A novel optimization algorithm for the missing data in HCC based on multiple imputation and genetic algorithm

Yasser Salaheldin
Information System Departmen,
Faculty of Computers and Information,
Menoufia Univerisity, Shibin el-Kom,
Menoufia, Egypt
yasser.salah@ci.menofia.edu.eg

Mohamed Hammad
Information Technology Departmen,
Faculty of Computers and Information,
Menoufia Univerisity, Shibin el-Kom,
Menoufia, Egypt
mohammed.adel@ci.menofia.edu.eg

Hatem Abdelkader
Information System Departmen,
Faculty of Computers and Information,
Menoufia Univerisity, Shibin el-Kom,
Menoufia, Egypt
hatem.abdelkader@ci.menofia.edu.eg

***Abstract*—Hepatocellular carcinoma (HCC) is a threat to the liver, which is considered one of the diseases devastating to human health that leads to death. Therefore, discovering HCC early is essential, this will not begin without complete, adequate, and reliable data. Hence, it is imperative to improve missing data completion processes to provide more reliable data in the detection phase. In this research, we offer a unique method that combines multiple imputations with a genetic algorithm to optimize multiple regression imputation processes and obtain the optimum fitness values for missing data from patients. We used 583 patient records from a public, available database to train and evaluate our proposed algorithm, separated into 416 liver patient records and 167 non-liver patient records. Results are proven that the proposed approach has the most improvement for missing data results. We were able to reach the optimal value which was measured by fitness value to 233 instead of using the normal equation in multiple imputations which gave 92.72 as the uttermost fitness value of it. The suggested model may be validated using a large database and used in HCC laboratories to assist doctors in making an accurate diagnosis.**

Keywords—HCC; Multiple Imputation; Fitness Value; Multiple Regression; Genetic Algorithm; Missing Data; Optimization.

I. INTRODUCTION

The liver is threatened by hepatocellular carcinoma (HCC) that most commonly occurs in persons who have chronic liver disease or who are at risk of developing cirrhosis [1]. Based on the most recent evidence, HCC is the deadliest malignant tumor on the planet, inflicting more than 600,000 fatalities each year [2]. In 2012 According to a report issued by the World Health Organization (WHO), around 14.1 million new malignancy cases were diagnosed, with about 8.2 million passing overall. Therefore, previous researchers proposed computer-aided diagnosis systems, which can be employed to aid the physician in their decision-making process [3]. Image processing, data analysis, and artificial intelligence technologies show potential in research applications for the effective characterization of liver cancer. Thus, in the medical and healthcare fields, automated systems may aid physicians in making accurate and timely diagnoses of their patients' different ailments (e.g., accurate clinical decision support systems (CDSS)) that must be constructed using patient data-driven [4]. The most curative intervention is liver transplantation. Transplantation options are restricted in individuals with HCC because of their advanced stage upon diagnosis. Thus, the most effective technique for improving patient prognosis is, in theory, early diagnosis and prevention of HCC development [5,6].

In the social, behavioral, and medical sciences, missing values are pervasive. Academics have relied on a variety of ad hoc methods to "repair" data for decades, such as eliminating incomplete instances or replacing missing values. A few missing worth attribution approaches execution relies upon the size of the dataset and the quantity of missing qualities inside the dataset [7]. Unfortunately, the bulk of these solutions are prone to severe bias since they rely on a relatively rigid assumption about the cause of missing data. Despite the fact that these approaches have fallen out of favor in the methodological literature [8], they are nevertheless widely used in research articles [9]. Additionally, missing data is a persistent issue in practically every area that makes use of empirical research methodologies [10]. In addition, missing data on multi-item instruments is prevalent in epidemiological and medical studies [11]. While missing data are an inevitable part of epidemiological and clinical research, their potential to damage the validity of study findings is sometimes neglected [12]. In decision science and the investigation of physical systems, optimization is a useful tool [13]. As a result, the goal of this study is to offer a novel efficient technique for solving global problems with no constraints. In the new optimized algorithm, a new directed crossover rule is introduced based on creation of combinations the best and the local individuals of a particular generation. The major contributions of this work are cleared as follow:

1- Propose a new approach for imputing missing data that uses the Multiple imputation (MICE) algorithm to fill in the gaps.

2- Optimize the multiple imputation algorithm based on the Genetic Algorithm (GA), which achieves high accuracy comparing with other previous methods.

3- Dealing with missing data in HCC using the optimized method.

II. LITERATURE REVIEW

Several academics have recently used machine learning algorithms to replace or impute missing data in datasets, particularly in medical diagnosis [11, 14-18]. Below, we have shown that as succinctly as possible:

- Wojciech et al. [14], used seven classifiers, including the k-nearest neighbor's algorithm (KNN), to fill in the missing examples in a machine learning model. 165 HCC patients were used to train and analyze the proposed

approach. They had the highest accuracy and F1-score, respectively, of 0.9030 and 0.8857.

- Iris *et al.* [11], examined the effectiveness of simple and more advanced procedures for dealing with missing data in multi-item instruments when some or all item scores are absent. To simulate real-world missing data conditions, a multi-item variable was used as a covariate in a linear regression model. The authors used to bias and coverage as performance criteria to compare fitted regression results. Mean imputation resulted in biased estimates in every missing data circumstance where more than 10% of individuals had missing data. Furthermore, when a large number of people were missing key components.
- Wojciech *et al.* [15], investigated a ML way to deal with identify HCC utilizing 165 patients. Ten notable ML calculations are utilized. In the preprocessing step, the standardization approach is utilized. The GA combined with a separated 5-crease cross-approval strategy is applied twice, first for boundary advancement and afterward for include choice. In this work, support vector machine (SVM) (type C-SVC) with new 2level genetic streamlining agent and element choice yielded the most elevated exactness and F1-Score of 0.8849 and 0.8762.
- Jared and Jerome [16], presented a nonparametric Bayesian joint model for multivariate continuous and categorical variables in order to provide a versatile engine for multiple value imputation. The model incorporated Dirichlet process mixtures of multinomial distributions and Dirichlet process mixtures of multivariate normal distributions for categorical data. The model integrated the Dirichlet process mixes of multivariate normal distributions for continuous variables.
- Mokrane *et al*. [17], used quantitative imaging characteristics retrieved from triphasic CT images to assist physicians in making more accurate diagnoses of HCC in cirrhotic patients with ambiguous liver nodules. They employed machine-learning approaches to train and calibrate the signature (finding cohort) and verify the signature (validation cohort) in order to categorize hepatic nodules as HCC vs. non-HCC, but the study's witness was unable to get the missing data, putting the study's veracity in jeopardy.
- To reduce the risk of bias, Janus *et al.* [18] suggested effective approaches for managing missing data. They looked into how to make the most of missing data management during the design stage of a randomized clinical trial, and they suggested analytical approaches that could assist remove bias caused by inevitable missing data. Finally, a practical guide and flowcharts describing when and how to use multiple imputations to address missing data in randomized clinical trials were provided; however, they only stated when and how multiple imputations should be used, not whether or not they should be optimized, which is exactly what we did in our study.

III. MATERIALS AND METHOD

The materials and, as a result, the technique used in our suggested investigation are discussed in this part. Furthermore, we frequently turn to the HCC dataset while putting our methodology into practice.

A. HCC Dataset

The Indian liver patient dataset [19] was utilized to assess prediction algorithms with the goal of alleviating physician workload. This data set includes 416 liver patient records and 167 non-liver patient records acquired in Andhra Pradesh's north-eastern area. A class label is used to categorize groups as liver patients in the dataset column (with or without liver disease). There are 441 records for male patients and 142 records for female patients in this data set. Any patient over the age of 89 is said to be in the 90s. The variables include the patient's age, gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, Albumin to Globulin Ratio, and Dataset: field, as shown in Table 1.

B. Methodology

This section describes the different stages of the proposed methodology with missing value, preprocessing, GA (GA features selection, GA parameter optimization) and classification steps. The main parts of each stage are briefly explained below:

1. Genetic Algorithm (GA)

GA are search techniques based on the notions of natural selection and genetics and inspired by the biological evolution of living creatures. They were first presented in the 1970s by J Holland. [3]. Genetic algorithms abstract the problem area into a population of people and iteratively search for the fittest person. GA grows from a population of starting people to a population of high-quality people, each of which offers a potential solution to the issue at hand. Each rule's quality is quantified using a fitness function, which is a quantitative description of the rule's adaptation to a particular environment. The technique begins with a randomly created starting population of people. In each generation, three fundamental genetic operators are applied sequentially to each individual with predetermined probabilities, namely selection, crossover, and mutation [3, 20]. We tweak the thetas and find the optimal solution using GA as shown in Figure 1.

Table 1: Sample of Indian liver patient (HCC Dataset)

Age	Gender	Total Bilirubin	Direct Bilirubin	Alkaline Phosphatase	Alamine Aminotransferase	Aspartate Aminotransferase	Total Proteins	Albumin	Albumin and Globulin Ratio	Dataset
65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
58	Male	1	0.4	182	14	20	6.8	3.4	1	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
26	Female	0.9	0.2	154	16	12	7	3.5	1	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1

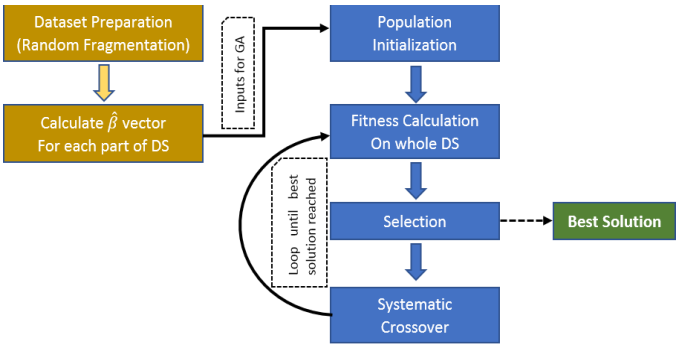

Fig. 1: Block diagram of the proposed model to optimize MI optimization on HCC dataset

2. Multiple Imputation (MI)

A particular execution of this methodology in which each factor is credited restrictive on any remaining factors is known as the multivariate imputations by chained equations (MICE) [19]. Multiple imputations are used in this step after the optimization phase. It is made to fill in the gaps in the HCC database's missing values. The MICE algorithm was used to fill in the missing values. MI is the only method for dealing with missing values that is computationally simple, versatile, relatively easy to apply, and increasingly available in standard statistical software [22]. However, MI is not the only principled method for dealing with missing values, nor is it always the best for any given problem. A weighted estimating approach [23] can be used to obtain good estimates in some instances. The dependent variable y_i is associated with two or more independent variables x_{i1}, x_{i2}, and x_{ik}, according to the multiple linear regression model. For *k* variables, the general model is of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \qquad (1)$$

The least-squares method is used to estimate the regression coefficients in Equ.1 in multiple linear regression analysis. The regression coefficients show the independent variables' unrelated contributions to predicting the dependent variable. In contrast to basic linear regression, judgments about the degree of interaction or correlation between the independent variables must be made [24]. In terms of vectors representing observations, levels of regressor variables, regression coefficients, and random errors, using matrices provides for a more compact framework. The model is of the form, and we have it when we write it in matrix notation.

$$y = x\beta \qquad (2)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \qquad (3)$$

Y is a n1 dimensional random vector containing the data, X is a n(k1) matrix generated by the predictors and is a (k+1) 1 vector of unknown parameters [24].

$$\hat{\beta} = (x^T x)^{-1} x^T y \qquad (4)$$

In multiple linear regression analysis, the initial step is to find the vector of least squares estimators, which gives the linear combination y that minimizes the error. These methods demonstrate how to find the least square estimators using matrix algebra. Recall the least squares estimators from the previous stage (4). Algorithm 1 shows the steps of the proposed technique.

In step 1 and 2 we just prepare our dataset to get a random sorting to start the processing. Then, step 3 we fragment the dataset to getting multiple squares estimators as shown in Fig. 1 (here we took 6 parts only as cleared in step 4), thus we determined the vector of least squares estimators. Then, we took the determined vector of least squares estimators as a population initialization and calculate the fitness function $F(c)$ (as shown up in step 8) on whole dataset using eq. 1 as follow:

$$F(c) = \frac{1}{|y - \hat{y}|} \qquad (5)$$

After calculation of fitness function for all individuals of the determined vectors of least squares estimators for all variables (in this case we have 11 variable), in (a) step 8 we make a comparison to get the individuals which give a highest qualities (we took top 3 individuals in (b) step 8) as a second phase in GA. And then we start our systematic crossover between these 3 vectors of least squares estimators in (c) step 8 as shown up, via replacing one by one in each individual, subsequently in our case we got 36 new individuals. From here we have been starting the iteration from eq. (5) to calculate the F(C) for all new individuals for electing the best individuals which give the highest qualities. In steps (9, 10) after we obtained the best individuals we resumption the MI steps to and calculate the mean value of our 6 results of missing value using eq. 6 (*Fitness* Value) equate.

IV. RESULTS

The Fitness Value for each least squares estimator was chosen as a fitness function of genetic algorithm as the measurement that we used. In step 5, you will find a formula. The model was created utilizing the records of 416 HCC patients and 167 non-HCC patients. Section III contains a detailed description of the dataset. Calculation of fitness values was rated as follow:

$$Fitness \text{ Value} = \left(\sum_{i=1}^{n} \frac{1}{|y_i - \hat{y}_i|} \right) / n \qquad (6)$$

This fitness value got where n = number of rows, *i* is number of iteration while reach to n which in this case is 583 records. We derived the results from previous equation in step 6 and show variation of fitness values for various epochs of fitness value function after our optimization as shown in Fig.2 that clarifies the variation of the outputs without any optimization, outputs after first generation and outputs from second generation.

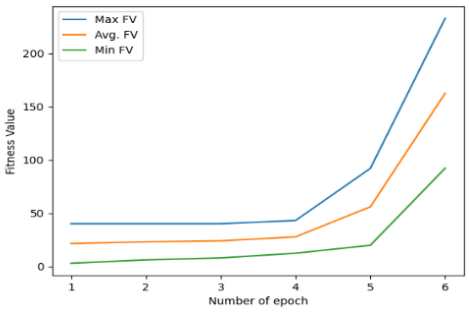

Fig. 2: Variation of fitness value for various epochs after GA optimization.

Algorithm 1: Steps of the proposed method

```
1-  Input: dataset <- read_csv("indian_liver_patient.csv")
2-  Sorting Dataset Randomly
3-  # Dataset defragment to 10 parts
        a. dataset_part1 <- dataset[:49] ... to dataset_part10 <- dataset[450:499]
4-  y <- 'Total_Bilirubin'          # desired output column
    theta1 <- mice.calc_theta(theta_dataset_part1, y) ... to theta10 <- mice.calc_theta(theta_dataset_part10, y)
    theta_dataframe <- DataFrame() # to store all theta individuals in dataframe
5-  # Start GA steps
    solution_per_population <- 36
6-  number_of_parents_individuals <- 3
7-  new_population <- theta_dataframe
8-  for iteration in range(10): # Until get the best solution
        a. # Calculate qualities (FC) to select the highst 6 individuals
           qualities <- ga.population_fitness(new_population, dataset, y, desired_output)
        b. # Selecting the best parents in the population for mating
           parents <- ga.select_mating(new_population, qualities, number_of_parents_individuals)
        c. # Generating next generation using crossover
           new_population <- ga.crossover(parents, solution_per_population)
    ENDFOR
9-  i <- 0
    yh <- 0              # imputed output
    for i < 6:
        yh <- mice.calc_y(new_population[i:i+1], dataset, r in rows, desired_output)
        yh <- yh +` yh
        i <- i++
    ENDFOR
10- Output: yh = yh/5            # the mean of fifth imputed values (last phase of multiple imputation)
```

As previously stated, when compared to multiple imputation, our proposed approach produced the best results. Our novel combination of genetic algorithms with multiple regression – systematic crossover produced the greatest

results in diagnosing HCC disease by impute the best values of missing data in the patients' dataset.

In table (2) a comparison between the techniques which used in missing data imputation phase of mentioned studies

and our method, therefore it turns us to implement the same experimental with these techniques to shine up our improvement, in fact we did it and we could have discovered the differences between our optimized algorithm and these algorithms as illustrative in table 4, It shows the degree of these algorithms results accuracy compared to the results which gained from our algorithm.

- **Evaluation metrics:**

We have a tendency to employ well-known basic measures like accuracy, which is generated in Equ.7, to evaluate the efficacy of our methods. The confusion matrix was supported by the measurements we computed. A basic example of a confusion matrix is shown in Table (3).

Table 2: The used techniques for missing data of mentioned studies:

Study	Missing Data Technique
Our study	Optimized MI
Wojciech *et al.* [14]	K-nearest neighbor (KNN)
Iris *et al*. [11]	Image dataset and not dealing with missing data
Wojciech *et al.* [15]	K-nearest neighbor (KNN)
Jared and Jerome [16]	Normal Eq. Of MI
Mokrane *et al*. [17]	No missing data mentioned
Janus *et al*. [18]	It just approaches for managing missing data.

In table (2) we also aimed to narrative the reasons of referring to these studies in our work, for instance Wojciech et al. [14] in HCC domain and dealing with missing data, Iris et al. [11] in HCC images dataset and known from them more about HCC detection, Wojciech *et al*. [15] in HCC domain and

dealing with missing data, we have been resorted to this paper to compare a lot of methods in the same issue, Jared and Jerome [16] in missing data MI which our goal to improvement and finally Janus *et al.* [18] whose prove that we dealing with clinical missing data correctly throughout their flowchart.

Table 3: Confusion matrix after using optimized algorithm:

Table: Confusion matrix after the optimization		
Actual	Predicted	
	P	N
T	255	66
F	48	210

$$Accuarcy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (7)$$

On the testing set, the optimization attained an accuracy rate of 80.30 percent, up from 19.7% previously. The confusion matrix for this optimized approach is shown in Table (3). When all of the stages were completed, the MI's performance rose by around 60%, also when compared our results with KNN results it gave us 16.92%, all differences are clearly through table (4) when implemented these algorithms on the same dataset.

Table 4: Algorithms results accuracy comparison:

Algorithm	compared results acc. with our results on the same dataset
Optimized MI	80.30 %
MI	19.70 %
KNN	16.92 %

V. CONCLUSION

We present here a novel approach based on a mix of multiple imputation and genetic optimization for a critical phase of HCC illness identification that imputes missing data in this study. We have used different approach in GA crossover which not random and get the best fitness value of least squares estimators on the HCC dataset. This model enabled more accurate prediction of missing values which in turn improve the detection of HCC than previous models presented in the literature. Our proposed method obtained an overall fitness value 233 from 2.686 in normal equation of multiple imputation and getting accuracy 80.30% with accuracy improvement nearly 60%. As a result, our optimization is a useful tool for filling in lacking data and performing an accurate and consistent HCC diagnosis.

REFERENCES

[1] Zhu, R. X., Seto, W. K., Lai, C. L., & Yuen, M. F. (2016). Epidemiology of hepatocellular carcinoma in the Asia-Pacific region. Gut and liver, 10(3), 332.

[2] Kumar, M., Zhao, X., & Wang, X. W. (2011). Molecular carcinogenesis of hepatocellular carcinoma and intrahepatic cholangiocarcinoma: one step closer to personalized medicine?. Cell & bioscience, 1(1), 1-13.

[3] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 23(1), 89-109.

[4] Chen, E. L., Chung, P. C., Chen, C. L., Tsai, H. M., & Chang, C. I. (1998). An automatic diagnostic system for CT liver image classification. IEEE transactions on biomedical engineering, 45(6), 783-794.

[5] Ho, E. Y., Cozen, M. L., Shen, H., Lerrigo, R., Trimble, E., Ryan, J. C., ... & HOVAS Group. (2014). Expanded use of aggressive therapies improves survival in early and intermediate hepatocellular carcinoma. Hpb, 16(8), 758-767.

[6] Fujiwara, N., Friedman, S. L., Goossens, N., & Hoshida, Y. (2018). Risk factors and prevention of hepatocellular carcinoma in the era of precision medicine. Journal of hepatology, 68(3), 526-549.

[7] Mostafa, Samih M., et al. "CBRG: A Novel Algorithm for Handling Missing Data Using Bayesian Ridge Regression and Feature Selection Based on Gain Ratio." *IEEE Access* 8 (2020): 216969-216985.

[8] Little, Roderick JA, and Donald B. Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019.

[9] Bodner, Todd E. "Missing data: Prevalence and reporting practices." *Psychological Reports* 99.3 (2006): 675-680.

[10] Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., & Sun, Q. (2018). Deep learning for image-based cancer detection and diagnosis− A survey. Pattern Recognition, 83, 134-149.

[11] Eekhout, I., de Vet, H. C., Twisk, J. W., Brand, J. P., de Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of clinical epidemiology, 67(3), 335-342.

[12] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338.

[13] Nocedal, J., & Wright, S. (2006). Numerical optimization. Springer Science & Business Media.

[14] Książek, W., Hammad, M., Pławiak, P., Acharya, U. R., & Tadeusiewicz, R. (2020). Development of novel ensemble model using stacking learning and evolutionary computation techniques for automated hepatocellular carcinoma detection. Biocybernetics and Biomedical Engineering, 40(4), 1512-1524.

[15] Książek, Wojciech, et al. "A novel machine learning approach for early detection of hepatocellular carcinoma patients." *Cognitive Systems Research* 54 (2019): 116-127.

[16] Murray, J. S., & Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. Journal of the American Statistical Association, 111(516), 1466-1479.

[17] Mokrane, F. Z., Lu, L., Vavasseur, A., Otal, P., Peron, J. M., Luk, L., ... & Dercle, L. (2020). Radiomics machine-learning signature for diagnosis of hepatocellular carcinoma in cirrhotic patients with indeterminate liver nodules. European radiology, 30(1), 558-570.

[18] Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. BMC medical research methodology, 17(1), 1-10.

[19] Dataset, Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. Inspiration, download link: https://www.kaggle.com/uciml/indian-liver-patient-records#indian_liver_patient.csv

[20] Haldurai, L., Madhubala, T., & Rajalakshmi, R. (2016). A study on genetic algorithm and its applications. International Journal of Computer Sciences and Engineering, 4(10), 139.

[21] Austin, Peter C., et al. "Missing data in clinical research: a tutorial on multiple imputation." *Canadian Journal of Cardiology* (2020).

[22] Liu, Y., & De, A. (2015). Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. International journal of statistics in medical research, 4(3), 287.

[23] Schafer, J. L. (1999). Multiple imputation: a primer. Statistical methods in medical research, 8(1), 3-15.

[24] Brown, S. H. (2009). Multiple linear regression analysis: a matrix approach with MATLAB. Alabama Journal of Mathematics, 34, 1-3.