Received: December 22, 2016

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Grey Fuzzy Neural Network-Based Hybrid Model for Missing Data Imputation in Mixed Database

Vijayakumar Kuppusamy[1]* Ilango Paramasivam[1]

[1] *School of Computer Science and Engineering,*
Vellore Institute of Technology University, Vellore, India
* Corresponding author's Email: kvijayakumar@vit.ac.in

Abstract: Nowadays, the missing data imputation is the novel paradigm to replace with the imputed value of the missing attribute. The missing data occurs due to bias information, non-response of the system. In the medical domain, it becomes the major challenge to impute the both categorical and numerical data. In this paper, the Grey Fuzzy Neural Network is proposed for missing data imputation in the mixed database. Initially, the WLI fuzzy clustering mechanism is utilized to generate the different clusters in which the medical data are grouped together. Then, we intend to integrate the Grey Wolf Optimizer (GWO) with the ANFIS network model, termed the Grey Fuzzy Neural Network (GFNN). The proposed method is mainly used to determine the optimal parameters to design the membership function. Finally, the hybrid prediction model is used to find out the imputed data for both categorical and numerical. In the hybrid prediction model, the categorical data is then imputed by the distance measure. The experimental results are validated, and performance is analysed by metrics such as MSE and RMSE using MATLAB implementation. The outcome of the proposed GFNN attains lower 0.13 MSE, and 0.35 RMSE ensures to impute the data significantly in the missing attribute of the mixed database.

Keywords: Categorical and Numerical missing data, WLI fuzzy clustering, Grey Wolf Optimizer, ANFIS, Hybrid prediction model.

1. Introduction

The imputation methods [1, 2, 3, 4] can be broadly classified into two types, which are single imputation and multiple imputations. The data in the missing attribute is imputed by one value, on the other hand, the imputed data is generated by natural variability and interpolation process [5]. Therefore, missing data arise when the dataset has several issues such as data loss, low signal-to-noise ratio, power limitations, limited storage space for data, expensive acquisition equipment, etc. [6]. Due to missing data in the database, the error is initialized and cannot obscure the information which results as a 'missing-ness.' The mechanism of missingness in the mixed database is categorized into Missing Completely At Random (MCAR), Not Missing At Random (NMAR) and Missing At Random (MAR) [7]. The imputation methods are named as i) listwise and pairwise deletion, ii) imputation procedure, iii) model-basedprocedure and iv) machine learning model.

The mixed attribute dataset poses categorical data, and numeric data becomes the major issue in the missing data imputation. Since the research has less attention to developing the method to assign the missing data in the input corpus, leads to acquiring the detrimental performance [8]. The most commonly used and efficient imputation method is k nearest neighbor (KNN) used to estimate the value to fill in the missing attribute. It exploits the k relevant instances of the data and provides simplicity, ease of implementation and achieved high accuracy. [9]. Also, the fuzzy rule based is widely in the data imputation method which sculpts the linguistic model structure which has the tendency to evaluate the value of missing data and mitigates the dimension reduction problem [10]. Furthermore, the machine learning model is a challenging task for the mixed data imputation

 DOI: 10.22266/ijies2017.0430.16

method. The key advantage of machine learning model is that achieves more flexibility and higher order interaction among the missing data in the attribute [11]. Some of the commonly used machine learning is Artificial Neural Network (ANN), Neural network (NN) [10], K-nearest neighbor (KNN) [12], support vector machine (SVM) [11] and so on. The drawback of conventional KNN is distance based learning and the computation cost is quite high. Similarly, the dimensionality problem may cause serious concern in the effectiveness and it finds difficult to find the imputed values if it contains multiple missing attributes.

The main objective of this paper is to design and develop the novel data imputation method using the grey fuzzy neural network and constraint-based hybrid prediction model. Here, the input medical dataset is given as input to the proposed model. The constraint-based hybrid prediction model is designed by the grey fuzzy neural network and WLI fuzzy clustering mechanism. Here, the input dataset constitutes the categorical data and numerical data. Firstly, the input dataset is undergone for the WLI fuzzy clustering mechanism where the data are grouped together to determine the data value for missing attribute. Secondly, the Adaptive Neuro-Fuzzy Inference System (ANFIS) is applied to the input database. The introduction section of this paper is followed by Section 2 demonstrates the approach of missing data imputation. The problem specification and challenges behind the data imputation is described in section 3. Then, the proposed methodology of missing data imputation is briefly explained in section 4. Section 5 demonstrates the experimental results and performance analysis. Finally, this paper concludes in Section 6.

2. Motivation behind the data imputation

2.1. Problem specification

The main problem of missing data arises in clinical and medical datasets. Due to missing data, the information about the patient gets lost which leads to bias in the system [7]. The missing data is caused by such as while entering the data, equipment bias and erroneous measurements. Due to missing, data, it becomes burdensome to several industrial and research applications [13]. Nowadays, the missing data imputation is cumbersome [8] in the mixed-attribute data sets since there is no algorithm or technique to fill the mixed value as categorical and numerical data.

The finding of missing attributes in the datasets is taken as problem in this paper. Also, the challenge of finding the missing attributes is that it should preserve the original data characteristics without losing of data originality.

2.2 Challenges

Data imputation for medical diagnosis [14] is the challenge one since some error occurs. It is caused by less equipment to diagnose the disease, which leads to providing the incorrect test results of certain patients.

The major challenge is to impute both categorical and numerical data since every dataset contains both continuous data and discrete data. In the medical domain, the patient's information is stored in both data such as age, height, weight, gender, etc. [15].

The challenge of using optimization algorithm for missing data imputation since it is used to determine the optimal value to determine the imputed data. This leads to significantly fill the data where the data is missed in the dataset.

3. Proposed Methodology: Missing data imputation method of constraint-based hybrid prediction model using grey fuzzy neural network and WLI fuzzy clustering

The ultimate aim of this paper is to design and develop the hybrid prediction model to impute the missing data in the dataset. Normally, the dataset poses categorical and numerical data where the data is missed in the attribute. The main challenge is to fill the data in both categorical and numerical data of the attribute. In order to achieve this objective, the constraint-based hybrid prediction model using WLI fuzzy clustering and the grey fuzzy neural network is proposed. Initially, the input dataset is undergone for the WLI fuzzy clustering mechanism in which the centroids are obtained. Due to the averaging process of the centroid, the data in the missing attribute is filled. Consequently, the input data is fed into the training algorithm. Thus, the training algorithm is newly designed by both grey wolf Optimizer (GWO) and Adaptive Neuro-Fuzzy Inference System (ANFIS). Figure 1 depicts the block diagram of proposed methodology.

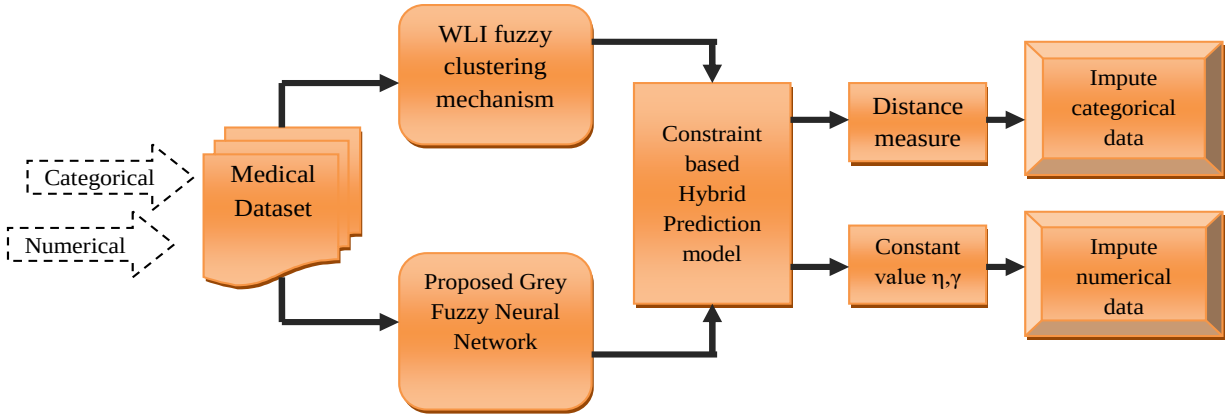


Figure 1. Block diagram of proposed methodology

3.1 Mixed database (Medical data)

Due to missing data in the medical data, the patient information gets lost. Since the bias measurement and incorrect entry of the patient's data, the missing value incurs. In the medical dataset, every attribute belongs to patient's age, gender, weight, height, etc. in which the data is either to be continuous or discrete. The core intent of this work is to impute both the categorical and numerical medical data using proposed method. Consider B is the input dataset includes n number of data objects and a number of attributes. Thus, the input data is defined as

$$B = b_{mn}(i); \quad 1 \le i \le o \tag{1}$$

Where, m and n represent the total number of attributes and data objects. Then, the input corpus contains d it represents the number of discrete data and c indicates the number of continuous data.

3.2 WLI fuzzy clustering for medical data

Generally, the clustering is defined as the process to group the homogenous data object and discriminated by the different number of clusters. It is widely used in data mining application where the missing data is detected to impute by the estimated value. Here, we utilize the WLI fuzzy clustering mechanism [16] to estimate the imputed data with the aid of mean value of its related attribute. The Cluster Validity Index (CVI) is the major concern in this clustering mechanism for the evaluation measure. Thus, the clustering is performed by the cluster validity index and average difference. The following steps are described the clustering mechanism to impute the missing value.

Step 1: Fuzzy compactness

The compactness and cluster separation are the two main key factors in the clustering mechanism. Initially, the categorical and numeric data is used to group the similar data objects. Subsequently, the weighted distance and fuzzy cardinality are the prerequisites to evaluate the compactness of the data object.

- Distance measure: The distanceis evaluated between the data object and cluster. Due to less distance measurement, we can achieve the vigorous compactness of the cluster. Thus, the fuzzy weighted distance [16]is expressed as:

$$d = \lambda_{xz} \| b_x - c_z \| \tag{2}$$

where, b is the input data, c defines the cluster and λ denotes the membership function.

- Membership function: The fuzzy cardinality or the membership function is used to impute the data. The membership function is employed to define the degree of truth mapping every data object with the centroid.

$$\sum_{i=1}^{V} \lambda_{xz} \tag{3}$$

- Finally, the fuzzy compactness of the WLI clustering mechanism is evaluated by the ratio of fuzzy weighted distances to the membership function of every data object. Thus, the fuzzy compactness of all cluster is determined by,

$$WL_f = \sum_{z=1}^{m} \left(\frac{\sum_{x=1}^{n} \lambda_{xz}{}^2 \| b_x - c_z \|^2}{\sum_{x=1}^{n} \lambda_{xz}} \right) \tag{4}$$

Step 2: Cluster separation

The prime characteristic of WLI fuzzy clustering mechanism is to validate the clustered index to find out the value used for data imputation.

- The $\frac{m(m-1)}{2}$ Distance is calculated between m centroids where we can estimate the minimum and median distance. The MIN is the term used to find the minimum distance among all the $\frac{m(m-1)}{2}$ distances. On the other hand, the MED is

defined as the median value of all $\frac{m(m-1)}{2}$ distance and $\left(\frac{m(m-1)/2}{2}\right)^{th}$ distance. Depends on the distance pair of centroid, the separation of cluster is represented by,

$$WL_s = \frac{1}{2}\left(\min_{x\neq y}\left\{\left\|c_x - c_y\right\|^2\right\} + median_{x\neq y}\left\|c_x - c_y\right\|^2\right) \quad (5)$$

Step 3: Cluster Validity Index

Since it contains the fuzzy compactness and separation, the CVI is employed to generate the different cluster number. Thus, from the medical dataset, the homogenous data are grouped together differentiates by the heterogeneous data. Hence, the WLI [16] is computed by the averaging process is calculated as:

$$WLI(m) = \frac{WL_f}{2\times WL_s} \quad (6)$$

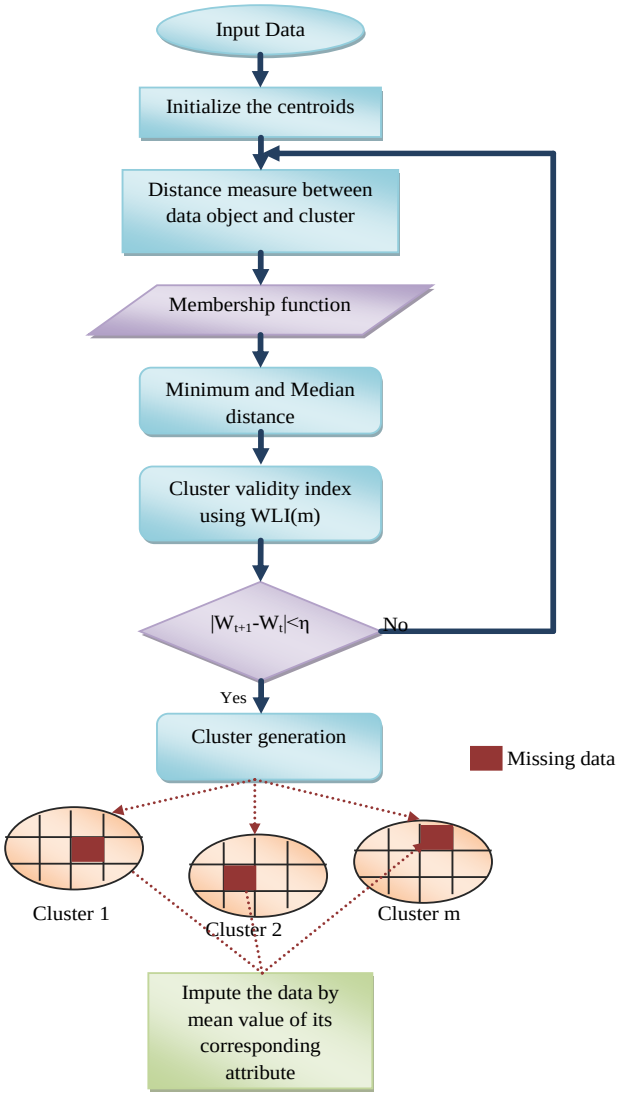


Figure 2. Flow Chart of WLI fuzzy clustering mechanism

Step 4: Threshold

The clustering mechanism is used to generate the clusters according to the number of iterations which is defined by the user-given threshold value. Here, η is represented as the user-given threshold value. Thus, the difference of cluster index between two iterations is less than the threshold value; theclusteris formed. Or else, the new centroid is generated to group the data.

$$\left\|W_{t+1} - W_t\right\| < \eta \quad (7)$$

Step 5: Mean computation for data imputation

Once the clusters are formed from the input medical dataset, the value is calculated for the missing data. In every cluster, the data in the missing attribute is used to measure the mean value. Then, the obtained value is utilized to impute where the data is missed in the input dataset. The mean value is computed by,

$$b^t_{ij(WLI)} = \frac{1}{n_c}\sum_{k=1}^{n_c} b_{kj} \quad (8)$$

where, b_{kj} is the data of j^{th} attribute in the k^{th} data object and also, n_c represents the total number of data in the corresponding missing attribute. Therefore, figure 2 represents the flow chart representation of the WLI fuzzy clustering mechanism.

3.3 Proposed Grey fuzzy neural network

The grey fuzzy neural network is newly designed with the aid of [17] grey wolf Optimizer (GWO) and adaptive neuro-fuzzy inference system (ANFIS) [18]. In ANFIS, the training algorithm is altered with the grey wolf optimizer algorithm. The grey wolf is one of the efficient optimization algorithms to find out the optimal weights to train the medical data in the training phase. On the contrary, the ANFIS is developed by the two main factors, which are neural network and fuzzy logic. Therefore, the core intent of proposed grey fuzzy neural network is to find out the optimal weight using grey wolf optimizer for learning the process.

Adaptive Neuro-Fuzzy Inference System (ANFIS)

The ANFIS is defined as a multilayer feed-forward neural network where the node performs the function as an activation function provide the desired output of the input training data. The ANFIS [18] model is sculpted by the fuzzy system and neural network in which the weight function plays a vital role in the training algorithm. The main advantage of ANFIS over other training algorithm is that has smoothness property, mitigates the search dimensions, attain more learning capability, etc. Figure 3 shows the architecture of the ANFIS network model.

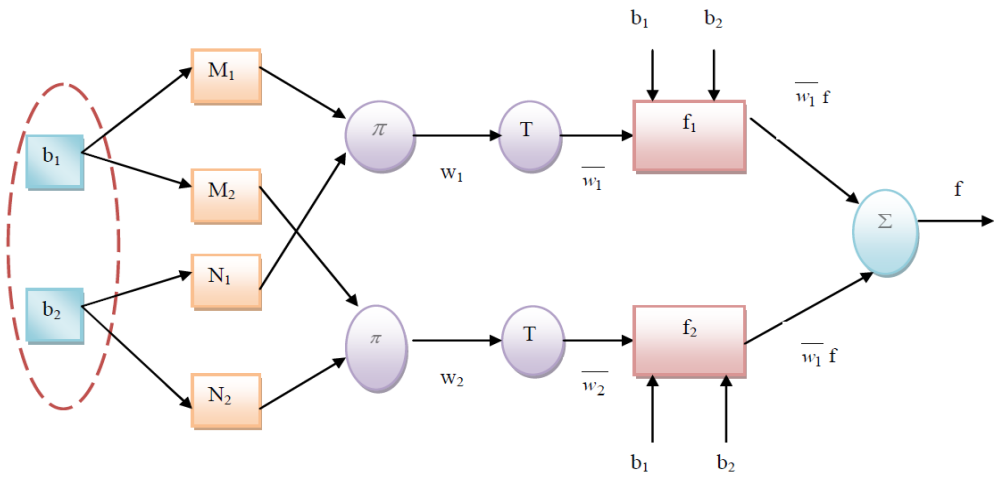


Figure 3. Architecture of ANFIS

In figure 3, the ANFIS [18] structure contains a five-layered structure in which g number of input data points are fed as input to determine the optimal value for missing attribute. It constitutes the antecedent and consequent parameters to train the ANFIS network and updates the input data. Initially, the fuzzy inference system contains two inputs b_1 and b_2 which leads to generating the fuzzy-if-then rules are defined by,

Rule 1: If b_1 is M_1 and b_2 is N_1, then

$$f_1 = p_1 b_1 + q_1 b_2 + r_1 \quad (9)$$

Rule 2: If b_1 is M_2 and b_2 is N_2, then

$$f_2 = p_2 b_1 + q_2 b_2 + r_2 \quad (10)$$

where, M_1,M_2 and N_1,N_2 represents the membership function and p, q, r are the antecedent parameters.

Layer 1:This layer is known as the fuzzification layer in which the input node behaves as an activation node caters the membership function of input data. Thus, the membership function plays a vital role in the ANFIS structure. Thus, the output of layer 1 is expressed by,

$$OL_i^1 = \lambda_{M_i}(b_1); \quad when \quad i = 1,2 \quad (11)$$

$$OL_i^1 = \lambda_{N_i}(b_2); \quad when \quad i = 3,4 \quad (12)$$

where, λ_{M_i} and λ_{N_i} defines the membership function (MF), which is derived as:

$$\lambda_{M_i}(b) = \frac{1}{1 + \left\{ \left(\frac{b - r_i}{p_i} \right)^2 \right\}^{q_i}} \quad (13)$$

where, p_i, q_i, and r_i indicate the parameters to design the membership function. But, the major drawback of the membership function is difficult to compute the arithmetic operations. Also, due to the changes of this parameter, it is critical to determine the optimized value. In order to achieve the optimized parameter, we intend to incorporate the Grey Wolf Optimizer (GWO) optimization algorithm into this layer.

Proposed GFNN: The proposed GFNN method comprises of both Grey Wolf Optimizer [17] and ANFIS [18] network model. The core intent of our proposed method is to find out the optimal membership parameters. Thus, the proposed model exploits this parameter to train the missing attribute of the input dataset. The proposed GFNN is apparently deliberated below.

a) Solution Encoding: The solution encoding is the major aspect in the optimization algorithm. Here we encode the solution with the aid of membership parameters of each input data. Here, our proposed GFNN exploits two input data b_1 and b_2 which contains the twelve numbers of parameters which get optimized by the grey wolf optimizer algorithm.

b) Fitness function: Once the solution is encoded for optimization, the fitness value is evaluated for every search agents. Here, the fitness function is calculated by the exponential function which is derived by,

$$F = -20 \times \exp\left(-0.2 \times \sqrt{\frac{\sum (V^2)}{d}} \right) - \exp\left(\frac{\sum (\cos(2\pi V))}{d} \right) + 20 \quad (14)$$

where, d represents the dimension of the solution.

c) Algorithmic Elucidation: The main characteristic of this optimization algorithm [17] is tracking, chasing and approaching the target (parameter). In general, the four wolves are alpha (α), beta (β), delta (δ) and omega (ω)grey wolves are used. The alpha wolf is the major concern to encircle and attack the target (premise parameter). The beta and delta are defined as the second and third best solution is followed by the alpha. The encircling and hunting are the two prerequisites in GWO which are described below.

i) Encircling: Initially, the solution is encircled by grey wolves. The mathematical formulation is expressed by,

$$U = \left| Y \cdot V_p(t) - V(t) \right| \quad (15)$$

Where t is the current iteration, then, the position of the wolf is updated by the next iteration is given by,

$$V(t+1) = V_p(t) - X \cdot U \quad (16)$$

where, X and Y indicate the coefficient vectors, V(t) defines the position vector of the grey wolf and $V_p(t)$ is the position vector of the target (parameter).

ii) Hunting: The grey wolves have the tendency to reach the target with the aid of position update. Thus, the position is updated iteratively to determine the optimal solution. The standard equation exhibits the position update of search agents. The first three wolves are mainly used for hunting behaviour in search space. It is formulated [17] with respect to three best search agents are

$$V(t+1) = \frac{V_1 + V_2 + V_3}{3} \quad (17)$$

where, V_1, V_2, and V_3 denote the three best agents which are expressed below.

$$V_1 = V_\alpha - X_1 \cdot (U_\alpha), \quad V_2 = V_\beta - X_2 \cdot (U_\beta), \quad V_3 = V_\delta - X_3 \cdot (U_\delta) \quad (18)$$

and

$$U_\alpha = \left| Y_1 \cdot V_\alpha - V \right|, \quad U_\beta = \left| Y_2 \cdot V_\beta - V \right|, \quad U_\delta = \left| Y_3 \cdot V_\delta - V \right| \quad (19)$$

The coefficient vector X and Y is considered as the key factor since it has the h parameter which is decreased from 2 to 0. Due to this reduction, we can determine the optimized parameter value. The X and Y are defined as:

$$\begin{aligned} X &= 2h \cdot z_1 - h \\ Y &= 2 \cdot z_2 \end{aligned} \quad (20)$$

where, z_1 and z_2 are random vectors ranges between 0 to 1. Finally, the GWO optimization algorithm attains the optimal parameter value to design the membership function in ANFIS network. Thus, p^g, q^g and r^g are the obtained parameter using GWO, termed as the Grey Fuzzy Neural Network (GFNN). The membership function for the layer 1 is derived by,

$$\lambda_{M_i}(b) = \frac{1}{1 + \left\{ \left(\frac{b - r_i^g}{p_i^g} \right)^2 \right\}^{q_i^g}} \quad (21)$$

The parameter in this functionis named as the GWO based premise parameter.

Layer 2:This layer is known as the rule layer. It is used to generate the rules strengthens our proposed grey fuzzy neural network. It is mainly used to determine the weights between the layers to represent the fuzzy sets. Thus, every node in rule layer caters the weight parameter for the subsequent layer. The output is given as below.

$$OL_i^2 \Rightarrow W_i = \lambda_{mi}(M) \cdot \lambda_{nj}(N); \quad i = 1,2,3,4 \quad (22)$$

Layer 3: Once the weight is computed, then it is undergone for the normalization purpose. Therefore, this layer is termed as the normalization layer. Each node in this layer is denoted by T. Due to the number of layer is equal to the number of rules, this layer is performed by the summation of two weights from the previous layer for the exploitation of fuzzy sets. In other words, the i^{th} rule strength is divided by the sum of four firing strengths. It is also known as the normalized firing strengths. Thus, the output of normalization layer [22] is

$$OL_i^3 \Rightarrow \overline{W_i} = \frac{\sum_{i=1}^{4} W_i}{W_1 + W_2 + W_3 + W_4} \quad (23)$$

Layer 4:The node in this layer is represented in square form where it poses normalized weight value and two input variables. Finally, this layer provides the trained output data to impute in the missing attribute. On the contrary to the premise parameter, this layer is used to generate the output with the aid of consequent parameters or linear parameters. In this layer, the normalized firing strength is multiplied with i^{th} order polynomial function. It is formulated by,

$$OL_i^4 = \overline{W_i} f_i = \overline{W_i}(x_i m + y_i n + z_i) \quad (24)$$

This layer is called the defuzzification layer since it has the consequent parameters x_i, y_i and z_i

Layer 5:This layer is termed as the sum layer since it contains single node where the data is computed by the summation of all incoming data.

$$OL_i^5 \Rightarrow \sum_i \overline{W_i} f_i = \frac{\sum_{i=1}^{4} W_i f_i}{W_1 + W_2 + W_3 + W_4} \quad (25)$$

Thus, our proposed grey fuzzy neural network constitutes premise and consequent parameters. Initially, the premise parameters are optimized to fix the values using GWO algorithm. Once the premise parameters are fixed, then the final output [18] is expressed by the combination of the linear parameter.

$$I = (\overline{W_i} b_1) x_i + (\overline{W_i} b_2) y_i + (\overline{W_i}) z \quad (26)$$

3.4 Data imputation using constraint-based hybrid prediction model

The input medical dataset consists of both categorical (discrete) data and numeric data (continuous). Both data is missed due to some technical error, bias information, etc. The advantage

of our proposed model over other existing technique is to impute both continuous and discrete data. Here, the hybrid prediction model is developed by the combination of WLI fuzzy clustering mechanism and proposed grey fuzzy neural network.

- The numeric data is defined the continuous data exhibits the value for age, weight, height, etc. Thus, the categorical data is imputed in the input medical dataset is determined by,

$$B'_{mn}(N) = \eta \cdot B_{mn(WLI)} + \gamma \cdot B_{mn(GFNN)} \quad (27)$$

where, η and γ represents the constant value and $B'_{mn}(N)$ is the imputed numerical data of m[th] attribute in the n[th] data object.

- Then, we also intend to impute the discrete data in the mixed database using WLI fuzzy clustering and proposed GFNN network. Thus, both methods acquire the desire categorical output data when the categorical missing attribute is given as input. The input dataset contains d discrete data. At first, the distance is computed between the desired output and input categorical data. Then, the minimum distance is taken out to impute the categorical data in its corresponding missing attribute. It is formulated as below.

$$B'_{mn(WLI)}(C) = \min_{i \in 0,1} D\left(B_{mn(WLI)}, b^i_{mn}\right) \quad (28)$$

where, $B_{mn(WLI)}$ is the output data of WLI fuzzy clustering mechanism and b'_{mn} is the categorical input data and $B'_{mn(WLI)}$ provides the categorical imputed data. Similarly, the minimum distance is figured out by the predicted value of proposed grey fuzzy neural network $B_{mn(GFNN)}$.

$$B'_{mn(GFNN)}(C) = \min_{i \in 0,1} D\left(B_{mn(GFNN)}, b^i_{mn}\right) \quad (29)$$

- After the two imputed values are computed, it is then subjected to following constraints. If both the data are same, then we impute the categorical data directly.

$$B'_{mn} = \{B'_{mn(GFNN)} \quad when \quad B'_{mn(WLI)} = B'_{mn(GFNN)} \quad (30)$$

On the other hand, we plan to calculate the frequency of obtained imputed data for both WLI and proposed GFNN. Based on the frequency value, the discrete data is imputed in the mixed medical dataset.

$$B'_{mn} = \begin{cases} B'_{mn(WLI)}; & fr\left(B'_{mn(WLI)}\right) > fr\left(B'_{mn(GFNN)}\right) \\ B'_{mn(GFNN)}; & Otherwise \end{cases} \quad (31)$$

4. Results and Discussion

This section demonstrates the experimental results and performance analysis of proposed GFNN model. It is then validated through MSE and RMSE parameters. The performance analysis is also compared with the existing methods.

4.1 Experimental Setup

a) Dataset Description: Here, we utilize two databases from UCI (UC Irvine) Machine Learning [19] for our experimentation to impute the medical data in the missing attribute. *Heart disease (Dataset 1):* The data is collected by the patient who was undergone for the heart diagnosis. This dataset consists of 76 attribute values which include the patient's age, sex, patient's ID, etc. *Pima Indian diabetes (Dataset 2):* This dataset is a collection of medical diagnostic reports of 768 examples from a population. The sample of this dataset consists of eight attribute values. The database now available in the repository has 512 training samples and 256 testing samples.

4.2 Performance Analysis

a) Analysis of dataset 1

The figure 4 depicts the MSE performance analysis of the proposed model. Depends on the different values of R, the proposed method achieves the lower error value. While using 30% of missing data, the proposed GFNN attains 10.68, 10.64, 10.73, 10.88 and 11.93 MSE for R=0.1, 0.2, 0.3, 0.4 and 0.5. While using R=0.1 of the proposed model, the mean square error obtains 5.95 which is then increased to 14.95 while increasing the R-value that is shown in figure 4. Consequently, the performance of root mean square error is represented in figure 5. The RMSE is the error measure between the actual and predicted output. The lower value of RMSE leads to provide the better performance. Initially, 2.45MSEis obtained which is gradually increased to 3.87with regard to different R value. When the percentage of missing data is 40, the error value 3.58is obtained for R=0.1, 3.56, 3.57 and 3.62 attains by the proposed method which is demonstrated in figure 5.

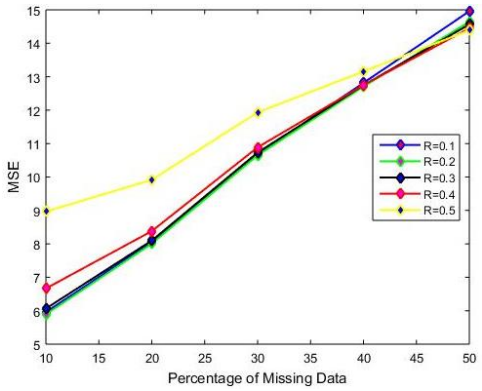


Figure 4. Performance analysis of dataset 1 using MSE

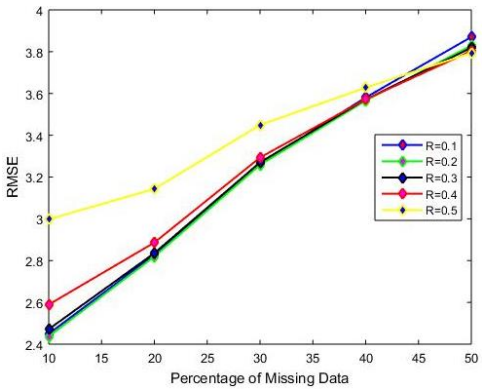
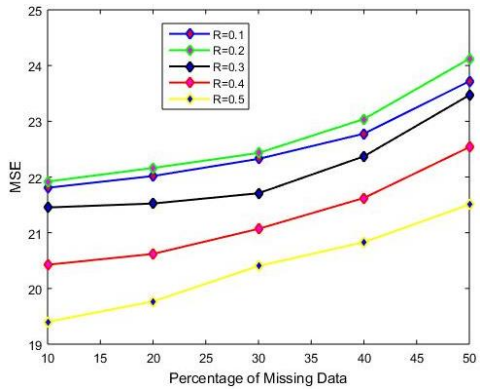


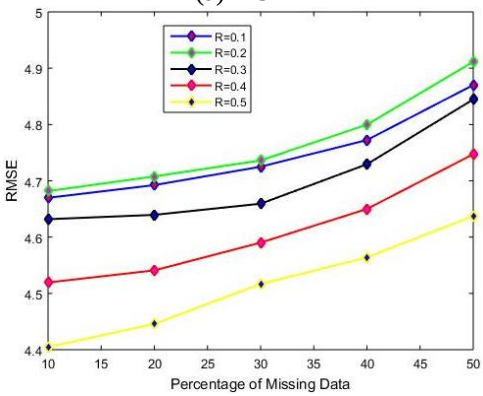
Figure 5. Performance analysis of dataset 1 using RMSE

b) Analysis of dataset 2

The performance analysis for the dataset 2 using proposed GFNN method is represented in figure 6. Here, we consider the dataset 2 as Pima Indian Diabetes dataset. The figure 6.a shows the MSE performance analysis. According to the percentage of missing data, the proposed method acquires minimum error value. While using R=0.4, the proposed GFNN method attains 20.42, 20.61, 21.06, 21.62 and 22.53with regard to the various percentage of missing data. When using the 40% of missing data in our proposed system, it provides 22.77, 23.03, 22.37, 21.62 and 20.82is achieved through various R values. Similarly, the figure 6.b depicts the RMSE performance analysis. When using ten percentage of missing data, the error value 4.66 and 4.68is obtained for R=0.1 and 0.2, then, 4.63 and 4.51 achieved by R=0.3 and 0.4 and finally, the 4.40is acquired for R=0.5. The 4.63 RMSE value is achieved initially and then it is moderately increased to 4.65, 4.72 and 4.84 by the proposed GFNN model that is shown in figure 6.b.



(a) MSE



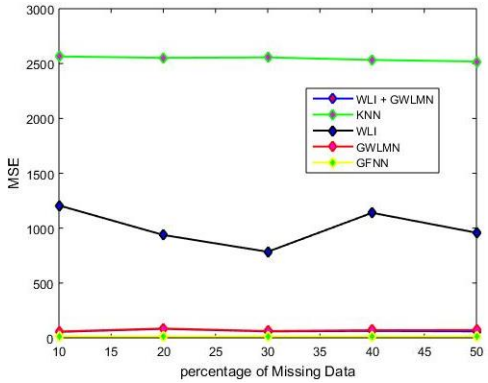
(b) RMSE

Figure 6. Performance analysis of dataset 2

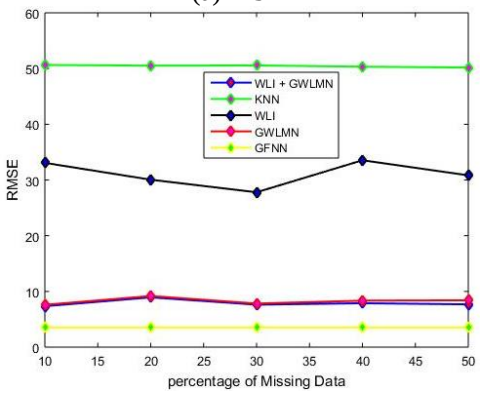
4.3 Comparative performance analysis

a) Comparative performance for dataset 1

Figure7 depicts the comparative performance for dataset 1. Based on the percentage of training data, the proposed method ensures the better performance when compared with the existing methods. The figure 7.a shows the MSE comparative performance for Heart disease dataset. When the percentage of missing data in the input dataset is 30, the existing WLI+GWLMN method acquires 1.20 mean square error, 0.937 and 0.785 MSE achieved by KNN and WLI algorithm and finally, the GWLMN method attains 1.139 error. Subsequently, the comparative RMSE performance is demonstrated in figure 7.b. The existing WLI clustering mechanism attains 33.07, 30.05, 27.79, 33.53 and 30.83 RMSE value is obtained based on the percentage of missing data. The other existing GWLMN method acquires 7.57 RMSE which is then gradually increased to 8.408. But, our proposed grey fuzzy neural network achieves lower 3.52 RMSE ensures to impute the data in the input dataset effectively.

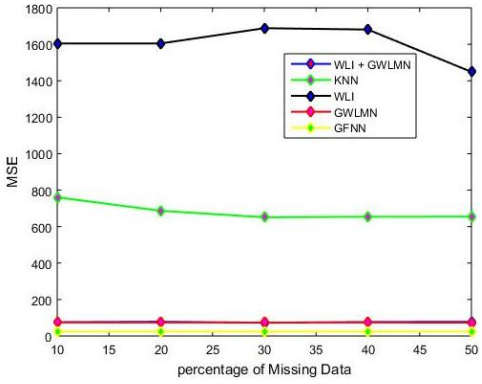


(a) MSE

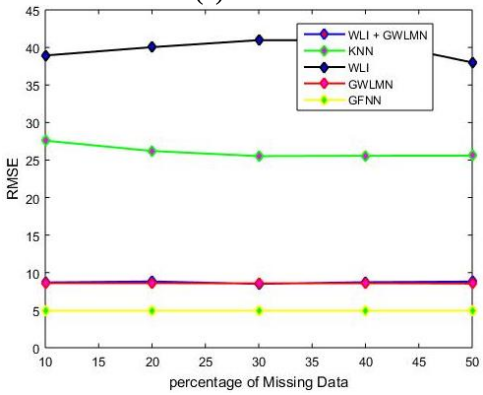


(b) RMSE

Figure 7. Comparative performance for dataset 1



(a) MSE



(b) RMSE

Figure 8. Comparative performance for dataset 2

Table 1. Comparative performance of different methods

Dataset	**Dataset 1**		**Dataset 2**	
Methods	MSE	RMSE	MSE	RMSE
WLI+ GWLMN	1.6044	38.9312	2.5650	50.646
KNN	1.6446	40.9387	2.5523	50.520
WLI	1.6883	40.9730	2.5564	50.560
GWLMN	1.6809	40.9675	2.5335	50.333
GFNN	1.4488	38.0003	2.5186	50.185
GRAANN	-	-	-	23.89
PSOAANN	-	-	-	21.72
PSOAAWNN	-	-	-	23.68
RBFAANN	-	-	-	32.28

b) Comparative performance for dataset 2

The comparative performance for dataset 2 is shown in figure 8. The figure 8.a depicts the MSE comparative performance analysis. The existing KNN algorithm achieves 0.76, 0.686, 0.651, 0.653 and 0.654 mean square error depends on the percentage of missing data. The proposed GFNN method attains minimum 0.0244 MSE value when compared to the existing algorithm. The proposed method sustains the 0.024 value for 10 to 50 percentage of missing data. Consequently, the figure 8.b represents the comparative performance analysis for RMSE. The RMSE is used to determine the error between the actual input data and desired output data in terms of the square root. While using 20 percentages of missing data, the existing WLI+GWLMN achieves 40.03; the KNN algorithm attains 26.2; the 8.826 and 8.611 are obtained by WLI and GWLMN algorithm. But, our proposed GFNN method obtains lower 4.93 RMSE when compared to the existing methods.

Table 1 demonstrates that the comparative performance of different methods. Here, the performance is analyzed through MSE and RMSE values by two data sets: Heart disease (Dataset 1) and Pima Indian diabetes (Dataset 2). Moreover, the GRAANN, PSOAANN, PSOAAWNN and RBFAANN methods for dataset 2is referred from [20].Table I depicts the performance of proposed model GFNN acquires low MSE (1.4488, 2.5186) and RMSE (2.5186, 50.185) values while comparing with other existing methods for both the datasets. Hence, the proposed method performs better than the conventional methods due to the hybrid behaviour of the clustering and neural network.

5. Conclusion

In this paper, we proposed the grey fuzzy neural network and WLI fuzzy clustering mechanism for missing data imputation. Here, we considered the mixed database which includes both categorical and numerical data. Firstly, the WLI fuzzy clustering mechanism was utilized in which the medical data

were grouped into different clusters. It was used to evaluate the mean value of the missing attribute to impute the data. Secondly, the input dataset underwent for the proposed GFNN method. The novel Grey Fuzzy Neural Network (GFNN) was designed and developed by integrating the ANFIS and grey wolf Optimizer (GWO). The proposed method was also used to provide the imputed data. Thirdly, the constraint-based hybrid prediction model composed of both WLI fuzzy clustering and proposed GFNN method. Thus, the experimental results were evaluated and performance was analysed through MSE and RMSE metrics. The proposed GFNN method achieved the lower 1.6 MSE and 38.93 RMSE error for the missing data imputation. In future, the respective experimentation will be expanded through advanced hybrid methods.

References

[1] C.O. Galan, F.S. Lasheras, F. J. Juez and A.B.Sanchez, "Missing data imputation of questionnaires by means ofgeneticalgorithms with different fitness functions",*Journal of Computational and Applied Mathematics,* Vol. 311, pp. 707-717, 2016.

[2] M.Amiria and R. Jensen, "Missing data imputation using fuzzy-rough methods", *Neurocomputing,* Vol. 205, pp. 152-164, 2016.

[3] M. MostafizurRahman and D.N.Davis, "Machine Learning-Based Missing Value Imputation Method for Clinical Datasets", *IAENG Transactions on Engineering Technologies,* vol. 229, pp. 245-257, 2013.

[4] P.J.Garcia-Laencina and P.H. Abreu, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values", *Computers in Biology and Medicine,* Vol. 59, pp. 125-133, 2015.

[5] J.Tian, B. Yu, D. Yu and S. Ma, "Missing data analyses: a hybrid multiple imputation algorithmusingGrey System Theory and entropy based on clustering", *Applied Intelligence,* Vol. 40, No. 2, pp. 376-388, 2014.

[6] X.P. Zhang, A. ShaharyarKhwaja and A.Anpalagan, "Multiple Imputations Particle Filters: Convergence and Performance Analyses for Nonlinear State Estimation with Missing Data", *IEEE Journal of Selected Topics in Signal Processing,* Vol. 9, No. 8, pp. 1536-1547, 2015.

[7] T. H. Lin, "Missing Data Imputation inQuality-of-Life Assessment", *Pharmaco Economics,* Vol. 24, No. 9, pp. 917-925, 2006.

[8] X. Zhu, S. Zhang, Z.Zhang, and Z.Xu, "Missing Value Estimation for Mixed-Attribute Data Sets", *IEEE Transactions on Knowledge and Data Engineering,* Vol. 23, No. 1, pp. 110-121, 2011.

[9] S. Zhang, "Shell-neighbor method and its application in missing data imputation", *Applied Intelligence,* Vol. 35, No. 1, pp. 123-133, 2011.

[10] W. Wei and Y. Tang, "A generic neural network approach for filling missing data in data mining", *In proceedings of IEEE International Conference on Systems, Man and Cybernetics,* pp. 862-867, 2003.

[11] Y. Zhang and Y. Liu, "Data Imputation Using Least Squares Support Vector Machines in Urban Arterial Streets", IEEE *Signal Processing Letters,* Vol. 16, No. 5, pp. 414-417, 2009.

[12] R. Pan, T. Yang, J. Cao, K. Lu and Z. Zhang, "Missing data imputation by K nearest neighbours based on grey relational structure and mutual information", *Applied Intelligence,* Vol. 43, No. 3, pp. 614-632, 2015**.**

[13] J. Luengo, J. A. Saez and F. Herrera, "Missing data imputation for fuzzy rule-based classification systems", Soft computing, Vol.16, No. 5, pp. 863-881, 2012.

[14] J. Luis, S.Gomez, A.F Vidal and M.Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation", *Neurocomputing,* Vol. 72, No. 7-9, pp. 1483-1493, 2009.

[15] A.Purwar and S.K. Singh, "Hybrid Prediction Model with missing value Imputation for medical data", *Expert Systems with Applications,* Vol. 42, No. 13, pp. 5621-5631, 2015.

[16] C-H. Wu, C.S.Ouyang, L.W Chen, and L.W.Lu, "A New Fuzzy Clustering Validity Index with a Median Factor for Centroid-based Clustering", *IEEE Transactions on Fuzzy Systems,* Vol. 23, No. 3, pp. 701 - 718, 2014.

[17] S.Mirjalili, S. M. Mirjalili and A. Lewis, "*Grey Wolf Optimizer", Advances in Engineering Software,* Vol. 69, pp. 46–61, 2014.

[18] N.Walia, H. Singh and A.Sharma, "ANFIS: Adaptive Neuro-Fuzzy Inference System- A Survey", *International Journal of Computer Applications,* Vol. 123, No. 13, pp. 32-38, 2015.

[19] UC Irvine Machine Learning Repository from http://archive.ics.uci.edu/ml/datasets.html".

[20] V.Ravi, M.Krishna,"A new online data imputation method based on general regression auto associative neural network", *Neurocomputing,* Vol.138, No. 22,pp. 106–113, August 2014.