Missing Data Imputation using Multiple Imputation with Adaptive LASSO for Parkinson's Disease Data

Repudi Pitchiah[1], Dr.T.Sasi Rooba[2], Dr.K. Uma Pavan Kumar[3]

[1]Resesrch Scholar, Dept. of Computer Science & Engineering, Annamalai University, Chidambaram.
[2]Dept. of Computer Science & Engineering, Annamalai University, Chidambaram.
[3]Dept of Computer Science and Engineering, MRIT, Hyderabad, Telangana.
Email:[1]pitchaiah99@gmail.com,[2]sasiruba@gmail.com,[3]dr.kethavarapu@gmail.com

Article History: Received: 5 April 2021; Accepted: 14 May 2021; Published online: 22 June 2021

Abstract: Nerve cells are the brain's building blocks for the nervous system. Once destroyed, they do not regenerate. When these nerve cells are damaged, the dopamine they contain is depleted, causing motor abilities and speech to deteriorate. Before the brain cells are impaired, the voice goes through a series of modifications. Voice shifts assist in the early detection of Parkinson's disease, avoiding injury to brain cells that would result in decreased balance and movement. However, this condition often suffers from missed data in clinical outcomes due to a variety of factors such as dropout, illness, and so on. Hence, imputation of these kinds of missing data is always performed prior to performing an intent-to-treat study. Indeed, predictive analysis of data relating to disease will not be feasible without the use of a suitable framework that efficiently manages missing data. The paper proposes an Adaptive LASSO Imputation approach based on item answer theory, which allows multiple imputations to be done when working with multiple sources of correlation. The accuracy of each imputation procedure was assessed using the Root Mean Square Error (RMSE) and Mean Absolute Error. The proposed method is applied on three types of Missing Data i.e MAR, MCAR, NMAR. The outcomes demonstrated that the suggested approach outperforms all other algorithms.

Keywords: *Parkinson's Disease, High-Dimensional Data, Multiple Imputations, Regression, Missing Data*

1. Introduction

Parkinson's disease (PD) is a neurodegenerative condition leading to increasingly limited physical movements as well as a worsening speech function. In 1817, Dr. James Parkinson [1] found and identified this disorder, which he dubbed the "Shaking Palsy." Neuro-degenerative disorders are inherited that are characterized as intermittent illnesses that are marked by a declining nervous system functions, almost resulting in debilitation and vegetative condition (JPND research, 2015). Parkinson's disease is the second most prevalent neurodegenerative disease, behind Alzheimer's disease, brain cancer, degenerative nerve disorders, and epilepsy [2]. The gradual reduction of dopamine neurons in the Substantianigra region of the midbrain – the brain's "movement regulation unit" – is the primary cause of Parkinson's disease. When dopamine is depleted, neurons fire erratically, resulting in hypokinetic movement dysfunction [3]. And though this condition can be easily detected in its early stages, successful treatment remains a challenge. There is still no cure or surgical therapy for Parkinson's disease. The triggers behind onset of Parkinson's disease remain unclear even after decades of research.Most researchers believe that PD is caused by a mixture of genetic [4] and environmental [5] triggers, including susceptibility to environmental toxins, head injury, rural

life, drinking water, manganese, and pesticide exposure. These indicators may differ from one individual to another. In addition, a person can experience certain symptoms that are usually not found in other patients.

Table.1. Different Stages of Parkinson Disease

S.No	Stage	Symptoms
1	Mildest Stage (Stage 1)	In this stage, the PD patients have least interference with routine tasks. Tremors and other symptoms are restricted toone side of the body
2	Moderate Stage (Stage 2)	In this stage, symptoms like stiffness, resting tremors and trembling can be sensed on both sides of the body. Also, facial expressions of PD patients may get changed
3	Mid-Stage (Stage 3)	During this stage, major changes like balance loss, decreased flexes in addition with stage II symptoms will be observed in PD patients. Occupational therapy combined with medication may help in decreasing the symptoms
4	Progressive Stage (Stage 4)	The condition of PD patient will get worse in this stage and it becomes difficult for the patient to move without some assistive device like a walker.
5	Advanced stage (Stage 5)	Stage 5 is the most advanced and debilitating stage of PD. Stiffness in legs may cause freezing when standing. Patientsare frequently unable to stand without falling. They may experience hallucinations and occasional delusions

Table 1 shows a summary of the various stages of Parkinson's disease. Tremor of the hands, head, thighs, jaw, and face, bradykinesia (slowness of movement), rigidity or weakness of the limbs and trunk, and postural dysfunction (decreased equilibrium and coordination) are all primary motor signs of Parkinson's disease [6–8Additionally, non-motor symptoms such as depression and memory loss may adversely impact one's quality of life [9,10]. In later stage detection, PD offers little chance to provide any effective therapy. Furthermore, if therapy is initiated at an early stage, it could be less successful in slowing the development of Parkinson's disease. This condition necessitates an early and definitive diagnosis of Parkinson's disease, allowing patients to sustain a high standard of life. There is currently no single blood or laboratory test that can be used to diagnose Parkinson's disease and monitor its development. After all, rating mechanisms like Hoehn and Yahr scale (1967), the Unified Parkinson Disease Rating Scale (UPDRS) and its updated variant MDS-UPDRS [11] can also be applied for detecting Parkinson's disease.Some drawbacks are as follows:

- Lack of technicians with the requisite expertise
- Need for excessive amounts of time and efforts from patients for long period [12] etc.

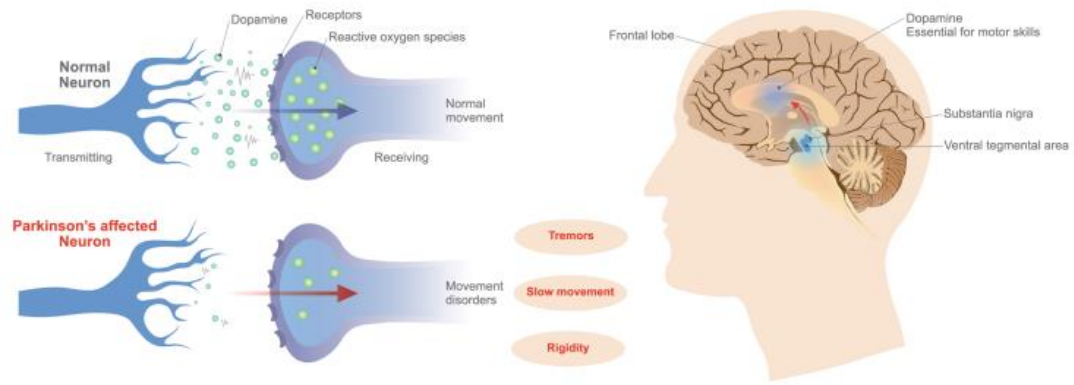

Fig.1. Parkinson's disease (normal movement vs. movement disorders)

Since certain neurological diseases have the same pathology, it can be difficult to differentiate between them. Idiopathic PD is reported in about 75% clinically diagnose patients suffering from Parkinson's disease. To increase diagnostic accuracy and assist doctors in making the best decisions, automated methods based on machine learning are needed.Majorly, illnesses or ailments cane be classified as communicable and non-communicable categories in the broadest sense. While certain illnesses can be healed, there are also those who do not have a cure. This illness affects almost 11 million people worldwide. Many studies have been carried out in order to find a better treatment for the disease.Machine learning is the process of analyzing historical data in order to come up with a suitable solution for the future. PD presents a difficulty in terms of data multidimensionality, as well as the issue of incomplete data in critical parameters. There are some approaches to dealing with such problems. The main aim of this paper is to work on Parkinson's disease by analyzing missing data, maintaining the dataset's essential nature and functionality, and sustaining the relationships between variables. Finally, it calculates the missing values using an approach that gives lesser errors.

2. Origin of Missing Data

Data representing consistently observed values in a full data set always fail due to issues like non-response, mistake, system malfunction, and other factors. Most modern promising fields suffer from insufficient data, which prevents good conclusions from being drawn. As a result, a large range of experiments and approaches must be proposed in order to recover the missing data and fill in the gaps in the results. The method's effectiveness is significantly dependent on a thorough analysis of the missing data, which involves the form of missing data, the cause, and the trend.

Let $Y=(y_{ij}):(n\times k)$ rectangular data set with no missing values then.

$$M=(m_{ij})\begin{cases} m_{ij}=1 \text{ if } y_{ij} \text{ is missing} \\ m_{ij}=0 \text{ if } y_{ij} \text{ is present} \end{cases} \quad (1)$$

2.1. Missing Models

A model is a dataset with the cumulative number of observations set as either missed or observed. It's also used to figure out the presence of incomplete data values. To put it another way, it indicates the exact location of "gaps" in the data sets. The univariate and multivariate categories make up the deficiency model. Consider the data that is form $n \times p$ and specify the complete matrix to n .Single variable data consists of the total j values among some of the values are missing and remaining are partially observed a data to represent as $Y_j = (Y_o, Y_m)$. The equivalent response matrix represented by '$\Re$'which took the entry' 0 'if values are missing, otherwise it must be' 1 '.

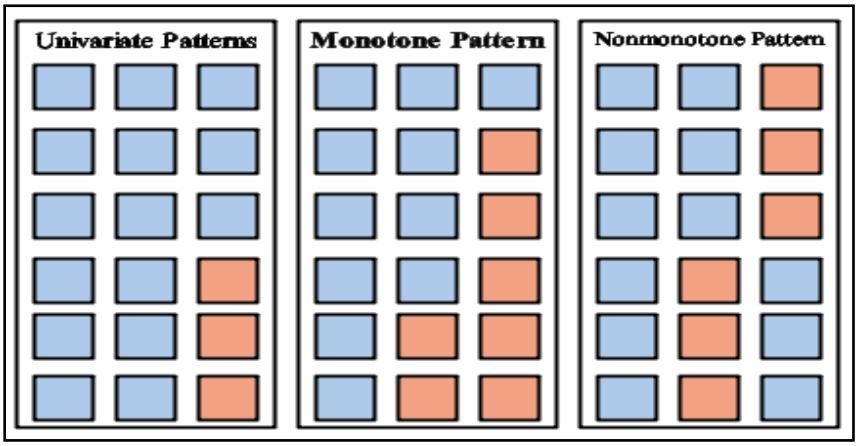

Fig..2. Representations of missing data models

There are two types of models involved in the withdrawal effect that is mentioned:

- **Univariate:** Missing data that is affected by a single variable called univariate data.
- **Monotonous and not monotonous:** model where the missing data where all the variables of Y_j be in the order in which the set of variables in Y_j not be visible in Y_j and $i > j$ said to be monotonous. The model that never follows monotonous says non-monotonous. There are different derived models in multivariate data and is shown in Fig.2.

2.2. Mechanisms of Missing Data

The framework for classifying missing data is labelled as ignorable or not ignorable. The risk of missing data in Ignorable situations is determined by the identified data rather than the missing data. The chance of missing data in Non-ignorable is determined by the missing data rather than the observable data. Figure 3 depicts the three categories of missing data frameworks that are in line with ignorable and non-ignorable missing data mechanisms. Here, it explains the process through which missing data is linked between missing variables and variable values in the data table.

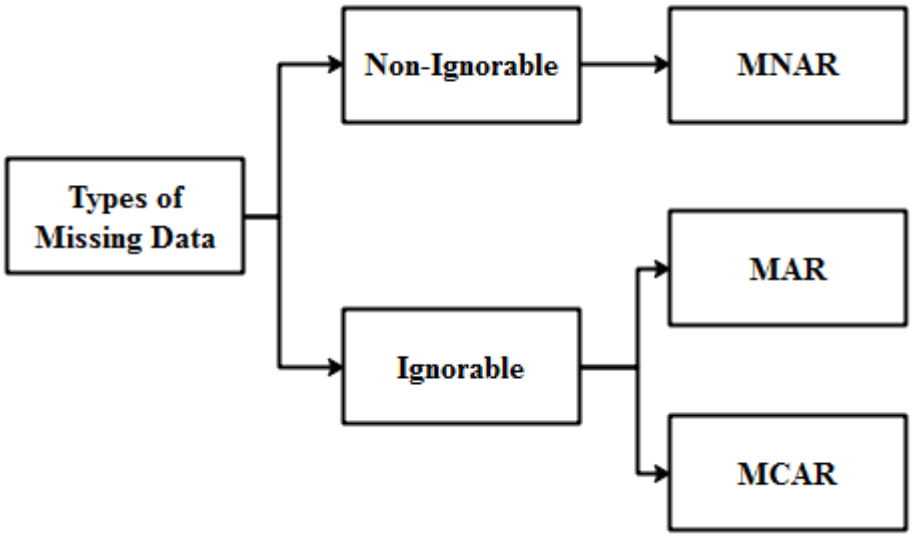

Fig.3. Classification of Missing Data

Let $Y=(y_{ij})$ represent the complete data of n individuals and k variables, with $y_i=(y_{i1},...,y_{ik})$ where y_{ij} is the value of j^{th} variable for individual i. Let $Y_{obs}=(y_{1obs}^T,...,y_{nobs}^T)^T$ and $Y_{mis}=(y_{1mis}^T,...,y_{nmis}^T)^T$ represent observed parts and missing parts of data, respectively. Then the i^{th} row (y_{iobs},y_{imis}). We define the missing indicator matrix $R=(r_{ij})$, with $r_{ij}=1$ when y_{ij} is missing and $r_{ij}=0$ otherwise. We define $Y_{-j}=(Y_1,....,Y_{j-1},Y_{j+1},...,Y_k)$ and $R_{-j}=(R_1,....,R_{j-1},R_{j+1},...,R_k)$, where Y_j and R_j represent the value and missing indicator for the j^{th} variable, respectively. Then we can use the distribution of R conditioning on Y to describe the missing data mechanism, as $f(R|Y,\phi)$ where ϕ represents the parameters. Let us consider three common missing data mechanisms:

2.2.1.Missing completely at random (MCAR): The missing values are independentof both the observed and missing data, such as flipping a coin before answeringa question. That is,

$$f(R|Y,\phi)=f(R|\phi) \text{ for all } Y,\phi \quad (2)$$

2.2.2.Missing at random (MAR): Given observed data, data are missing independently of unobserved data, e.g., male participants refuse to respond to depression questions, but it does not depend on their level of depression. That is,

$$f(R|Y,\phi)=f(R|Y_{obs},\phi) \text{ for all } Y_{mis},\phi \quad (3)$$

2.2.3.Missing not at random (MNAR): The distribution of missingness depends on the missing values. For example, people who do not like to say about their disease are less likely to respond to the question. That is,

$$f(R|Y,\phi)=f(R|Y_{obs},Y_{mis},\phi) \text{ for all } Y,\phi \quad (4)$$

2.2.4.Item wise conditionally independent nonresponse (ICIN): Given other items and missingness indicators, the missingness of each item is independent of the value itself. That is,

$$Y_j \perp R_j \mid Y_{-j}, R_{-j} \ (j=1,...,k) \quad (5)$$

There are currently different methods of dealing with missing data. One such method is Multiple Imputation.

3.Multiple Imputation

Imputation, post-imputation examination, and pooling of data are the three important phases in MI. Each missing value is replaced by m plausible values during the imputation stage. In essence, various ways are explored to derive the possible values, like a Bayesian posterior statistical distribution is used to derive a random draw. The imputation stage yields m datasets that have been completed. The finished dataset differs from a "complete dataset" wherein the former is seen representing data having imputed values. The latter is shown representing the dataset that is possibly seen in the absence of any missing value. Consider the following scenario in which the population parameter of significance is Q. Every m dataset is fitted witha complete data method in the analysis stage, leading to m estimated parameters $\hat{Q}_i$ and itsrelative variance $U_i (i = 1, 2, ..., m)$. The m analysis results are applied in the results-pooling stage togenerate the ultimate inference while applying Rubin's combining rule.The final estimate refers to the average of the estimates:

$$\bar{Q}_m = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i \quad (6)$$

The final variance estimation is composed of the within-imputation variance $\bar{U}_m$ as well as the imputation variance B_m.The average of the individual variance estimates generated from each imputed dataset is referred to as the within-imputation U_m.

$$\bar{U}_m = \frac{1}{m}\sum_{i=1}^{m} U_i \quad (7)$$

The between imputation variance B_m is defined as follows:

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m} (\hat{Q}_i - \bar{Q}_m)^2 \quad (8)$$

The adjustment between MI imputation and within-MI imputation is summed up as the final estimated variance of the MI estimate $\bar{Q}_m$. As per Rubin's rule, it is represented as $T_m = (1+\frac{1}{m})Bm + \bar{U}_m$. We can define $r_m = (1+\frac{1}{m})Bm / \bar{U}_m$ as the relative hike in variance on account ofnonresponse, while $\gamma_m = \frac{r_m}{r_m + 1}$ refers to the fraction of missing information caused by nonresponse.The following advantages are seen in Multiple imputation:

- Multiple imputation will help maintain the inherent relationship between variables by simulating the distribution of missing data.
- Multiple imputation contains a considerable volume of detail to calculate the variance of the estimation outcomes compared to single-value imputation's simplistic estimation results.
- Multiple imputation is used to produce the filling values. The difference between them may mean that the missing data is random.

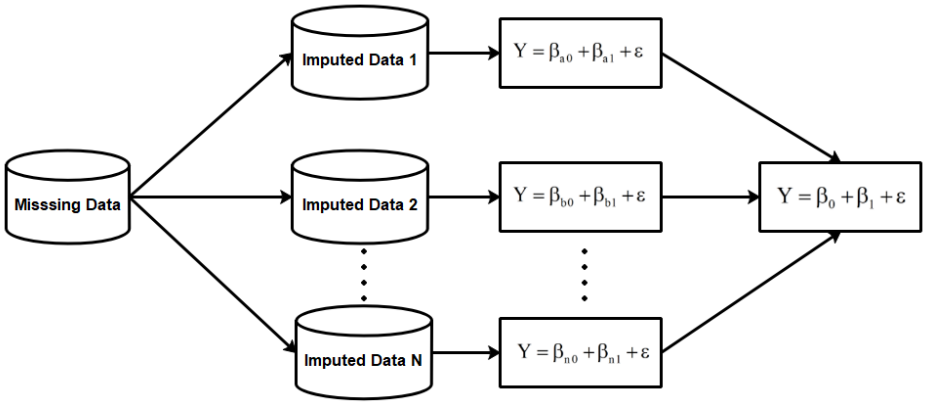

Fig.4. Classification of Missing Data

4. Proposed Work

LASSO, a regularization method is used as a variable selection tool as well. It performs variable selection by allowing certain coefficient estimates to equal zero and shrinking coefficient estimates towards zero (unlike ridge regression). When basic linear regression notation and conditions with p predictors are applied, it makes LASSO coefficients to minimize the quantity as given below:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (8)$$

which equals the residual sums of squares plus a penalty term:

$$RSS + \lambda\sum_{j=1}^{p}|\beta_j| \quad (9)$$

The λ in the penalty term is a tuning parameter that is used to controlthe extent of shrink estimated by the coefficient. Variable selection is performed when some coefficient estimations are equalized to zero forced by large values of λ.

We present the adaptive multiple imputation lasso (MIAL), a novel approach that incorporates multiple imputation and adaptive lasso. In terms of variable selection and estimation, the adaptive lasso shows better performance than the closely clustered predictor variables. The developed approach can manage an arbitrary missing model without monotone under the assumption of absence at random (MAR) and consider the $p>>n$ Case. MIAL now has a few additional capabilities. To begin, MIAL uses several imputations to expand the random lasso to data with missing entries. Second, stability selection increases the adaptive lasso's tight

threshold, resulting in higher prediction precision, improved variable selection efficiency, and variable value rating.MIAL is divided into four stages. Multiple imputation is used in the first step to produce multiple sets of imputed results. For each imputed data set, bootstrap samples are collected in the second stage, and a measure of value for each variable is generated. In the third stage, Lasso-ALS estimates are generated for bootstrap datasets with variables sampled from measurements of significance. In the fourth step, final estimators are extracted by aggregation and stability selection to create a final sparse model.

4.1. Proposed Method: An Adaptive Lasso for Multiple Imputation

Consider data set with number of samples to be represented as ' n ' andpredictors/features to be denoted as 'p' with missing data.The imputation is the popular method to fill missing values with the linear model and it is represented as $Y=\beta_0+X\beta+\varepsilon$. From the linear model Y is known as continuous response variable and set of predictor variables represented as X with the $n\times p$ dimension, and error to be denoted as ε. The β_0 represents intercept and $\beta=(\beta_1,\beta_2...\beta_p)$ is parameter estimates of X correspondingly. However, replacing single imputation values in multiple times considered as better method and is known as Multiple Imputation. The method which can applied to large studies depends on Regularization Criteria. Number of Regularized methods were applied to MI which includes Lasso, Ridge, Elastic Net and Adaptive Lasso. In this study applied Adaptive Lasso to MI to impute missing value in the data set. The complete idea described in the Algorithm 1.

Algorithm 1:MIAL(Multiple Imputation with Adaptive Lasso) algorithm

Step 1: Consider the missing data set X with the dimension of $n\times p$ with n samples and p predictors. Number of imputations to be considered as'm'.

Step 2: Next applied MI for the missing data X. The resultant imputed data set is I_m Im and normalize the values of each feature in the range of 0 to 1.

Step 3: Applied Bootstrap approach for the imputed data set I_m. Total B bootstrap samples generated for every imputed data set.

Step 3.1: calculate predictor importance measures $\hat{\beta}_{ij}^{(b)}$ for the b_j predictors using Adaptive Lasso (ALS) as follows:Thebootstrap sample $b=\{1,2,.....,B\}$ and resultant predictors b_j , j=1,2,...,p with $i\in\{1,2,....m\}$ imputation.

Step 3.2: Next estimated importance of every predictor variable as follows:

$$I_j=(mB)^{-1}\left|\sum_{i=1}^{m}\sum_{b=1}^{B}\hat{\beta}_{ij}^{(b)}\right|$$

Step 4: Estimation of the initial MIAL to be represented as:

Step 4.1: Among the b_j bootstrap sample, selected only $\lceil j/2\rceil$ number of the variables from the X_j and also calculated I_j to know the importance of the every

sample.

Step 4.2: Next, estimated K **number of** exponential deteriorations to be represented as Ω . Of course, this is controlled by theparameters λ .

Step 4.3: Now, used AdaptiveLasso (ALS) and calculated set of predictors with $\hat{\beta}_{ij\lambda}^{(b)}$ for β_j, and $\lambda \in \Omega$.

Step 5: For the MIAL estimate with the $m \times B$ to be represented as:

$$\hat{\beta}_j^{init} = (mB)^{-1} \sum_{i=1}^{m} \sum_{b=1}^{B} \hat{\beta}_{ij\lambda_{ib}}^{(b)}$$

Where λ_{ib} is the optimized parameter and is calculated using the 10-fold cross validation.

Step 6: Next estimated the selection probability of the MIAL estimates to be represented as:

$$\hat{\Pi}_j^{\lambda} = (mB)^{-1} \sum_{i=1}^{m} \sum_{b=1}^{B} I\left\{\hat{\beta}_{ij\lambda}^{(b)} \neq 0\right\}$$

Step 7: from the result of the previous step, retrieved maximum value of the selection probability to be estimated as :

$\max_{\lambda \in \Omega} \hat{\Pi}_j^{\lambda}$.

Step 8: Next, estimated the important variables from the selection probability and not exceed to the threshold π_{thr}

$$\hat{S}^{stable} = \left\{ j : \max_{\lambda \in \Omega} \hat{\Pi}_j^{\lambda} \geq \pi_{thr} \right\}$$

Where threshold π_{thr} to be estimated using 10-foldcross-validation.

Step 9: At, the end return the MIAL estimates and is represented as:

$$\hat{\beta}_j = \hat{\beta}_j^{init} \times I\left\{ j \in \hat{S}^{stable} \right\}$$

From the algorithm 1, step 2 to 3 is represented as a two-step procedure. In the procedure, corresponding regularized estimator so called as lasso to be estimated as $\hat{\beta}=\arg\min \left\|y-X\beta\right\|_2^2 + \lambda \left\|\beta\right\|_1$. From the lasso estimation λ to be calculated using 10-fold cross validation.

4.2. Performance Measures

Five separate measures were used to compare the different MI approaches by calculating a unidirectional marginal factor, as well as all the joint bidirectional and tripartite probabilities.

4.2.1. Mean Squared Error

An estimator's mean squared error (MSE) is the average of the squares of the errors. The MSE is a probability function that represents the estimated value of a squared loss or error. The mean square error (MSE) was a measure of the difference between $\hat{q}_M$ and Q. This study used a scaled

version of MSE, relative MSE, for ease of analysis as MSE values for bivariate and trivial probability estimates are typically very poor, which is defined as

$$\text{MSE}(\hat{q}_M)=\frac{\text{MSE}(\hat{q}_M) \text{ based on imputed data for MI method}}{\text{MSE(q) based on pre-missing data}} \quad (10)$$

This is a comparison of the elevated MSE for a study with no missing data to the MSE for the same sample using one of the three MI processes. Intuitively, if a method's imputations are fine, the method's MSE should be as low as the method's MSE based on the pre-missing results. Therefore, quality imputations should yield values around 1.

4.2.2. Root Mean Square Error (RMSE)

RMSE is square root of MSE. The root mean square error (RMSE) is a useful metric for assessing the precision or overall accuracy of each imputation technique. The disparity between the values expected by a formula or estimator and the observed values is measured using the root mean square error (RMSE). In this study, the standard deviation of the variations between the expected and observed values is represented by the RMSE. When the measurements are done on the data set that was used for the estimation, these individual deviations are called residuals, and when they are measured out of sample, they are called prediction errors. The RMSE of predicted values $\hat{y}_t$ for times t of the dependent variable of a regression y_t is calculated for n different predictions as the square root of the mean of the squares of the deviations:

$$\text{RMSE}=\sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t-y_t)^2}{n}} \quad (11)$$

4.2.3. Mean Absolute Error (MAE)

The mean absolute error (MAE) is measured as the average of the absolute variations between actual and predicted missing values. The best estimate tool is the one with the lowest MAE weight. Hence, the method is defined as

$$\frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t-y_t| \quad (12)$$

where n is the number of imputations, $\hat{y}_t$ are the imputed values; and y_t are the observed data values. MAE values will range from 0 to infinity, with zero denoting a perfect match.

4.3. Parkinson's Disease Dataset

This research applied a Parkinson's disease dataset from the UCI Machine Learning repository (https://archive.ics.uci.edu/ml/datasets/parkinsons). The PD dataset includes 195 voice recordings from 31 individuals, each with 22 characteristics (referred to as "predictors" or "explanatory variables" in this study). Out31 persons, 23 of them enjoyed good health. The average of speech fundamental frequency, maximum, low, change in fundamental frequency, change in amplitude, ratio of noise to tonal components over voice, dynamic nonlinear sophistication, concentration fractal size of the signal exponents, and nonlinear measures of

fundamental frequency fluctuations are among the 22 features used in biomedical voice measurements.Speech and voice signals are used to measure all the features defining the characteristics of the speech displayed in the recordings. In Table.1, the detailed characteristics of each variable are listed. Moreover, each observation has a response variable that indicates whether the person has Parkinson's disease. In the PD dataset, the answer variable is known as "status." In addition, each observation has a response variable that indicates whether the person has Parkinson's disease.

Table.1. Characteristic characteristics of the PD dataset

Function Number	Function Name	Description
X1	MDVP: Fo (Hz)	Middle vocal fundamentals
X2	MDVP: Fhi (Hz)	Maximum fundamental speech frequency
X3	MDVP: Flo (Hz)	Minimum fundamental speech frequency
X4	MDVP: Jitter (%)	Gigue Kay Pentax MDVP in percentage
X5	MDVP: Jitter (Abs)	Kay Pentax MDVP absolute jitter in microseconds
X6	MDVP: RAP	Kay Pentax MDVP Relative Amplitude Disturbance
X7	MDVP: PPQ	Kay Pentax MDVP Five Point Period Disturbance Quotient
X8	Jitter: DDP	Average absolute difference in differences between cycles divided by the average period
X9	MDVP: shimmering	Pentax MDVP local shimmer key
X10	MDVP: Shimmer (dB)	Pentax MDVP key local flicker in decibels
X11	Shimmer: APQ3	3-point amplitude disturbance quotient
X12	Shimmer: APQ5	5-point amplitude disturbance quotient
X13	MDVP: APQ	Kay Pentax MDVP Eleven Point Amplitude Disturbance Quotient
X14	Shimmer: DDA	Mean absolute difference between consecutive differences between amplitude of consecutive periods
X15	NHR	Noise / harmonic ratio
X16	HNR	Harmonic / noise ratio

X17	RPDE	Return period density entropy
X18	DFA	Analysis of dissolved fluctuations
X19	Propagation1	Non-linear measurement of the fundamental frequency
X20	Propagation2	Non-linear measurement of the fundamental frequency
X21	D2	Entropy of the height period
X22	EAR	Not

5. Results

5.1. Explore Missing Data

For better exploration of the missing data the data set is reduced from 22 to 6. Fig.4. shows the lack of data on Parkinson's disease.This paper uses python 3.5.8.for generating the results. The Pitch (PPE) has a high percentage missing of 29.74% while the Jitter.DDP has 10.26% of absence, etc. This paper uses python 3.5.8.for generating the results.

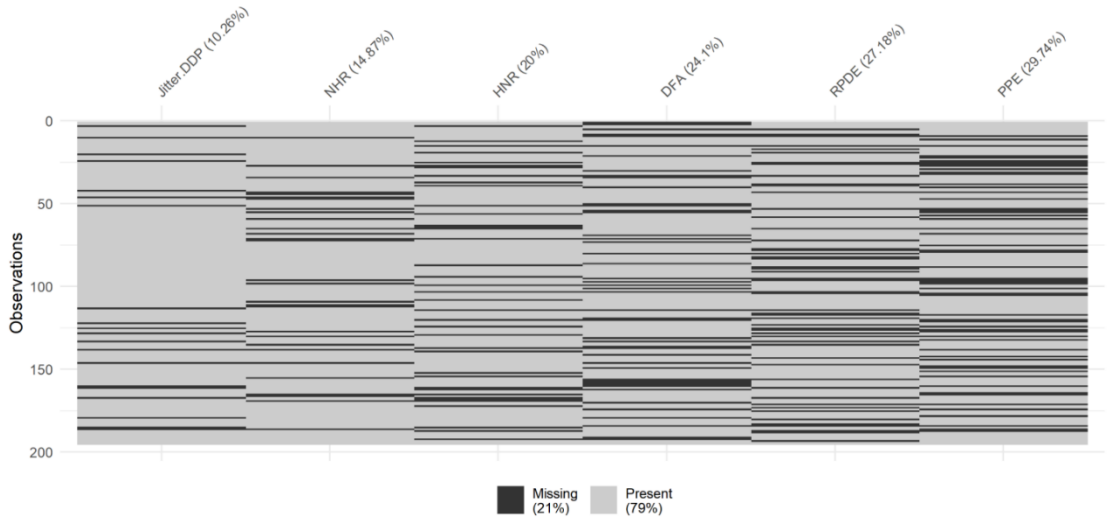

Fig.5. Percentage of absence in Parkinson's disease data

5.2. Aggregation Graph

The graph is used to locate the distribution of missing values in MCAR which provides an overview of missing data with patterns of missing values. It also examines in detail aspects related to the location of missing samples as well as their frequency. In Figure 4.5, the results of exploring missing data on the absolute frequency and proportion of each variable in the data set are included. However, the description of the missing observations is provided in the graph to the left. On the right, the projection is done by combining red and yellow boxes, representing the lack of variables. But the frequency of absences in each combination of variables is derivedplot as projected in a separate box. It is seen that the plot improves the lack model when aligned with a linear order of frequency. Likewise, the lower left row contains cases with missing values roughly the same as those in the upper left corner.

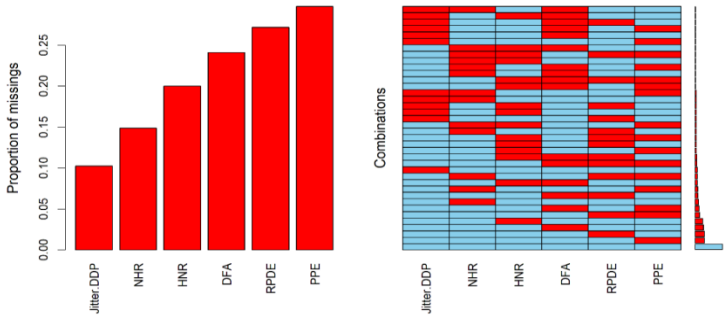

Fig.5 Aggregation diagram for exploring missing data in Parkinson's disease data.

The given data set is embedded with 10%, 20%, 30%, 40% and 50% of missingness. Then proposed multiple imputation with adaptive lasso algorithm is used for imputing missing data. The correlation coefficient calculated between actual and imputedvalues in PDD dataset is given inTable.3.

Table.3. Correlation Coefficient between actual and imputed values of PDD Dataset

Missing Rate (%)	Mean Imputation	Multiple Imputation	Lasso Imputation	MIAL Imputation
5	0.54	0.86	0.92	0.97
10	0.51	0.83	0.89	0.96
20	0.5	0.8	0.84	0.92
30	0.47	0.75	0.81	0.89
40	0.44	0.71	0.77	0.86
50	0.42	0.67	0.75	0.81

Compared to Mean Imputation, Multiple Imputation and Lasso Imputation, MIAL Imputation achievesvalue of r close to 1. It indicates that the correlation between actual valuesand the values imputed by using MIALare highly correlated.It is observed that from the Table.4,5,6 when there is less than 10 % missingness, RMSE of MIAL is less than 1%. When % of missingness increases, RMSE value also increases, but it is comparatively better than other methods.

Table.4. RMSE of MIAL method with other methods under MAR

Missing Rate (%)	Mean Imputation	Multiple Imputation	Lasso Imputation	MIAL Imputation
5	5.24	1.86	0.68	0.59
10	7.47	2.97	0.97	0.82
20	13.01	5.24	1.35	1.2
30	16.82	6.83	2.79	2.48
40	19.57	8.44	4.26	3.76
50	22.45	9.78	6.48	5.75

Table.5. RMSE of MIAL method with other methods under MCAR

Missing	Mean	Multiple	Lasso	MIAL

Rate (%)	Imputation	Imputation	Imputation	Imputation
5	5.57	2.04	0.78	0.63
10	7.21	2.98	0.87	0.79
20	13.85	5.58	2.26	1.89
30	15.89	6.52	2.75	2.48
40	20.23	8.86	4.24	3.8
50	23.36	10.16	5.69	4.61

Table.6. RMSE of MIAL method with other methods under NMAR

Missing Rate (%)	Mean Imputation	Multiple Imputation	Lasso Imputation	MIAL Imputation
5	5.74	3.56	2.08	0.79
10	7.38	5.47	3.04	0.91
20	14.07	9.86	5.73	2.36
30	16.69	12.24	6.68	2.66
40	20.78	15.48	8.97	4.47
50	23.15	19.79	10.35	6.03

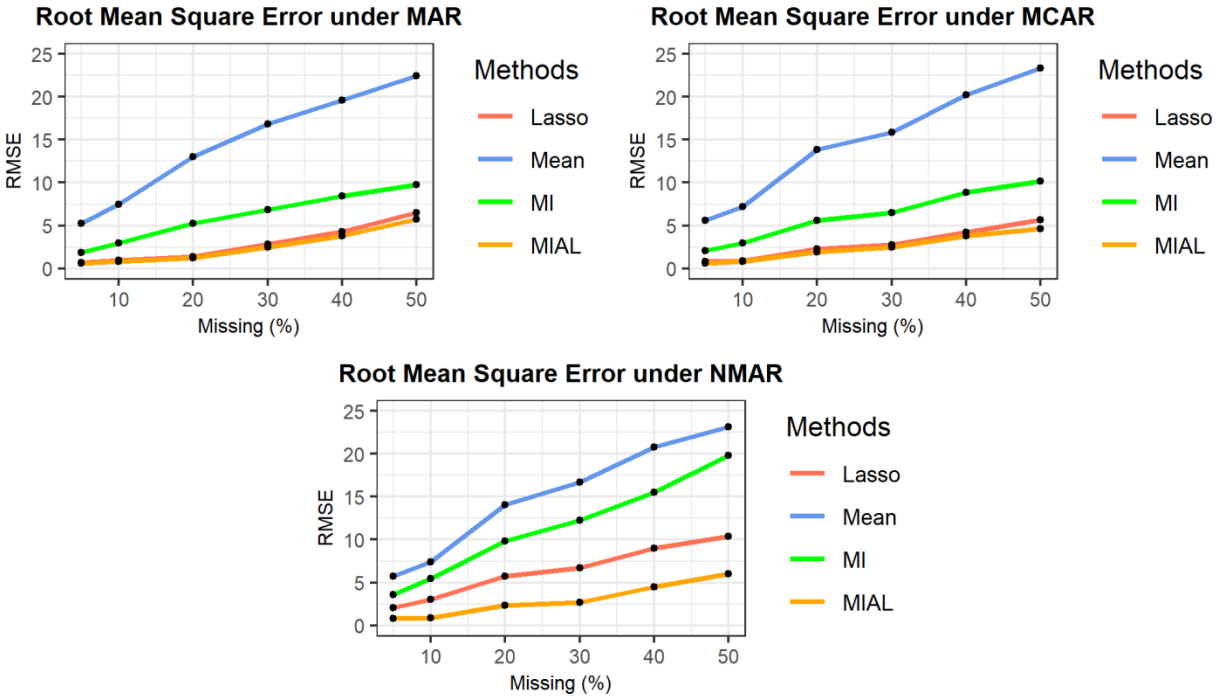

Fig.6.Root Mean Square under MAR, MCAR, NMAR.

From the results obtained in fig.6 , it is confirmed that irrespective of the type of missing attribute, MIAL shows very good performance at different missing rates varying from 5% to 50%.The classification error % of MIAL, Mean Imputation, Multiple Imputation and Lasso

Imputation are given in Table.7,8 for both MAR and MCAR. MIAL shows comparatively much better performance than other missing data handling methods.

Table.7. MAE of MIAL method with other methods under MAR

Missing Rate (%)	Mean Imputation	Multiple Imputation	Lasso Imputation	MIAL Imputation
5	8.27	4.59	1.68	1.55
10	12.68	9.37	2.36	2.15
20	20.36	16.3	6.95	6.3
30	22.48	18.68	8.86	8.44
40	29.51	20.07	11.37	10.96
50	37.26	23.79	13.59	12.44

Table.8. MAE of MIAL method with other methods under MCAR

Missing Rate (%)	Mean Imputation	Multiple Imputation	Lasso Imputation	MIAL Imputation
5	9.45	5.82	2.14	2.02
10	12.85	8.57	2.66	2.54
20	19.17	17.34	6.08	5.93
30	21.89	18.8	10.72	9.21
40	29.76	21.13	12.41	11.36
50	38.52	25.86	13.98	12.89

Table.9. MAE of MIAL method with other methods under NMAR

Missing Rate (%)	Mean Imputation	Multiple Imputation	Lasso Imputation	MIAL Imputation
5	9.12	5.38	2.67	2.08
10	12.34	9.87	2.34	2.15
20	16.5	13.07	5.98	4.99
30	20.91	16.78	8.05	6.44
40	22.04	19.44	10.07	9.68
50	27.75	20.9	11.26	11.19

MAE of MIAL in imputing non-ignorablemissing values is compared with the existing methods and given in Table 9. It is observed that MIAL shows very good performance than other methods. From the fig.6 it is observed that MIAL shows very good performance than other methods. It also observed the classification error increases as the missing proportion increases.

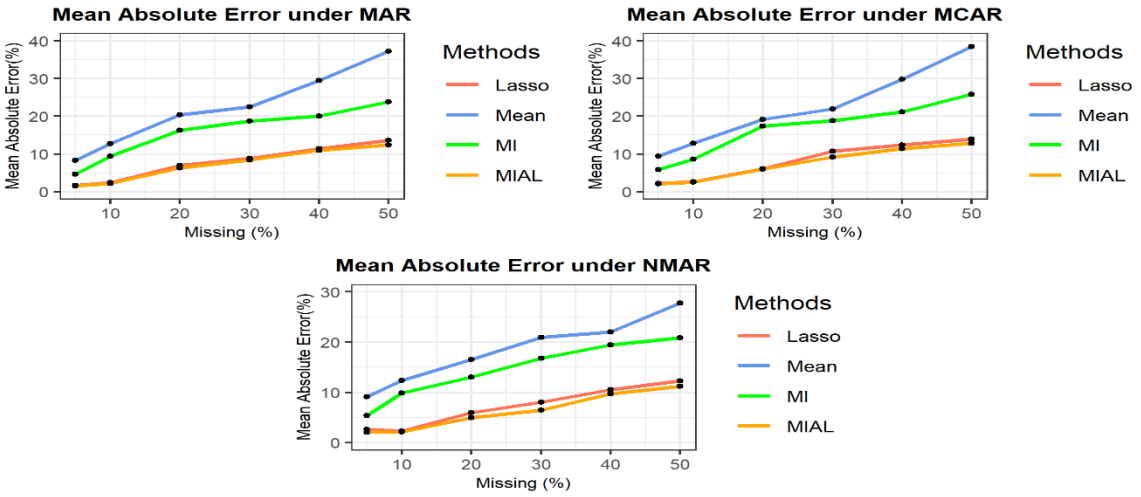

Fig.7.Mean Absolute Error under MAR, MCAR, NMAR.

By looking at the results it is observed that MIAL methods generates better RMSE and MAE values in MAR, MCAR when compared to NMAR.

6.Conclusion

The missing covariate values can be effectively imputed as the datasets contain detailed patient information, thereby ensuring that and correct parameter assumptions are made. At separate missing frequencies, MIAL actually imputes the missing values of both constant and discrete attributes. Missing data is inevitable in clinical trials. Hence, MIAL may be extremely useful in such studies, as making the correct inferences from data with missing values is a critical element in medical research. The key benefit is that MIAL uses Bayesian analysis, which is a straightforward approach for discrete and continuous attributes that uses all of the data in the dataset. According to the findings of three cases examined in this study, MIAL works better and is optimal for all types of missing data processing in clinical trials.

References

1. S. Aich, H. Kim, K. younga, K. L. Hui, A. A. Al-Absi and M. Sain, "A Supervised Machine Learning Approach using Different Feature Selection Techniques on Voice Datasets for Prediction of Parkinson's Disease," 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang Kwangwoon_Do, Korea (South), 2019, pp. 1116-1121.DOI: 10.23919/ICACT.2019.8701961

2. G. Kiss, A. B. Takács, D. Sztahó and K. Vicsi, "Detection Possibilities of Depression and Parkinson's disease Based on the Ratio of Transient Parts of the Speech," 2018 9th IEEE International Conference on Cognitive Info communications (CogInfoCom), Budapest, Hungary, 2018, pp. 000165-000168.DOI: 10.1109/CogInfoCom.2018.8639901

3. S. Aich, M. Sain, J. Park, K. Choi and H. Kim, "A text mining approach to identify the relationship between gait-Parkinson's disease (PD) from PD based research articles," 2017

International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 481-485.DOI: 10.1109/ICICI.2017.8365398

4. R. Guzman-Cabrera, M. Gomez-Sarabia, M. Torres-Cisneros, M. A. Escobar-Acevedo and J. R. Guzman-Sepulveda, "Parkinson's disease: Improved diagnosis using image processing," 2017 Photonics North (PN), Ottawa, ON, 2017, pp. 1-1. DOI: 10.1109/PN.2017.8090549

5. S. Sivaranjini and C. M. Sujatha, "Analysis of Parkinson's Disease SPECT Images Using Geometric Measures and Orthogonal Moments," 2018 Fourth International Conference on Biosignals, Images and Instrumentation (ICBSII), Chennai, 2018, pp. 206-212.DOI: 10.1109/ICBSII.2018.8524601

6. M. Heijmans, J. Habets, M. Kuijf, P. Kubben and C. Herff, "Evaluation of Parkinson's Disease at Home: Predicting Tremor from Wearable Sensors," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 584-587**.**DOI: 10.1109/EMBC.2019.8857717

7. A. F. Henao-Martinez, K. Colborn, and G. Parra-Henao. Overcoming researchbarriers in Chagas disease-designing effective implementation science. *Parasitol.Res.*, 116(1):35–44, Jan 2017.

8. N. M. Bowman, V. Kawai, R. H. Gilman, C. Bocangel, G. Galdos-Cardenas,L. Cabrera, M. Z. Levy, J. G. Cornejo del Carpio, F. Delgado, L. Rosenthal, V. V.Pinedo-Cancino, F. Steurer, A. E. Seitz, J. H. Maguire, and C. Bern. Autonomicdysfunction and risk factors associated with Trypanosoma cruzi infection amongchildren in Arequipa, Peru. *Am. J. Trop. Med. Hyg.*, 84(1):85–90, Jan 2011.

9. Y. Z. Castellanos-Dominguez, Z. M. Cucunuba, L. C. Orozco, C. A. Valencia-Hernandez, C. M. Leon, A. C. Florez, L. Munoz, P. Pavia, M. Montilla, L. M.Uribe, C. Garcia, W. Ardila, R. S. Nicholls, and C. J. Puerta. Risk factors associated with Chagas disease in pregnant women in Santander, a highly endemicColombian area. *Trop. Med. Int. Health*, 21(1):140–148, Jan 2016.

10. Z. M. Cucunuba, A. C. Florez, A. Cardenas, P. Pavia, M. Montilla, R. Aldana,K. Villamizar, L. C. Rios, R. S. Nicholls, and C. J. Puerta. Prevalence and riskfactors for Chagas disease in pregnant women in Casanare, Colombia. *Am. J.Trop. Med. Hyg.*, 87(5):837–842, Nov 2012.

11. R. E. Gurtler, R. Chuit, M. C. Cecere, M. B. Castanera, J. E. Cohen, andE. L. Segura. Household prevalence of seropositivity for Trypanosoma cruzi inthree rural villages in northwest Argentina: environmental, demographic, andentomologic associations. *Am. J. Trop. Med. Hyg.*, 59(5):741–749, Nov 1998.