (3) as follows:

$$\bar{y}_k = \alpha_o + \sum_{i=1}^m \alpha_i x_{ki} + e_i, \quad (4)$$

where e_i represents the residuals of the constructed data set.

3.1.3. Multiple Interpolation. Multiple interpolation is a missing data interpolation method based on repeated simulation thought first proposed by Rubin in 1978 [38]. The idea is derived from Bayesian inference, which includes m ($m > 1$). The interpolation vector of the interpolation values of the missing data replaces each missing data, and then m complete data sets are constructed according to the interpolation vector. The m data sets obtained are analyzed according to the statistical standard. Finally, the overall evaluation of the analysis results was carried out to obtain the final estimate of the target variable. In summary, the multiple interpolation method mainly includes three steps: missing data interpolation, overall analysis, and evaluation and merger.

Multiple interpolation can maintain the relationship between the variables of the original data set well and can provide a lot of possible information for the uncertainty of the estimated results.

Let us say that Y is the nm matrix with m variables. The existing data part in Y is Y_o , and the missing part is Y_m , then $Y = (Y_o, Y_m)$, and Y obeys the dimensional normal distribution $\mu = (\mu_1, \mu_2, \dots, \mu_m)$, $\varepsilon = (\delta_{jM})$. The interpolation method is as follows:

- (1) Select a vector arbitrarily θ^* from the parameter vectors θ to be estimated.
- (2) $Y^*i(m)$ in the conditional distribution $P(Y^*i(m)/Y^*i(o), \theta^*)$.
- (3) Assuming that the parameter to be estimated is α , then $\alpha = \alpha(Y) = \alpha(Y_o, Y^*m)$ can be obtained from the constructed complete data set θ . The variance between the interpolations is $U = \text{var}(\alpha)$.
- (4) Repeat the above three steps M times to get $\alpha_{(j)}, U_{(j)}$, where $j = 1, 2, 3, \dots, m$. Integrating the above analysis and derivation results, the estimated value of multiple interpolation parameter α , interpolation internal variance, interinterpolation variance, and total variance can be obtained, as shown in the following formulas, respectively:

$$\bar{\alpha} = \frac{1}{M} \sum_{j=1}^m \alpha_{(j)}, \quad (5)$$

$$U_1 = \frac{1}{M} \sum_{j=1}^m U_{(j)}, \quad (6)$$

$$U_2 = \frac{1}{M-1} \sum_{j=1}^m (\alpha_{(j)} - \bar{\alpha})(\alpha_{(j)} - \bar{\alpha})^T, \quad (7)$$

$$U_\tau = U_1 + \left(1 + \frac{1}{M}\right)U_2. \quad (8)$$

- (5) $F_0 = (\alpha_{(j)} - \bar{\alpha})^T U^{-1/2} (\alpha_{(j)} - \bar{\alpha})$: if the missing data of parameter α contains the same information, then the distribution of F_0 is close to the distribution of F . The 95% confidence interval estimation of parameter α approximates the t-distribution. The parameters are expressed as follows:

$$d = (M-1) \left[1 + \frac{U_1}{1 + M^{-1}U_2} \right]^2. \quad (9)$$

- (6) The estimation λ of the missing information of parameter α can be obtained:

$$\lambda = \frac{\gamma + 2/(d+3)}{\gamma + 1}, \quad (10)$$

where γ represents the increment of variance generated by missing data, which can be expressed as

$$\gamma = \frac{(1 + 1/M)U_2}{U_1}. \quad (11)$$

3.1.4. Support Vector Machine Interpolation. Support Vector Machine (SVM) is a classification method based on statistical learning and minimal structural risk theory. The application idea of interpolation and missing data is similar to SVM function [39]. For the linear regression problem, the data set is firstly obtained from the data set, and then the model is trained and learned. The data is fitted by the method of linear regression, and the final regression function is obtained after error fitting and conditional constraints:

$$f(x) = \sum_{i=1}^n (\alpha_i^* - \alpha_i)(x_i \cdot x) + b^*. \quad (12)$$

α_i is the Lagrange factor. Finally, the data set D_{miss} containing missing data is brought into the trained model to predict the final interpolation result.

3.2. Missing Data Interpolation. Starting from each feature of the data set, this paper realizes the interpolation of each feature one by one and puts forward the method of mixed interpolation according to column. The flow chart is shown in Figure 2:

- (i) Starting from the data set containing missing data in practical application, record the feature R_i containing missing data from the data set R containing missing data, and sort $R_1, R_2, R_3, \dots$ in an ascending order according to the amount of missing data.

FIGURE 2: Flow chart of missing data interpolation.

- (ii) Delete the individual sample with missing data from the data set R to obtain the complete data set R' .
- (iii) Select the corresponding R_i feature attribute in the data set R' , and blank the R_i feature attribute in the data set R' according to the proportion of missing data of R_i feature attribute in the original data set R' to generate a new data set R'' .
- (iv) The mean interpolation method, multiple linear regression interpolation method, support vector machine interpolation method, and Markov chain Monte Carlo multiple interpolation method are used to interpolate the data set R'' , and the interpolation results of each method X_i' are compared with the original accurate value X_i , and the best interpolation effect is selected as the interpolation method of the feature attribute.
- (v) Use the interpolation method determined in 4 to interpolate the corresponding R_i feature attributes in the data set R .
- (vi) Traverse all missing features $R_1, R_2, R_3, \dots$. Repeat steps 3, 4, and 5 until the data set R is complete.

There may be a high proportion of missing data or scattered missing data in the original data set, and the data samples obtained may be out of balance or the amount of data may be too small. Therefore, it is necessary to choose whether or not to expand the data according to the actual situation of the data set.

On the basis of column-by-column mixed interpolation method, this paper puts forward a missing data processing flow for AD data by combining the direct deletion method of missing data and the expert knowledge in the field of Alzheimer's disease medicine, as shown in Figure 3:

- (i) First, the missing rate of each sample data in the data set containing missing data is counted, and then some sample variables are deleted according to the set threshold of missing rate of sample data.
- (ii) On the basis of deleting some sample variables, the data missing rate of each characteristic variable is counted, and the characteristic variable to be deleted is selected according to the threshold value of missing rate of the characteristic variable set.
- (iii) On the basis of selecting feature variables to be deleted, some feature variables are deleted according to the opinions of experts.
- (iv) On the basis of deleting some sample variables and characteristic variables, interpolation of missing data is carried out to obtain a complete data set.

4. Experimental Analysis

4.1. Missing Data Processing. The data source used in this experiment was the *ADNIMERGE* data set from ANDI database [40]. A total of 15 related features were extracted from the complete data set, including DX, AGE, ADAS13, MMSE, Rimmed, R.learn, R.p.forget, LDEL, FAQ, Hippo, WBrain, Entorhinal, MidTemp, MOCA, and TAU. A total of 11087 sample data were included. The extracted complete ADNIMERGE data set was randomly divided into 90% training data and 10% test data.

Existing research results show that when the data missing rate of the data set is higher than 60%, the utilization value of the data set is basically zero, and no matter what processing method is adopted at this time; it is useless. Therefore, the loss rate will be 50% of the data, as a general data processing data set is missing rate limit, and the data loss rate of 30% or more, after processing accuracy of the data set, will be a serious decline; on this basis, this article under the lack of random model constructed, respectively, missing data at a rate of 5%, 10%, 20%, 30%, and 40%, the lack of data collection, and lack of experimental analysis in different rates under the effect of different interpolation methods.

4.2. Comparison of Interpolation Methods with 5% Data Missing Rate. According to the characteristic variables in the complete data set, a missing data set containing 5% of the missing data is generated. Figure 4 shows the distribution of the missing data in the missing data set.

As shown in Figure 4, the left part is the proportion of missing data contained in each feature in the total sample. It can be seen from the figure that the missing rate of each feature variable is about 5%, indicating that the missing data is distributed randomly and approximately uniformly among all the feature variables. In the right half, green

FIGURE 3: Missing data processing flow.

(majority) represents the complete data sample, and red represents the sample with missing data. Mean difference interpolation, multivariate linear back interpolation, support vector machine interpolation, multiple interpolation, and the column based mixed interpolation proposed in this paper were used to interpolate the above-mentioned missing data sets, respectively, and the interpolation results were analyzed for qualitative and quantitative data characteristics, respectively. Table 1 shows RMSE and MAE values of quantitative data interpolation results obtained by the five interpolation methods.

It can be seen from the table that when the data missing rate is 5%, the MAE value of mean interpolation and mixed interpolation is small; that is, the accuracy is high, while the accuracy of multiple interpolation is relatively low, and the effect is the worst. Although the accuracy of mixed interpolation and mean interpolation is similar and relatively high, the RMSE value of mean interpolation is higher; that is, the degree of dispersion between the interpolation result data and the real data is higher.

Figure 5 shows the Boxplot of MAE values obtained from 50 experiments with five interpolation methods under 5% data loss rate, and the box line distribution diagram obtained. The upper and lower limits of the box, respectively, represent the maximum and minimum values of the interpolation results, and the middle line is the median of the interpolation results. According to the figure, the MAE value of mean interpolation and mixed interpolation is the lowest, but the deviation range of

mixed interpolation is relatively smaller; that is, the interpolation results are more accurate.

Figure 6 is the box-line distribution diagram of RMSE value corresponding to Figure 5. As can be seen from the figure, the box height of multiple interpolation method is the smallest, which represents the lowest degree of dispersion of interpolation results and the most stable, but its RMSE value is high. Among them, the RMSE value of the mixed interpolation result is the lowest, and the relative addition of the box height is lower; that is, the interpolation effect of this method is optimal.

Table 2 shows the misclassification rates of the five interpolation methods on the results of the interpolation of qualitative variables. When the miss rate is 5%, the misclassification rates of multiple interpolation and mixed interpolation are low, while those of the other three methods are high; that is, the interpolation results are poor.

Considering the interpolation effects of the five interpolation methods on qualitative data and quantitative data, when the data missing rate is 5%, the interpolation results are not significantly different, but the mixed interpolation method has the best interpolation effect compared with the other four methods.

4.3. Comparison of Interpolation Methods at 10%, 20%, 30%, and 40% Data Missing Rates. On the complete data set, missing data sets with missing rates of 10%, 20%, 30%, and 40% are generated, respectively. As shown in Figure 7, missing data distribution diagrams of data sets with different missing rates are, respectively, generated. According to the figure, there is no obvious feature dependence among the missing data, and the data missing rate of each feature is approximately the overall missing rate of the data set; that is, the missing data is randomly and approximately evenly distributed among the data features of 16 independent variables.

Five interpolation methods are used to interpolate the missing data for four data sets with different data missing rates. MAE value, RMSE value, and miss-division rate of the interpolation results are shown in Tables 3–5, respectively.

According to the data in Table 3, when the miss rate is 10% and 20%, the MAE value of the results obtained by regression interpolation, support vector machine interpolation, and mixed interpolation is significantly lower than that obtained by mean interpolation and multiple interpolation, indicating that the better the interpolation effect is. The MAE value of the mixed interpolation method is slightly lower than that of the regression interpolation method and the support vector machine interpolation method, which indicates that the mixed interpolation method has the highest absolute accuracy. When the data missing rate is 30%, the MAE value of the mixed interpolation method is significantly lower than that of the other four methods; that is, the absolute accuracy of the interpolation results is the highest. When the data missing rate reaches 40%, the MAE value of mixed interpolation and multiple interpolation is lower than that of the other three interpolation methods, and the MAE value of mixed interpolation is slightly lower than

FIGURE 4: Missing data distribution map at 5% data missing rate.

TABLE 1: Comparison of RMSE and MAE values of five interpolation methods.

Evaluation	Mean value interpolation	Regression interpolation	SVM interpolation	Multiple interpolation	Mixed interpolation
RMSE	0.647	0.632	0.631	0.638	0.628
MAE	0.612	0.621	0.617	0.624	0.611

that of multiple interpolation, but it is not significant. Combined with the above results, when the data missing rate is no more than 40%, the absolute accuracy of the interpolation results obtained by the mixed interpolation method is higher than that of the other methods; that is, the interpolation results are the most accurate.

Table 4 shows the RMSE values of the interpolation results obtained by five interpolation methods under different data loss rates corresponding to Table 3. The data in the table show that, under the same data loss rate, the mixed interpolation method is better than the other four methods. The results show that the discretization between the interpolation results and the real data is the lowest, and the interpolation effect is the best.

Table 5 is five interpolation methods under different data loss rate on qualitative variable interpolation results fault

rate, and it can be seen that the mixed interpolation method in missing data is at a rate of 10%, 20%, 30%, and 40%, and the mixed interpolation method of interpolation results for fault points rate compared with other four ways has a certain degree of decline; it shows that, in the inserted by loss of qualitative variables missing data in data collection, the mixed interpolation method of interpolation effect is the best.

4.4. Change and Comparison of the Interpolation Methods under Different Data Missing Rates. As shown in Figure 8 for five kinds of interpolation methods in the interpolation results under different data loss rate of MAE value, the data loss rate is low (5%), and that of the mean value interpolation method and the mixed interpolation method of

FIGURE 5: Boxplot of MAE values of five interpolation methods under 5% data miss rate.

FIGURE 6: Boxplot of RMSE values of five interpolation methods under 5% data miss rate.

TABLE 2: Error rate of five interpolation methods under 5% data missing rate.

Evaluation	Mean value interpolation	Regression interpolation	SVM interpolation	Multiple interpolation	Mixed interpolation
Error rate	31.37	32.56	31.24	29.57	28.49

FIGURE 7: Continued.

(c)
FIGURE 7: Continued.

FIGURE 7: Missing data distribution under different data missing rates. (a) 10% data missing rates. (b) 20% data missing rates. (c) 30% data missing rates. (d) 40% data missing rates.

TABLE 3: MAE values of the results of the five interpolation methods under different miss rates.

Data missing rate (%)	Mean value interpolation	Regression interpolation	SVM interpolation	Multiple interpolation	Mixed interpolation
10	0.668	0.649	0.641	0.662	0.632
20	0.692	0.656	0.649	0.676	0.643
30	0.721	0.682	0.679	0.685	0.661
40	0.787	0.732	0.711	0.704	0.691

TABLE 4: RMSE values of the results of the five interpolation methods under different miss rates.

Data missing rate (%)	Mean value interpolation	Regression interpolation	SVM interpolation	Multiple interpolation	Mixed interpolation
10	0.697	0.679	0.668	0.671	0.661
20	0.752	0.708	0.697	0.696	0.683
30	0.791	0.746	0.721	0.714	0.701
40	0.869	0.812	0.784	0.776	0.758

interpolation results of absolute error is minimum. However, as the data loss rate rises, the error of the mean value interpolation method obviously rises and is completely

higher than that of the other four kinds of interpolation methods. However, the MAE value of mixed interpolation method is lower than that of the other four methods at

TABLE 5: Error rates of interpolation results of five interpolation methods under different data miss rates.

Data missing rate (%)	Mean value interpolation	Regression interpolation	SVM interpolation	Multiple interpolation	Mixed interpolation
10	32.58	33.49	31.34	31.38	30.32
20	35.92	35.78	33.74	34.82	33.57
30	39.44	37.43	36.91	36.01	35.21
40	46.83	42.37	40.16	38.15	37.16

FIGURE 8: MAE values of the five interpolation methods under different data miss rates.

FIGURE 10: Error rate of five interpolation methods under different data missing rates.

FIGURE 9: RMSE values of the five interpolation methods under different data miss rates.

different data missing rates, and from 5% to about 30%, the advantages of this method gradually increase compared with the other four methods. When the data missing rate reaches 40%, the advantages of this method decrease, but the MAE is still lower than that of the other methods.

Figure 9 shows the RMSE values of the interpolation results of five interpolation methods under different data loss rates. When the data loss rate is about 30% or less, the interpolation method of RMSE value basically presents the increasing trend with the increase of loss rate, and that of the support vector machine (SVM) interpolation method and the mixed interpolation method of overall RMSE value is lower than the other way, but the change of RMSE value of mixed interpolation method tends to be more stable.

Figure 10 shows the error rate of five interpolation methods for the results of missing data interpolation of qualitative variables under different data loss rates. The result of the mixed interpolation method of overall error rate is lower than the other way. In the range of 5% to 40% data missing rate, the error rate of the mixed interpolation method for qualitative variables is basically less than or equal to the minimum value of the other four methods under the data missing rate, which proves that this method is the best to fill the missing data of qualitative variables.

5. Conclusions and Future Work

Firstly, a mixed interpolation mechanism based on mean interpolation, regression interpolation, support vector machine (SVM) interpolation, and multiple interpolation is proposed, and the interpolation effects under different data loss rates are compared with the four interpolation methods.

Experiments show that the mixed interpolation by column can improve the accuracy of interpolation results effectively. Secondly, a more reasonable missing data processing method is formed on the basis of the proposed mixed interpolation mechanism by column and the missing data deletion method.

For the missing data in the data set, this paper proposes a mixed interpolation mechanism based on feature differences. In this interpolation mechanism, the effective mean interpolation, regression interpolation, support vector machine (SVM) interpolation, and mixed interpolation are selected as the basic methods. Before the interpolation of missing data, each feature containing missing data is simulated, and the optimal method in the corresponding state is selected to complete the missing data interpolation. In this paper, the accuracy of the interpolation results obtained by this interpolation mechanism under different data missing rates is analyzed. The experimental results show that, within a certain range, the interpolation results obtained by this interpolation mechanism are superior to other interpolation methods, and the higher the data missing rate is, the more obvious the advantage is. In addition, the deletion method and interpolation method of missing data are combined in this paper, and the constraint of expert knowledge is added to make the data more accurate and scientific.

There is still much work to be done in the processing of missing data: (1) take the nonrandom missing model of data into account in future learning and research; (2) consider studying the same longitudinal data of Alzheimer's disease over a period of time; (3) consider the impact of more basic interpolation methods on column-by-column mixed interpolation method. We will further study the processing method of missing value of Alzheimer's disease data to provide more accurate help for clinical diagnosis.

Data Availability

Data used in preparation of this article were obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at <http://adni.loni.usc.edu>.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the present study.

Acknowledgments

This study was funded by the National Natural Science Foundation of China Youth Fund Project (71701056) and Heilongjiang Provincial Postdoctoral Funding Project (LBH-Z15100).

Supplementary Materials

In the Data Availability section, We provide a brief supplementary data description. Dataset is provided under figures & tables section due to larger file size. (*Supplementary Materials*)

References

- [1] K. Hahn, N. Myers, S. Prigarin et al., "Selectively and progressively disrupted structural connectivity of functional brain networks in Alzheimer's disease - revealed by a novel framework to analyze edge distributions of networks detecting disruptions with strong statistical evidence," *NeuroImage*, vol. 81, no. 5, pp. 96–109, 2013.
- [2] M. Liu, D. Cheng, K. Wang, and Y. Wang, "Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis," *Neuroinformatics*, vol. 16, no. 3-4, pp. 295–308, 2018.
- [3] C.-Y. Wee, P.-T. Yap, D. Zhang et al., "Identification of MCI individuals using structural and functional connectivity networks," *NeuroImage*, vol. 59, no. 3, pp. 2045–2056, 2012.
- [4] Y. Yu, M. Li, L. Liu, Y. Li, and J. Wang, "Clinical big data and deep learning: applications, challenges, and future outlooks," *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 288–305, 2019.
- [5] H. Zetterberg, T. Skillbäck, N. Mattsson et al., "Association of cerebrospinal fluid neurofilament light concentration with Alzheimer disease progression," *JAMA Neurology*, vol. 73, no. 1, pp. 60–67, 2016.
- [6] C. D. Lehman, R. D. Wellman, D. S. M. Buist, K. Kerlikowske, A. N. A. Tosteson, and D. L. Miglioretti, "Diagnostic accuracy of digital screening mammography with and without computer-aided detection," *JAMA Internal Medicine*, vol. 175, no. 11, pp. 1828–1837, 2015.
- [7] E. Zhang, S. S. M. cheng, W. Lu, and X. Gu, "Birads features-oriented semi-supervised deep learning for breast ultrasound computer-aided diagnosis," *Physics in Medicine and Biology*, vol. 65, no. 12, 2019.
- [8] H. Ramshini, A. Ebrahim-Habibi, S. Aryanejad, and A. Rad, "Effect of cinnamomum verum extract on the amyloid formation of hen egg-white lysozyme and study of its possible role in Alzheimer's disease," *Basic and Clinical Neuroscience*, vol. 6, no. 1, pp. 29–37, 2015.
- [9] I. Ezzine and L. Benhlila, "A study of handling missing data methods for big data," in *Proceedings of the 2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, October 2018.
- [10] T. Iliou, C. N. Anagnostopoulos, and M. Nerantzaki, "A novel machine learning data preprocessing method for enhancing classification algorithms performance," in *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*, pp. 1–5, Halkidiki, Greece, September 2015.
- [11] S. Sarwar, Z. Ul-Qayyum, and A. Kaleem, "Machine learning based intelligent framework for data preprocessing," *The International Arab Journal of Information Technology*, vol. 15, no. 6, pp. 1010–1015, 2018.
- [12] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing & Applications*, vol. 19, no. 2, pp. 263–282, 2010.

- [13] S. Zhang, "Shell-neighbor method and its application in missing data imputation," *Applied Intelligence*, vol. 35, no. 1, pp. 123–133, 2011.
- [14] P. Schmitt, J. Mandel, and M. Guedj, "A comparison of six methods for missing data imputation," *Journal of Biometrics & Biostatistics*, vol. 6, no. 1, 2015.
- [15] J. Engels and P. Diehr, "Imputation of missing longitudinal data: a comparison of methods," *Journal of Clinical Epidemiology*, vol. 56, no. 10, pp. 968–976, 2003.
- [16] M. Belger, J. M. Haro, C. Reed et al., "How to deal with missing longitudinal data in cost of illness analysis in Alzheimer's disease-suggestions from the GERAS observational study," *BMC Medical Research Methodology*, vol. 16, no. 1, p. 83, 2016.
- [17] C. Ribero and A. Freitas, "A data-driven missing value imputation approach for longitudinal datasets," *Artificial Intelligence Review*, 2021.
- [18] V. Audigier, F. Husson, and J. Josse, "A principal component method to impute missing values for mixed data," *Advances in Data Analysis and Classification*, vol. 10, no. 1, pp. 5–26, 2016.
- [19] A. MSeftidian and N. Daneshpour, "Estimating missing data using novel correlation maximization based methods," *Applied soft Computing Journal*, vol. 91, no. 2, 2020.
- [20] M. Z. Islam and M. G. Rahman, "Missing value imputation using a fuzzy clustering based EM approach," *Knowledge and Information Systems*, vol. 46, no. 822, pp. 389–422, 2016.
- [21] T. Chih-Fong and F. Y. Chang, "Combining instance selection for better missing value imputation," *Journal of Systems and Software*, vol. 122, no. 8, pp. 63–71, 2016.
- [22] X. Liu, X. C. Lai, and L. Zhang, "A hierarchical missing value imputation method by correlation-based K-nearest neighbor," *Intelligent Systems and Applications*, vol. 1037, pp. 468–496.
- [23] F. Lobato, C. Sales, I. Araujo et al., "Multi-objective genetic algorithm for missing data imputation," *Pattern Recognition Letters*, vol. 68, no. 8, pp. 126–131, 2015.
- [24] Z. Zhang and Y. Luo, "Restoring method for missing data of spatial structural stress monitoring based on correlation," *Mechanical Systems and Signal Processing*, vol. 91, pp. 266–277, 2017.
- [25] Y. C. Jiang, S. Yin, J. D. Dong, and K. N. Okyay, "A review on soft sensors for monitoring, control and optimization of industrial processes," *IEEE Sensors Journal*, vol. 21, no. 11, pp. 12868–12881, 2010.
- [26] Y. Jiang and S. Yin, "Recent advances in key-performance-indicator oriented prognosis and diagnosis with a MATLAB toolbox: db-kit," *IEEE transactions on industrial informatics*, vol. 15, no. 5, pp. 2849–2858, 2019.
- [27] I. S. P. Mary and L. Arockiam, "Imputing the missing data in IoT based on the spatial and temporal correlation," in *Proceedings of the IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, Bangalore, India, 2017.
- [28] T. Zhou, M. Liu, K.-H. Thung, and D. Shen, "Latent representation learning for Alzheimer's disease diagnosis with Incomplete multi-modality neuroimaging and genetic data," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2411–2422, 2019.
- [29] Y. B. Liu, L. X. Fan, C. Q. Zhang, and T. Zhou, "Incomplete multi-modal representation learning for Alzheimer's disease diagnosis," *Medical Image Analysis*, vol. 69, 2021.
- [30] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye, "Bi-level multi-source learning for heterogeneous block-wise missing data," *NeuroImage*, vol. 102, pp. 192–206, 2014.
- [31] M. Liu, J. Zhang, P.-T. Yap, and D. Shen, "Diagnosis of Alzheimer's disease using view-aligned hypergraph learning with incomplete Multi-modality Data," *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*, vol. 9900, pp. 308–316, 2016.
- [32] T. Zhou, K. H. Thung, M. Liu, F. Shi, C. Zhang, and D. Shen, "Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging data," *Medical Image Analysis*, vol. 60, p. 101630, 2020.
- [33] A. M. Dong, Z. G. Li, M. L. Wang, D. G. Shen, and M. X. Liu, "High-order laplacian regularized low-rank representation for multimodal DementiaDiagnosis," *Frontiers in Neuroscience*, vol. 15, 2021.
- [34] M. Liu, Y. Gao, P.-T. Yap, and D. Shen, "Multi-hypergraph learning for incomplete multimodality data," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1197–1208, 2018.
- [35] Y. Pan, M. Liu, C. Lian, Y. Xia, and D. Shen, "Spatially-constrained Fisher representation for brain disease identification with incomplete multi-modal neuroimages," *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2965–2975, 2020.
- [36] A. H. Baligh, Q. Chen, B. Xue, and M. J. Zhang, "A hybrid GP-KNN imputation for symbolic regression with missing values," *Advances in Artificial Intelligence*, vol. 11320, no. 2, pp. 345–357, 2018.
- [37] E. Lsilva, P. M. Rafae, and L. C. Manuel, "Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and K-nearest neighbours for monotone patterns," *Applied Soft Computing*, vol. 29, no. 9, pp. 65–74.
- [38] H. Y. Wang, H. Li, and J. Y. Shen, "A novel hybrid fractal interpolation-SVM model for forecasting stock price indexes," *Fractals*, vol. 27, no. 4, 2019.
- [39] M. Scutari, C. Vitolo, and A. Tucker, "Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation," *Statistics and Computing*, vol. 29, no. 5, pp. 1095–1108, 2019.
- [40] <http://adni.loni.usc.edu/>.