Uncertain Data Analysis using Gray Set based Multiple Imputation with Penalized Optimization Algorithm

G.V.Suresh[1]
Research Scholar
Jawaharlal Nehru Technological University, Hyderabad
Hyderabad, Telangana., INDIA.
vijaysuresh.g@gmail.com
Dr. E. Sreenivasa Reddy[2]
Professor
University College of Engineering,
Acharya Nagarjuna University, A.P., INDIA
edra92@gmail.com

Article Info
Page Number: 252 – 267
Publication Issue:
Vol. 71 No. 3 (2022)

Abstract
Many real-world applications face an uncanny predicament of uncertainty plaguing the available knowledge.In the normal run of things, some uncertainties arise due to the prevalence of incorrect measurements and inaccurate decision-making, resulting in unreliable data transmission and data storage.Furthermore, the inevitable randomization that occurs during the physical data generation and gathering process contributes to the aforementioned issue.However, data imperfection issues create lots of trouble for people while working on real-world data wherein data insufficiency emerges as the primary source of trouble. Invalid approaches to handle missing data results in biased outcomes, over-confident intervals, and inaccurate inferences. In recent times, multiple imputation has emerged as a potential alternative to the traditional approaches in missing data analysis. This paper proposes a missing data imputation method i.e., Multiple Imputation–Penalized Optimization Algorithm(MIPOA). By removing the missing-valued items, it first divides the non-missing data instances into various clusters. The entropy of the proximal category is then used to calculate the similarity metric for each incomplete instance using gray relational analysis.Our algorithm is superior to previous approaches in terms of validity, according to experiments on UCI datasets.

Keywords: Missing Data, Uncertainty, Grey Sets, Clustering, and Multiple Imputation.

Article History
Article Received: *12 January 2022*
Revised: *25 February 2022*
Accepted: *20 April 2022*
Publication: *09 June 2022*

1. INTRODUCTION

The available knowledge tends to be uncertain in many real applications. Measurement and decision errors, inaccurate data transmission and data storage, as well as the random nature of data generation and gathering, may lead to such uncertainty. There are many faces for uncertainty i.e., inconsistency, imprecision, ambiguity, incompleteness, vagueness, unpredictability, noise, unreliability etc. It is significant to handle the uncertainty in various data mining applications to achieve acceptable results [1][2].Even when faced with uncertainty, human professionals are usually able to come at reasonable conclusions about their work.There are many ways to deal with uncertainty, including rough set theory, fuzzy set theory, evidence theory, Bayesian theory, statistical functions, and the certainty factor [9].But all these theories have some advantages and disadvantages[21-23]. These theories are somehow valid for some specific purpose only i.e., each technique is applicable to a particular problem only. As a result, a hybrid strategy must be developed that incorporates two or more theories.Shortcomings in individual theories can be mitigated by integrating various methods.

1.1. IMPERFECT INFORMATION SYSTEM

An incomplete information system is one that contains missing values. Objects in incomplete information systems, such as Table.1 [8], may have numerous unknown attribute values. Unknown values are denoted by special symbol "*". In an imperfect information system $\mathbf{I=(U,A)}$, let $\mathbf{t}^{x}_{a,i} \subseteq \mathbf{V}_{a}$ be the $\mathbf{i}^{th}$ set of overall $\mathbf{s}$ possible value sets of $\mathbf{x}$ on $\mathbf{a}$ and $\left\{\mathbf{p}^{x}_{a,i}\right\}>\mathbf{0}$ be its probability. Then the pair $\left(\mathbf{T}^{x}_{a},\mathbf{P}^{x}_{a}\right)$,where $\mathbf{T}^{x}_{a}=\left\{\mathbf{t}^{x}_{a,i} \mid \mathbf{1}\le \mathbf{i}\le \mathbf{s}\right\}$, $\mathbf{P}^{x}_{a}=\left\{\mathbf{p}^{x}_{a,i} \mid \sum_{i}\mathbf{p}^{x}_{a,i}=\mathbf{1}\right\}$,represents imperfect values of object $\mathbf{x}$ and $\mathbf{a}$ In the above, $\mathbf{t}^{x}_{a,i}$ are not necessarily be mutually disjoint[7]. When any set of possible values is a singleton i.e., $|\mathbf{t}^{x}_{a,i}|=\mathbf{1}$., the value is unclear. [11] Some sorts of missing values have a predetermined probability distribution that could be considered uncertainty[11].

Table1Incomplete Information System

U	A_1	A_2	A_3	A_4
x_1	H	Y	Y	Y
x_2	M	Y	*	*
x_3	*	*	Y	Y
x_4	H	*	N	N
x_5	L	Y	*	*
x_6	*	N	N	N

Note: H-High, M-Medium, L-Low; Y-Yes, N-No ;"" represents Missing Data*

When there is just one set of several possible values and the probability of this set is likewise one, a value is imprecise, formally $|\mathbf{T}^{x}_{a}|=\mathbf{1}, \mathbf{p}^{x}_{a,1}=\mathbf{1}$. Two extreme types of imprecision are an exact value and a missing value with no pre-defined probability distribution [10].An exact

value can only be found in a singleton set i.e., $|t^{x}_{a,1}|=1$ [4][5]. Imprecise information could be viewed as missing values without pre-defined probability distribution if the attribute domain contains the whole set of possible values that $t^{x}_{a,1}=V_{a}$. Some attribute values are frequently missing or partial in real-world data sets.Any of the following are possible explanations for such an occurrence: i) The value is not applicable in this situation,ii) was not recorded when the data was taken, or is not specified by users due to privacy concerns in the third case.Problems arise when missing values are present, including decreased efficacy, the difficulty of the system to process data with missing values, and biasing of the results in comparison to the original dataset. Prior to selecting an algorithm, it is important to evaluate the type of missing values in a dataset.

1.2. CATEGORIES OF MISSING VALUES

Classifying missing data into ignorable or non-ignored categories. In this scenario, the probability of missing data is based solely on the seen data, not the missing data itself. Non-ignorable, the probability of missing data depends on the absence of data, not on the presence of data [13]. Let $\mathbf{X} = (x_{ij}) : (n \times k)$ rectangular data set without missing values. If $M=\left(m_{ij}\right)$ then $m_{ij}=1$ if x_{ij} is missing and $m_{ij}=0$ if x_{ij} is present. Further in line to ignorable and non-ignorable missing data mechanisms classified into three types MCAR, MAR and NMAR [8].

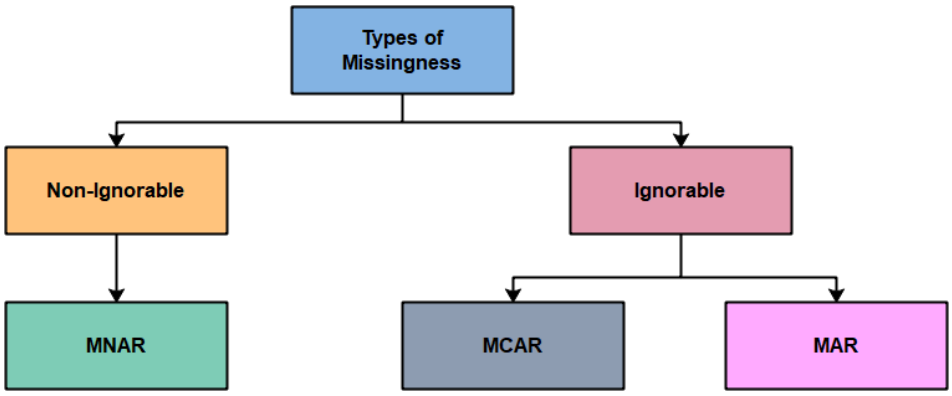


Figure 1 Classification of Missing Data

- **Missing Completely at Random (MCAR):** Missingness mechanism independent of the values of the data $\mathbf{X}$ (Missing- $\mathbf{X}_{miss}$ - or Observed- $\mathbf{X}_{obs}$)[5].
- **Missing at Random (MAR):** Missingness mechanism depends only on $\mathbf{X}_{obs}$, not on $\mathbf{X}_{miss}$
- **Not Missing at Random (NMAR):** Missingness mechanism depends on $\mathbf{X}_{miss}$ [6].

Various techniques are available for treating missing data; a few techniques are described below [3]. Missing data treatment techniques can be classified into three classes, Traditional Approaches and Modern Approaches as shown in Fig.2.

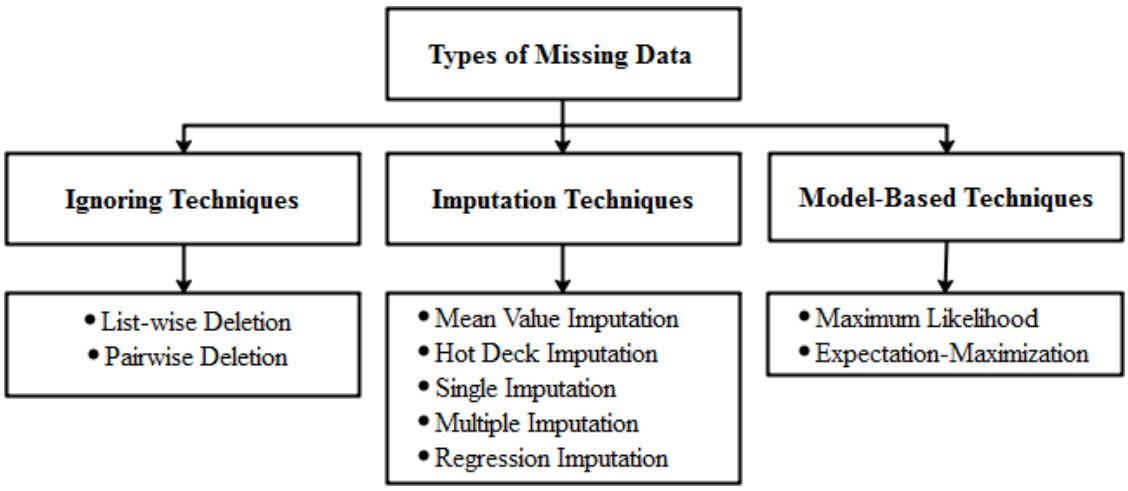


Figure 2 Methods for Handling Missing Data

The presence of incomplete data might potentially cause major issues for researchers. In fact, the treatment of missing data in data analysis can lead to bias and confusing results from a research study, as well as limiting the generalizability of the research findings [5]. Missing data in DM typically causes three types of issues:

• Loss of efficiency.
• Difficulty in processing and analysing data.
• Inequity stemming from discrepancies between missing and comprehensive data.

The rest of the paper is described in the various sections which includes, The Methods and Materials is covered in Section 2. The proposed method is shown in Section 3. The results and evaluation is shown in Section 4. The study conclusion is described in Section 5.

2. Methods and Materials

In this study, data imputation is presented as a problem of filling in the gaps in a data set that would otherwise be meaningless by using statistical methods. (a) using binary data to estimate missing values. (b)GRA is based on attributes rather than specific instances. (c) Instead of merging instances, merge attributes. (d) After each missing attribute element is imputed.Next time we will utilise the new imputation result (imputation by PA) instead of the mean to determine the missing values of a given attribute for the reminders. [14]

Let $\mathbf{X}_{miss}$ denote an incomplete dataset with $\mathbf{n}$ attribute $\mathbf{X}_{miss}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$ and $\mathbf{k}$ instances. Foreach elements of incomplete dataset are defined by $\mathbf{M}_{ij}\mathbf{(i=1,2,3,...,k;j=1,2,...,n)}$, It is divided into two parts: $\mathbf{M}_{ij}=\{\mathbf{m}_{ij}^{obs},\mathbf{m}_{ij}^{miss}\}$ where $\mathbf{m}_{ij}^{obs}$ are the observed values and $\mathbf{m}_{ij}^{miss}$ where are the missing values [15].

$$\mathbf{M}=\begin{bmatrix} \mathbf{m}_{11} & \mathbf{m}_{12} & .. & \mathbf{m}_{1n} \\ \mathbf{m}_{21} & \mathbf{m}_{22} & .. & \mathbf{m}_{2n} \\ \vdots & \vdots & .. & \vdots \\ \mathbf{m}_{k1} & \mathbf{m}_{k2} & .. & \mathbf{m}_{kn} \end{bmatrix}$$

In the case of an incomplete dataset $\mathbf{X}_{\mathbf{miss}}$, a binary matrix $\mathbf{R}$ is constructed by converting each seen value $\mathbf{X}_{\mathbf{miss}}$ to one and each missing value $\mathbf{m}_{\mathbf{ij}}^{\mathbf{miss}}$ to zero. When this binary matrix has the same number of rows and columns as the data matrix $\mathbf{M}$, it is referred to as a matrix of missing data indications.

$$\mathbf{R}_{ij} = \begin{cases} \mathbf{1 \text{ if } m_{ij} \text{ is observed}} \\ \mathbf{0 \text{ if } m_{ij} \text{ is missing}} \end{cases}$$

For example:

$$\mathbf{M}=\begin{bmatrix} 2 & * & 4 \\ * & 7 & 8 \\ 9 & * & * \end{bmatrix} \rightarrow \mathbf{R}=\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

After Enhanced Fuzzy C-Means has allocated a new attribute to the nearest cluster each time (EFCM). Lastly, a single instance was added to the binary matrix and referred to as class (target): $\mathbf{R}_{\mathbf{ij}}=\left(\mathbf{r}_{\mathbf{ij}}\right)_{\mathbf{nK}}$ it is linked to the cluster's data matrix. The Enhanced Fuzzy C-Means (EFCM) starts as follows:

2.1. CLUSTERING STRATEGY

A better goal function is necessary to produce more accurate clustering results. Enhanced Fuzzy C-Means (EFCM), which considerably increases FCM performance by parameter ϕ learning using a prototype.

During each iteration of the learning process ϕ , the strength of the exponential separation between clusters is adjusted.[17] It is possible to calculate the parameter ϕ as follows:

$$\phi=\exp\left[-\min_{i\neq k}\frac{\left\|\mathbf{v}_i-\mathbf{v}_k\right\|^2}{\beta}\right] \quad (1)$$

As a sample variance, the β is normalized term in the preceding equation is used. According to β and ϕ defined by:

$$\beta=\frac{\sum_{j=1}^{n}\left\|\mathbf{x}_j-\bar{\mathbf{x}}\right\|^2}{\mathbf{n}} \text{ where } \bar{\mathbf{x}}=\frac{\sum_{j=1}^{n}\mathbf{x}_j}{\mathbf{n}} \quad (2)$$

It is critical to note that the same value is utilized by all of the data in each iteration, because otherwise an error [18] may occur if the parameter is not modified.In order to replace the common value, a new parameter is introduced, such as adding a weight to each vector or to each dataset point in respect to each cluster.As a result, this weight allows for more accurate categorization, which is particularly useful when dealing with noisy data. The weight is calculated using the following equation [19].

$$W_{ij}=\exp\left[-\frac{\left\|x_j-v_i\right\|^2}{\left[\sum_{j=1}^{n}\left\|x_j-v_i\right\|^2\right]}*c/n\right] \quad (3)$$

(3) W_{ij} Denotes the point's weight j in relation to the class i . The fuzzy and typical partitions are modified using this weight.Both of the following expressions appear in the objective function and are related to one another: There are two forms of probabilistic functions: the fuzzy function, which employs a fuzziness weighting exponent, and the possibilistic function, which employs a conventional weighing exponent. The fuzzy function is the more common type of probabilistic function.There is no need to use such coefficients in the objective function as exhibitors of membership or typicality in the objective function because those coefficients only appear in the objective function as exhibitors of membership and typicality.It is established that when the tendency is toward 1, a new and slightly different relationship is established, which allows for a more rapid decline of the function while simultaneously increasing the membership and typicality, and that when the tendency is toward 0, a more rapid decline of the function while simultaneously increasing the membership and typicality are established.For the purposes of this relationship, it is necessary to include the Weighting exponent as a distance exhibitor in the two objective functions under consideration. The following are the terminologies that can be used to express the EFPCM's aim function:

$$J_{EFPCM}=\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij}^{m}w_{ij}^{m}d^{2m}(x_j,v_i)+t_{ij}^{\eta}w_{ij}^{\eta}d^{2\eta}(x_j,v_i)) \quad (4)$$

$U=\{u_{ij}\}$ represents a fuzzy partition matrix, is defined as:

$$u_{ij}=\left[\sum_{k=1}^{n}\left(\frac{d^2(x_j,v_i)}{d^2(x_j,v_k)}\right)^{2m/(m-1)}\right]^{-1} \quad (5)$$

$T=\{t_{ij}\}$ represents a typical partition matrix, is defined as:

$$t_{ij}=\left[\sum_{k=1}^{n}\left(\frac{d^2(x_j,v_i)}{d^2(x_j,v_k)}\right)^{2\eta/(\eta-1)}\right]^{-1} \quad (6)$$

$V=\{v_{ij}\}$ represents c centers of the clusters, is defined as:

$$v_i=\frac{\sum_{j=1}^{n}(u_{ij}^{m}w_{ij}^{m}+t_{ij}^{\eta}w_{ij}^{\eta})}{\sum_{j=1}^{n}(u_{ik}^{m}w_{ji}^{m}+t_{ik}^{\eta}w_{ij}^{\eta})} \quad (7)$$

The process would be terminated as soon as the number of clusters reached the predetermined threshold.

2.2. IMPUTATION STRATEGY

When it comes to imputation, the process begins as follows: After classifying each instance, calculate the expected information (entropy).

Let's consider the following probability distribution $\mathbf{P=(p_1,p_2,...,p_v)}$ and a dataset $\mathbf{D}$ and define the information carried by the distribution otherwise known as the entropy of $\mathbf{P}$

$$\mathbf{Entropy(m_j)=-\sum_{i=1}^{v} p_i \log_2 (p_i)} \quad (8)$$

Where $\mathbf{p_i}$ is the likelihood of event occurring

$\mathbf{n}$ is number of records

$\mathbf{v}$ is number of clusters $\mathbf{(v \geq 2)}$

Information needed after split $\mathbf{n}$ due to $\mathbf{j}$

$$\mathbf{E\text{-}split=Exp}\left|\left(\sum_{j=1}^{v}\frac{|D_j|}{|D|}\log_2\left|\left(\frac{|D_j|}{|D|}\right)\right|\right)\right|\mathbf{*Entropy(m_j)} \quad (9)$$

The coefficient of difference for the $\mathbf{f^{th}}$ records is calculated in the following step. $\mathbf{t_f=1-I_f\ ;\ f\text{-}1,2,..,n}$

Contrast intensity is applied to the corresponding parameter $\mathbf{t_f}$, $\mathbf{f^{th}}$. The more meaningful an $\mathbf{t_f}$ expression is, the greater its worth.Then calculate the $\mathbf{f^{th}}$ copy's coefficient of weight.

$$\mathbf{w_f=\frac{t_f}{\sum_{f=1}^{n} t_f}} \quad (10)$$

In the first imputation, the mean mode substitution (MMS) is used to replace missing data. An easy-to-understand method would only work if the data was already divided. Then, estimate the missing value attributed to

$$\mathbf{X_{ij}^{miss}=\sum_{q=1,q^1j}^{n} w_q x_{iq}^{miss}} \quad (11)$$

Modified Mean Imputation(MMI) can be used to fill in the gaps in attributes that are missing . Here's a description of the MMI methods:

2.3. MODIFIED MEAN IMPUTATION (MMI)

Mean value substitution is used in the beginning of this algorithm. The algorithm, assuming the original values are erroneous, recalculates the new values based on the Euclidean distance between the missing value records and the remaining records. [20] The records with the shortest Euclidean distance from the missing value record were omitted from the mean value computation.

Algorithm: Modified Mean Imputation (MMI)

1. Begin
2. For **i=1:n**
3. $\mathbf{X}_{obs} \leftarrow \mathbf{M}_i \cap \mathbf{X}_{miss}$ //where is the column of attributes without missing values is the set of missing values in $\mathbf{M}$//
4. Let Δ_i be the mean of $\mathbf{X}_{miss}$
   Replace all the missing elements of $\mathbf{M}_i$ with Δ_i
5. end
6. Let $\mathbf{X}_{imp} = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n\}$

// where $\mathbf{X}_{imp}$ be the approximately imputed data set of $\mathbf{M}$; $\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n$ are the m rows of the data set $\mathbf{M}$.//

7. for **j=1:n**
8. $\delta \leftarrow \mathbf{dist}(\mathbf{X}_{imp}, \mathbf{r}_j)$
9. $\mathbf{J} \leftarrow \mathbf{find}(\mathbf{M}, \Delta(\delta))$

// where δ is the distance matrix $\mathbf{J}$ is the index of elements which are having distance higher than mean $\Delta(\delta)$ .//

10. for **k=1:n**
11. if $\mathbf{X}_{imp}(\mathbf{m,n})$ is originally a missing element
12. begin
13. Let Δ_j be the mean of elements $\mathbf{X}_{imp}(\mathbf{J,n})$
14. $\mathbf{r}_j(\mathbf{k}) \leftarrow \Delta_j$
15. end
16. end
17. end
18. $\mathbf{X}_{imp} \leftarrow$ **Imputed Data Set**

The MIPOA imputation algorithm was utilised instead of the mean imputation technique to determine the imputation of the missing values of certain attributes after each missing element was imputed.

3. MULTIPLE IMPUTATION–PENALIZED OPTIMIZATION ALGORITHM (MIPOA)

A $(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$ random sample is taken from the distribution $\mathbf{f(X|\Theta)}$, which Θ is represented by the vector of parameters, $\mathbf{X}_i = (\mathbf{X}_i^{obs}, \mathbf{X}_i^{mis}), \mathbf{i=1,2,...,n}$ and $\mathbf{X}_i^{obs}$ is the observed and $\mathbf{X}_i^{mis}$ missing data, respectively.

Let $X_i=(X_1,X_2,.....,X_n)$ $X_i^{obs}=(X_1^{obs},X_2^{obs},....,X_n^{obs})$ and $X_i^{mis}=(X_1^{mis},X_2^{mis},...,X_n^{mis})$to

As a way to underline how much a dimension depends on the sample size, we also write Θ and Θ_n identify by $\Theta_n^{(t)}$ the estimate of Θ that is derived during the iteration of the MIPOA method, which we call. The MIPOA approach operates by iterating back and forth between the imputation MI- and post-imputation PO phases.

MI Step: Let m denote the number of imputations. Impute the sample m times by chained equations. For each imputation draw X_i^{mis} from the predictive distribution $h(X_i^{mis}|X_i^{obs}),\Theta_n^{(0)}$ given X_1^{obs} and the current estimate $\Theta_n^{(t)}$.

a. PO Step: Using pseudo-complete data $\bar{X}=X_i^{obs},\bar{X}_i^{mis}$, determine an updated estimate that contributes to the formation of a consistent estimate of

b.

$$\text{c. } \Theta_*^{(t)}=\arg\max_{\Theta} E_{\Theta_n^{(t)}}\left[\log\left\{f_{\Theta}(\vec{x})\right\}\right] \quad (12)$$

where $E_{\Theta^{(t)}}\left[\log\left\{f_{\Theta}(\vec{x})\right\}\right]=\log\left\{f(X_{obs},X_{mis|\Theta})\right\}$, $\int f(X_{obs}|\Theta_*)\,h(\vec{X}_{mis}|X_{obs},\Theta_n^{(t)})dX_{obs}|dX_{mis|\Theta},\Theta_*$

denotes the true value of the parameters and $f(X_{obs}|\Theta_*)$

The marginal density function of X_{obs} is denoted by the symbol When performing the RO step, it is necessary to discover the minimizer of the Kullback–Leibler divergence from $f(X_{obs},\vec{X}_{mis|\Theta})$ the joint density $f(X_{obs}|\Theta_*)h(X_{mis}|X_{obs},\Theta_n^{(t)})$. Used as Θ_* unknown and estimated using objective function called log-likelihood function.

$$\sum_{i=1}^{n}\log\left\{f(X_1^{obs},\vec{X}_i^{mis}|\Theta)\right\}/n \quad (13)$$

A regularisation term is included in the parameters to account for the high-dimensional scenario, in which the number of parameters can be much greater than the sample size, i.e, when the number of parameters is significantly greater than the sample size.we propose to estimate $\Theta_*^{(t)}$ the parameters using a larger sample size than is available.

$$\Theta_{n.p}^{(t+1)}=\arg\max_{\Theta}\left[\frac{1}{n}\sum_{i=1}\log\left\{f(X_{obs},X_{mis|\Theta})\right\}-P_\tau(\Theta)\right] \quad (14)$$

Where X_{mis} are drawn from $h(X_{mis}|X_{obs},\Theta_n^{(t)}),P_\tau$ denote the penalty function and τ denote regularization parameter.

4. Results and Evaluations.

4.1. The Evaluation Criterion

4.1.1. Mean Absolute Error (MAE) and RMSE (Root Mean Square Error)

Among the different data imputation procedures in which the characteristics are quantitative, the mean absolute error (MAE) and the root mean square error (RMSE) are typically used in evaluating the predictive capacity, and both are widely recognized.For the purpose of simplicity, let us assume that we already have samples of model errors calculated in the manner described as $(e_i; i=1,2,...,n)$ .No consideration is given to the uncertainties introduced by observational errors or the approach used to compare the model and data in this study. We also make the assumption that the error sample set is impartial. The root mean square error (RMSE) and the mean absolute error (MAE) for the data set are calculated as follows:

$$\mathbf{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|e_i\right| \quad \mathbf{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2} \tag{15}$$

The accuracy of classification is used to evaluate the performance of categorical attribute algorithms in terms of their overall performance.

4.1.2. Classification Accuracy (CA)

In order to evaluate the performance of categorical attribute algorithms, the classification accuracy (CA) must be met or exceeded

$$\mathbf{CA}=\frac{1}{n}\sum_{i=1}^{n} l(EC_i, TC_i) \tag{16}$$

In the case of a missing value, where EC_i and TC_i are the estimated and true class labels, respectively, with representing the total number of i^{th} missing values.

4.2. Experimental Evaluation

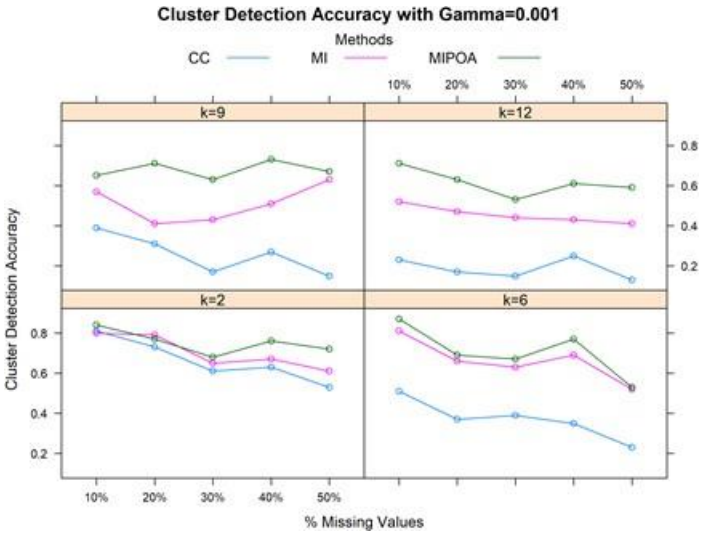
Data that is multivariate and has established mixing distributions can be used to test our hypotheses, allowing us to compare the performance of different methods. Data can be generated from a Gaussian mixture with a predetermined overlap between the distributions, as demonstrated in this study. It is possible to modify a dataset's clustering difficulty by manipulating this pair wise overlap, which is the sum of two incorrect classification probabilities. The overlap between the i^{th} and j^{th} component can be defined as

$$\gamma_{ij}=\gamma_{i|j}+\gamma_{j|i}$$

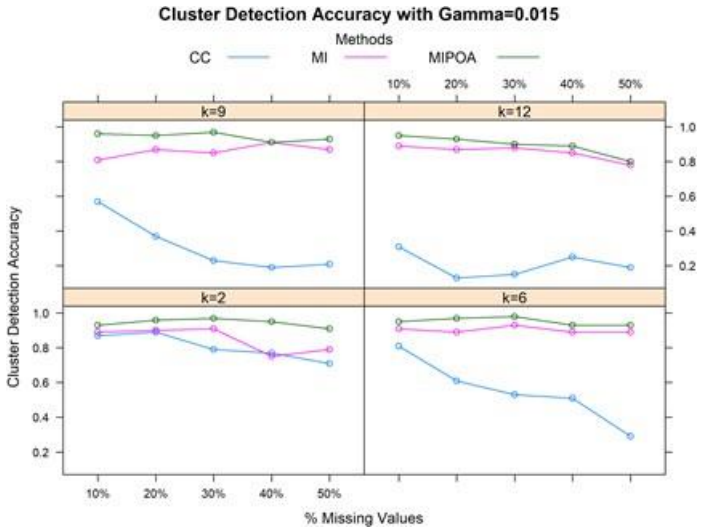
When an observation $\gamma_{i|j}$ is misclassified to a $\mathbf{X}$ to the i^{th} cluster, what is the likelihood that it is actually a j^{th} component.

$$\gamma_{ij}=Pr^*\pi_j\phi(X,\mu_j,\Sigma_j)<\pi_j\phi(X,\mu_i,\Sigma_i)\,|\,X:N_p(\mu_j,\Sigma_j)+$$

Random sampling was used to produce missing data from a uniform distribution, and the percentages of missing data were 10%, 20%, 30%, 40%, and 50%.



(a)



(b)

Figure 2 Cluster Detection Accuracy with (a) γ=0.001 (b) γ=0.015

In the presence of MCAR missing values, three methods (Complete Case (CC), Multiple Imputation (MI), and the suggested method Multiple Imputation–Penalized Optimization Algorithm (MIPOA)) were examined for their capacity to properly determine the number of components of the data.A preliminary simulation with 300 data sets with significant overlap **γ=0.001** as well separated cluster, **γ=0.015**,average separated cluster, **γ=0.05** clusters with substantial overlap was done and it was found that the overlap measure had an immense impact.There is a general decrease in detection ability as the number of missing people increases, as predicted. In the initial simulations, we found that the overlap had a large impact, and now we can see that dimensionality has a significant impact as well.Now, the method MIPOA is applied on Statlog data, then total 15 iterations applied to set of clusters and results showed that proposed MIPOA is better than existing algorithms in terms of RMSE. Especially, proposed method linearly decreases error rate with the extension on number of imputations. Fig 3 shows that this method outperforms KNN Imputation with

Mutual Information (KNNIMI) by 7 percent and performs marginally better than the Clustering-based Multiple Imputation (CMI) method.

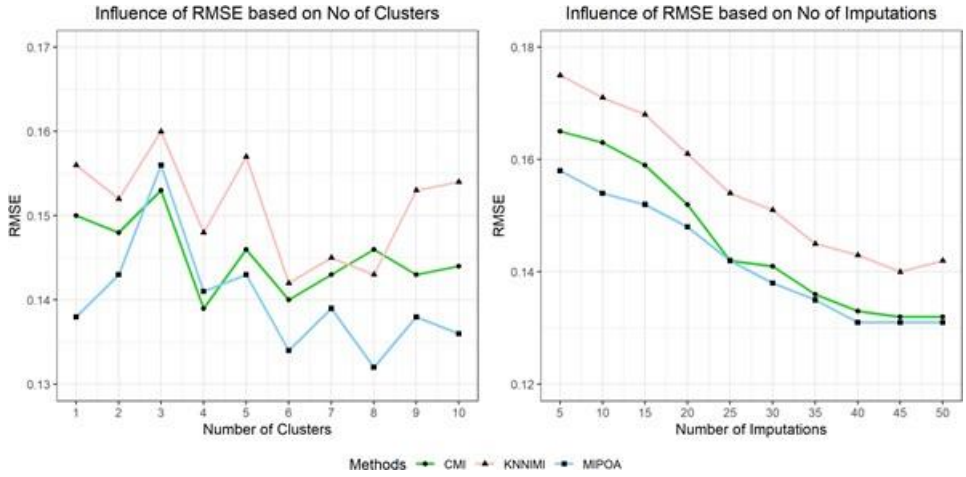


Figure3. Influence ofRMSE on(a) No.of Imputations(b) Number of Clusters on Statlog Data

In view of clustering principles, the relation among RMSE and clusters in cluster- based imputation techniques is shown in In Fig 3 from these plots, it is discussed that the complete data to be organized into total 10 clusters and tested performance, where RMSE produced to be minimum (i.e., 0.134) using proposed MI method MIPOA. In case of,KNNIMI the lower RMSE = 0.141 produced at cluster 7. Similarly, in CMI, RMSE to be 0.139 with 4 clusters.MIPOA produced least RMSE value than CMI and KNNIMI and results shownin Fig 3.

As the number of iterations before convergence emerges increases, it is clear that the cluster-based imputation methods used in this work yielded improved CA, as seen in Fig 4.Meanwhile, MIPOA achieved the best accuracy compared to other two techniques. The number of iterations increased, the accuracy raised and mention in the range of 0.83 to 0.88. In addition,

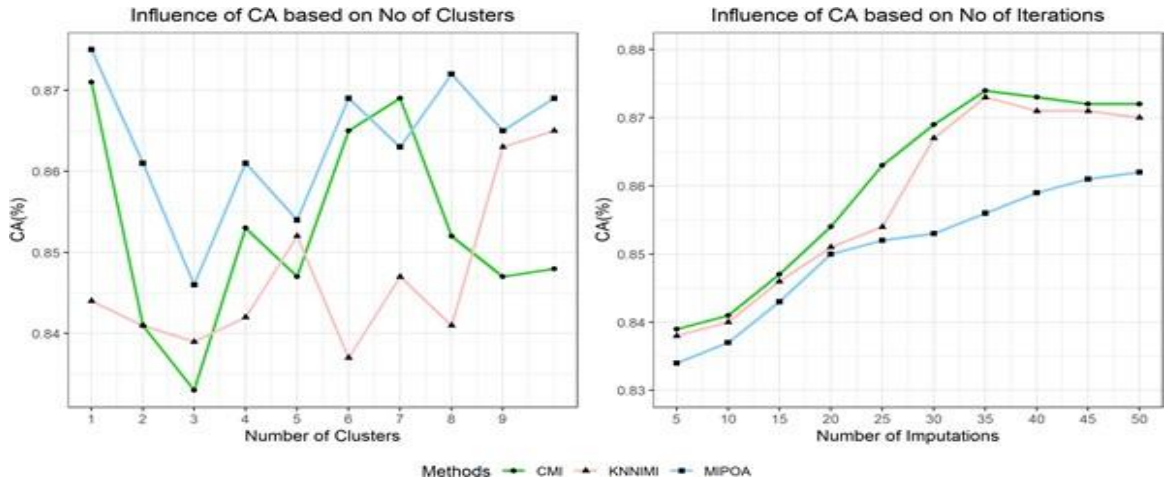
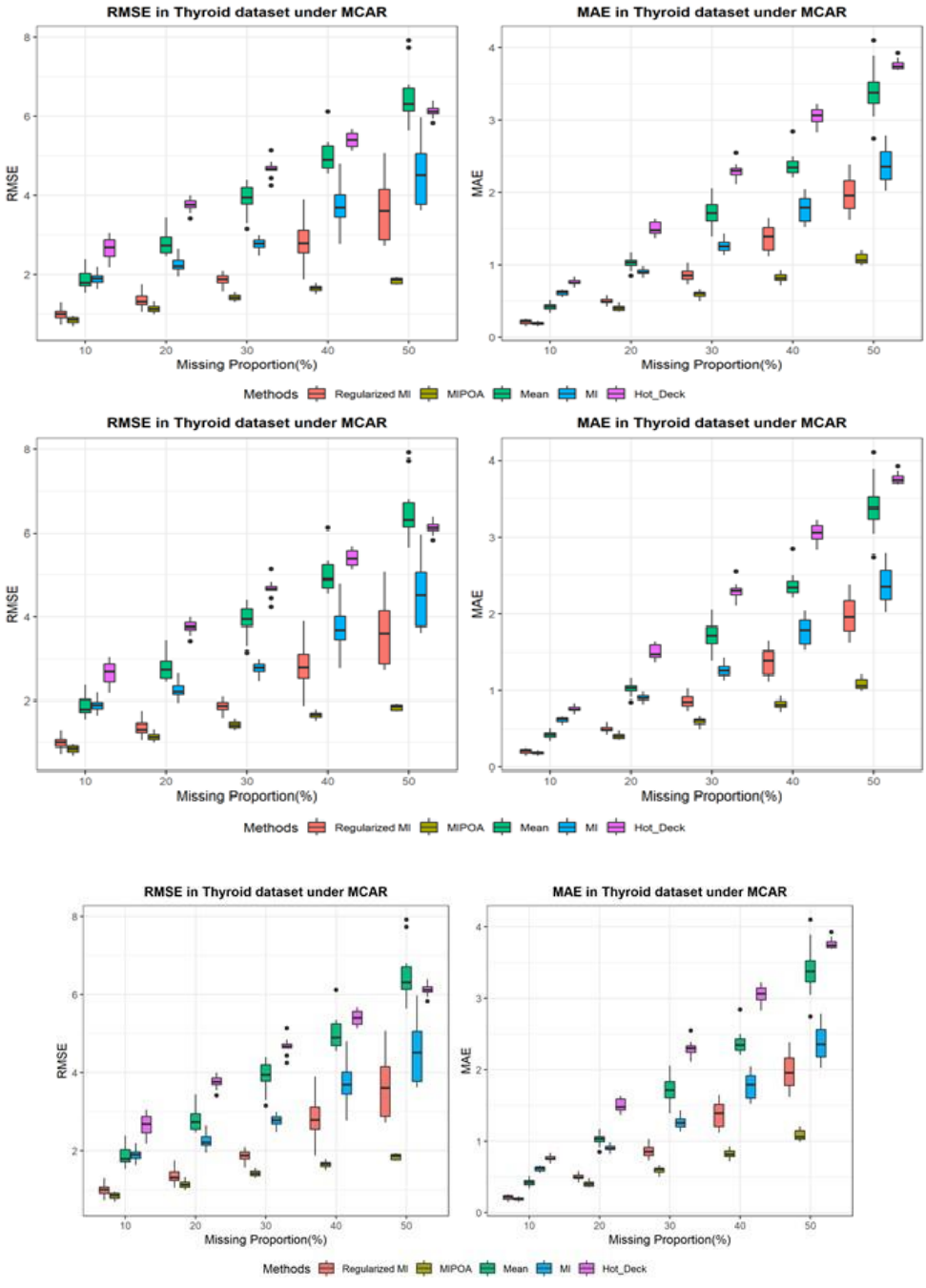


Figure 4. The CA influenced on (a) No.of Imputations (b) Number of Clusters on Liver Disorder Data.

Also, at cluster 8 proposed method produced better CA value 0.872 compared to and results shown in Fig4.Generally, KNNIMI and CMI proved to be inaccurate compared to MIPOA with the optimal number of clusters and iterations, respectively.Thyroid disease data set includes 7200 cases and 21 attributes. To test the performance of the missing analysis, the study focused on a missing component in artificial data with different missing reports, which are 10%, 20%, 30%, 40% and 50% associated with different missing mechanisms mentioned like MAR, and MCAR. The performance analysis of cluster-based imputation methods with a measure of the RMSE on Thyroid disease data set is shown in Fig 5. Fig 5 shows that MIPOA gives better results than other standard approaches including Regularized Multiple Imputation, Multiple Imputation, Hot-Deck Imputation, and Mean Imputation and in the absence of two mechanisms MCAR and MAR over the data set.



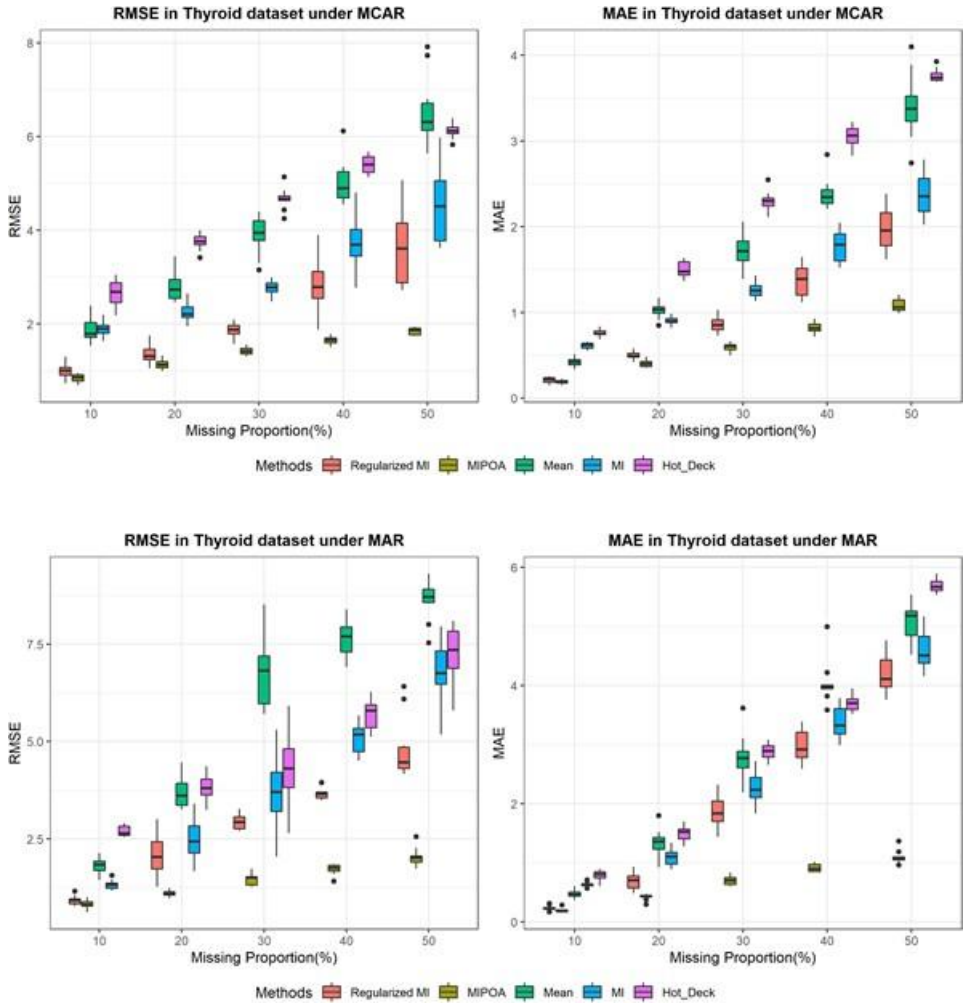


Figure 5. The RMSE and MAE in Thyroid Data for (a)MCAR (b) MAR

It is observed from the results, raising the RMSE values with impact on high missing proportion. Although, it is known from the study the RMSE yields maximum in case of MCAR compared to the MAR. The result , from the variation of missing proportion 10% to 50% , it is observed that error value derived low with range 1.00 to 2.20 in case of the MIPOA compared to single imputations(i.e, Mean and Hot Deck) with range 2.00 to 8.25 and Multiple Imputations(i.e, MI and Regularized MI) in range 1.25 to 7.5. The study also tested imputation techniques performance with MAE on Statlog data with MAR and MCAR mechanisms. From the figure (6), it is shown MAE yields maximum in case of MAR compared to the MCAR. The proportion of missingness 10% to 50% , results MAE value low with range 0.12 to 1.00 in case of the MIPOAcompared to single imputations(i.e, Mean and Hot Deck) with range 0.18 to 5.85 and Multiple Imputations(i.e, MI, and Regularized MI) in range 1.00 to 5.00. Then study focused on analyzing the performance of cluster-based imputation methods on thyroid diseasewhich is shown in Fig6. Also, it is mentioned that the proposed method has generated promising results with respect to the standard single and multiple imputations. In addition, it is the opinion of the phenomenon suggested that the performance classification accuracy (CA) deteriorates as of improvement over the missing proportion. From the results of the study, found that in the MCAR proposed method produces a 90% accuracy as compared to standard single and multiple imputation methods (i.e., 72% in

single imputations 83% in MI).Then in MAR, an accuracy of 80% resulted in the proposed method and whereas in the single imputation methods (i.e., Mean and Hot Deck) it is 66% and in MI(i.e., MI and Regularized MI) it IS 84%.

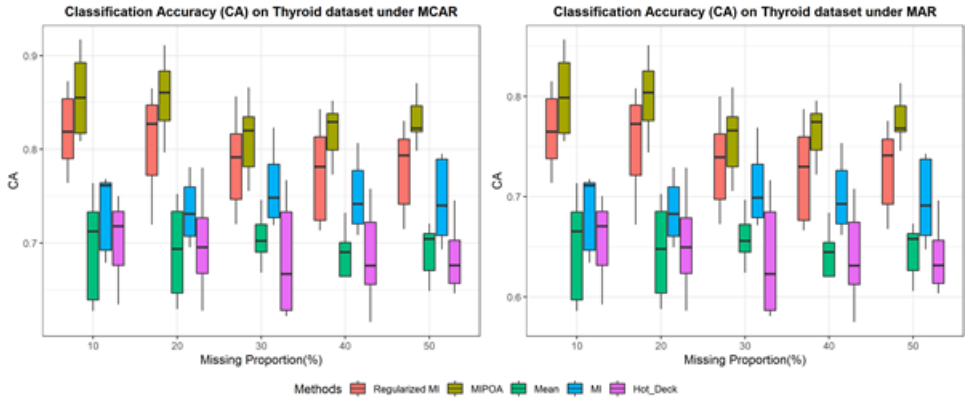


Figure 6. Classification Accuracy in Thyroid Data for (a)MCAR (b) MAR

5. CONCLUSION

Using missing data processing approaches, this work offers clustering-based imputation by splitting original data into two no overlapping subsets: missing-valued subsets and complete-valued subsets.The iterative imputation approach is paired with the identified groups after each missing value is incorporated into the nearest cluster using a gray relational analysis-based distance measure.Experiment results show that MIPOA surpasses the existing single and multiple imputation approaches in the field of thyroid disease. Continuous and discrete missing attribute RMSEs and MAEs were measured at various missing ratios. CAs were measured for discrete missing attributes.The effects of iteration times on the RMSE, MAE, and CA show that MIPOA converges faster and with better accuracy in the real application setting than other iterative imputation techniques. The MNAR Mechanism will be used to estimate and impute missing data in the future, and this will be the focus of future study.

References

[1]Hariri, R.H., Fredericks, E.M. & Bowers, K.M. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 6, 44 (2019).

[2] C. Liu Schafer JL. Analysis of incomplete multivariate data. London: Chapman and Hall; 1997.

[3]Gupta, M.K., Chandra, P. A comprehensive survey of data mining. Int. j. inf. tecnol. 12, 1243–1257 (2020).

[4] Liu Xingyi.Filling missing value algorithm based on Mahalanobis distance and gray analysis ,.Journal of Computer Applications, 2009, (9):2502-2506.

[5] K. Lavanya, L. S. S. Reddy and B. Eswara Reddy, ”Modelling of Missing Data Imputation using Additive LASSO Regression Model in Microsoft Azure”, Journal of Engineering and Applied Sciences,2018,Vol 13,Special Issue 8,pp:6324-6334.

[6]Schafer J.L. (2003) Multiple imputation in multivariate problems when the imputation and analysis models differ. Statistica Neerlandica, 57, 19–35.

[7]Liu Xingyi, Tan Yao, Zeng Chunhua.Filling missing data method based on the Mahalanobis distance ,Microcomputer Information, 2010,(9) :225-226.

[8]H. de Silva, A. Perera, "Missing data imputation using Evolutionary k- Nearest neighbor algorithm for gene expression data", 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), p. 141, 2016

[9] Monard, Maria-Carolina. (2002). A Study of K-Nearest Neighbour as an Imputation Method.

[10] Sengupta S., Das A.K. (2012) Dimension Reduction Using Clustering Algorithm and Rough Set Theory. In: Panigrahi B.K., Das S., Suganthan P.N., Nanda P.K. (eds) Swarm, Evolutionary, and Memetic Computing. SEMCCO 2012. Lecture Notes in Computer Science, vol 7677.

[11]F. Pacheco, M. Cerrada, R. V. Sánchez, D. Cabrera, C. Li and J. V. de Oliveira, "Clustering algorithm using rough set theory for unsupervised feature selection," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 3493-3499, doi: 10.1109/IJCNN.2016.7727647.

[12] Zhang, L.; Lu, W.; Liu, X.; Pedrycz, W.; Zhong, C. Fuzzy C-Means clustering of incomplete data based on probabilistic information granules of missing values. Knowl. Based Syst. 2016, 99, 51–70.

[13] Harrel, O, Zhou, XH. Multiple imputation: review of theory, implementation, and software. Stat Med 2007; 26: 3057–3077.

[14] Patil, Bankat & Joshi, R. & Toshniwal, Durga. (2010). Missing Value Imputation Based on K-Mean Clustering with Weighted Distance. Communications in Computer and Information Science. 94. 600-609. 10.1007/978-3-642-14834-7_56.

[15]Wiharto, Wiharto & Suryani, Esti. (2020). The Comparison of Clustering Algorithms K-Means and Fuzzy C-Means for Segmentation Retinal Blood Vessels. Acta Informatica Medica. 28. 42.

[16] C. C. Huang and H. M. Lee, "A grey-based nearest neighbor approach for missing attribute value prediction," Applied Intelligence, vol. 20, no. 3, pp. 239–252, 2004.

[17] Krzysztof Simiński. 2013. Clustering with Missing Values. Fundam. Inf. 123, 3 (July 2013), 331-350.

[18]Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang, "POP algorithm: kernel-based imputation to treat missing values in knowledge discovery from databases," Expert Systems with Applications, vol. 36, no. 2, pp. 2794–2804, 2009

[19] Little RJ, Rubin DB. Statistical analysis withmissing data. 2nd ed. New York:Wiley; 2002.

[20]Rubin DB. Multiple imputations for nonresponse in surveys. New York: Wiley; 1987.

[21]K. Lavanya, G.V.Suresh," An Additive Sparse Logistic Regularization Method for Cancer Classification in Microarray Data", The International Arab Journal of Information Technology, Vol. 18, No. 2, March 2021.https://doi.org/10.34028/iajit/18/10, ISSN: 1683-3198E-ISSN: 2309-4524.

[22]K.Lavanya, K. Harika, D. Monica, K. Sreshta,"Additive Tuning Lasso (AT-Lasso): A Proposed Smoothing Regularization technique for Shopping Sale Price Prediction", International Journal of Advanced Science Technology, Vol 29,No 05,pp 878-886,2020.

[23]Lavanya K., Reddy, L., & Reddy, B. E. (2019). Distributed Based Serial Regression Multiple Imputation for High Dimensional Multivariate Data in Multicore Environment of Cloud. International Journal of Ambient Computing and Intelligence (IJACI), 10(2), 63-79. doi:10.4018/IJACI.2019040105.