IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-11, Nov 2019 E-ISSN: 2347-2693

A Novel Approach for Missing Value Replacement in MLP-RMSProp Based Classification Model

G. Suresh[1*], S. Saraswathi[2]

[1]Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India
[2]Dept. of Information Technology, Pondicherry Engineering College, Puducherry, India

[*]*Corresponding Author: sureshg2233@yahoo.co.in Tel.: +917010693133*

DOI: https://doi.org/10.26438/ijcse/v7i11.619 | Available online at: www.ijcseonline.org

Accepted: 10/Nov/2019, Published: 30/Nov/2019

Abstract— Data Preprocessing has become a vital task to be carried out in the Data Mining process. The data becomes the most important resource due to its significance in various domains. However, it is hard to gather every data and saves it in ^{real}-time that lead to few missing data. It is not preferable to omit the missing data due to the fact that even a few amount of data acts as a significant part in the outcome. Missing value replacement acts as a main process to handle missing data prior to the prediction of hidden pattern, that exist in the dataset. This paper presents a new, Linear Regression based missing valve replacement in the MLP-RMSprop based classification model to handle missing data. Here, linear regression model is applied to predict the values to replace the missing data, which will help to improve the classification process. Then, multilayer perceptron (MLP) classifier is applied to classify the data which further tuned by the use of root mean square propagation (RMSProp) model. An extensive implementation takes place on three benchmark dataset to showcase the betterment of the presented model. The resultant values from simulation indicated that the projected model offered supreme performance over the other models.

Keywords— Missing value; Classification; RMSProp; Linear Regression

I. INTRODUCTION

Generally, in many types of scientific analysis using data, loss of data is a common problem. The idea of examining the correct model for the presence of missing data is a common issue, and determining the right approach for mitigating this often grows as a most promising issue [1]. Using the growing interest in data dependent tools such as machine learning (ML), quality of data always becomes a main driving unit for final result, i.e., better data quality and outcome. Loss of data from dataset, i.e. missing values in dataset harmfully influences the quality of data and also the resultant knowledge is identified. A main task in data search is to gather the related data. Practically, grouped datasets are not in full-fledged form as well as raddled with missing values. This happens due to different reasons such as physical error during the time of data entry, no values are found, machine errors, inaccurate value, natural problems, etc. The values present in datasets are lost more than one value over definite number of record known as datasets with missing values. These missing values are classified into 3 categories and each one requires various way of handling the issue. They are termed as Missing Completely at Random (MCAR), where the missed value of an attribute is independent of value from attribute of its own, but on some other feature. In Missing Not a Random (MNAR), the attribute lost is based on the corresponding value of the own attribute. For handling missed values, deletion mechanism is utilized whereas definite records consist of variables using missed values which are not assumed for analyzing purposes.

Imputation models are developed for managing the missing values, adaptable value is applied for replacing the concerned missing values [2]. The imputed value may be mode, mean, median, or different predefined value of the parameter consist of missing values, or obtained using few prediction models. Imputation techniques are utilized in handling missing values, could be MAR or MCAR type, and whenever every record or variable in dataset is most important and individual record does not contains missing values over several parameters. In MCAR type, missing values in database could be managed using deletion or preferably list wise deletion. There are no techniques for handling MNAR type of missing values.

There is a need of using some techniques for taking care of the missing values from database; many methods from deletion to imputation are suggested in this study. Additionally, increased number of likelihood methods is used for managing missing values. Deletion could be complete deletion, list-wise deletion, else Complete Case Analysis, where each rows comprises of more than one attribute values

missing or removed or specific deletion, where the rows deleted contains predefined percentage about their attribute missing value. There could also be variable deletion or pair-wise deletion, where every row includes missing values for variables in the current analysis is removed. Then, these rows anyway going to be utilized for other analyzing purpose that is not involved in concern variable contain missing value. In a very rare situation, every parameter having missing values over different records might end in deleting the entire database.

An imputation method applies the data accessible in elected dataset for estimating the missing value, whereas appropriate value is imputed instead of missed value. In [3], the value can be retrieved from case substitution. Imputed value can also be processed with the help of regression methods, KNN, expectation maximization (EM) imputation, hot deck and cold deck imputation. Prediction models comprises of some techniques, a model is designed on the basis of previous data that is used for predicting appropriate values for missing data. When it comes to mean and mode imputation models, the missing value of an attribute gets substituted with mean or mode of familiar values of the attribute [4]. Case substitution is utilized for handling missing value in testing surveys, whereas single example of missed data is imputed with alternate non-sampled instances. In Hot deck method work, the missed values are evaluated from present data, whereas in cold deck model, it is evaluated with the help of data source rather than present dataset. KNN applies the k nearest cases for imputing missing values. Between KNN and EM imputations, KNN imputation results outperform the EM. Future imputation can be undergone either single or multiple imputation.

Single imputation has one appropriate value that is imputed instead of missed value. Multiple imputation has complete datasets, which is obtained by the imputation of missing values by m times, the resultant imputed dataset are fixed as weighted average of these m datasets. Multiple imputationscontain different beneficial elements compared with secondary models, like single imputation, maximum likelihood techniques, and deletion; it is in need of several resources. Single imputation model deals every values including imputed value as true and do not consider for missing value which is indefinite that leads for inflated type I error metrics. Maximum likelihood techniques have complexity in implementing nonstandard models.

There are also Inverse Probability Weighing (IPW) approaches for managing missing data that utilizes the opposite of monitored probability for weight observed records, also represents the entire data with missing values. Imputation method performs in a better way [5]. Local Least Squares (LLS), Least Squares Adaptive (LSA) and Bayesian Principal Component Analysis (BPCA) [6], which are few optimal imputing models that are reliable to other two dataset. Few alternate methods are used in multiple imputations, Ordinary Least Squares (OLS), Singular Value Decomposition (SVD) and Partial Least Squares (PLS). Globally used imputation methods are PLS, SVD and BPCA offers extended results on database with lower difficulty, while neighbour dependent models namely KNN, OLS, LLS and LSA performs better with maximum difficulty.

Existing studies consist of other models and the calculations local Least Square_gene (LSI_gene), LSI_adaptive, LSI_array, LSI_combined, Local Least Square imputation (LLSI) and EM_gene techniques outperform BPCA. Fully Conditional Specification (FCS) and Multi Variant Normal Imputation (MVNI) techniques are usually low biased and generate same solution in spite of datasets having binary and ordinal parameters. MVNI generates easy method of specification; therefore some people might face the issue with unrealistic behaviour of the multivariate normal assumption. FCS requires an independent regression model for every variable whose value undergoes imputation and thus includes tedious model specification [7]. FINNIM is efficient nonparametric iterative multiple imputation method that is performed with the help of KNN for estimating missing value [8].

Kernel extension methods denotes competitive performance while handling missing values in datasets that consist of binary values when compared with multiple imputation with the help of SVM [9]. Multiple imputation use sequential regression tree as conditional scheme that is capable of capturing difficult relation and needs minimum tuning by the client [10]. Normal Imputation (Nm) Method, posterior normal distribution technique is applied for imputing missing values and depends on regression coefficients. Mean and variance (MV) approach utilizes empirical distribution of standardized residual, whereas each missing value is evaluated with the addition of predictive means of residuals. In the Predictive Mean Matching (PMM) method, the missing value is imputed through monitoring randomly picked from a group of observed cases whose predictive mean is adjacent to that of missing values. Local Residual Draw (LRD) method operates imputation of missing value using the predictive mean in case of PMM, in conjunction with residual randomly drawn from set of observed residual case using predictive means that is nearby for identifying missing value.

Bounded missing value imputation, Proportioned Residual Draw (PRD), Predictive Mean Matching-Proportioned Residual Draw (PMM-PRD) techniques plays extended performance when compared to Normal imputation Method (NM), Mean and Variance (MV), PMM and Local Residual Draw (LRD) methods, when the count of boundaries are maximum [11]. Decision tree and Sampling based missing value Imputation (DSMI) method use correlation technique for managing missing values by imputation. It categorizes

dataset in various segments horizontally based on non-missing attributes of missing records and missing value, imputation is performed with the help of samples from the distributions induced through correlation process. DSMI performs better than k Nearest Neighbour Imputation (KNNI) and secondary well-known imputation technique [12]. Multiple imputation method is obtained by the combination of Gaussian Mixture Model (GMM) and Extreme Learning Machine (ELM) that offers extended accuracy compared with other models based on conditional mean imputing performance but consumes maximum operational time [13].

Imputation acts as a main process to handle missing data prior to the prediction of hidden pattern exist in it. This paper presents a new imputation and classification model to handle missing data. Here, linear regression model is applied to predict the values to replace the missing data which will helps to improve the classification process. Then, multilayer perceptron (MLP) classifier is applied to classify the data which further tuned by the use of root mean square propagation (RMSProp) model. An extensive implementation takes place on three benchmark dataset to showcase the betterment of the presented model. The resultant values from simulation indicated that the projected model offered supreme performance over the other models.

II. THE PROPOSED MODEL

The working process of the projected model is shown in Fig. 1 which comprises several processes namely, pre-processing, missing value replacement, 10 fold cross validation (CV), and classification. Initially, the provided data is offered as input to the projected model. The data will be preprocessed through format conversion and data transformation task. Then, 10 fold cross validation (CV) process takes place to split the dataset into training and testing part. Finally, classifier is applied to classify the test cases under proper labels.

2.1. Preprocessing

In this phase, format conversion and data transformation processes are carried out. In the beginning, the actual dataset is transformed to an appropriate .csv or .arff format to make it well-suited for classification task. Afterwards, data transformation procedure will be invoked where the categorical values or any other kind of values will be transformed to a numerical values.

2.2. Linear Regression based Missing value replacement

Linear regression is a linear model, e.g. a model which considers a linear relationship among the input (x) and output parameters (y). Particularly, y can be determined from a linear amalgamation of the input parameters (x). In the existence of an individual input parameter (x), it is known as linear regression. In case of the existence of many input parameters, it is called as multiple linear regression. Many methods are devised to organize or train the linear regression equation from data and the most widely available one is known as Ordinary Least Squares (OLS).

2.2.1. Model Representation

Linear regression is a novel approach due to the nature of simpler representation. It integrates a particular collection of input values (x) of the solution to which is the predicted output for the corresponding collection set of input values (y). Hence, all the input and output values are in numeric format. The linear equation will allocate one scale factor to every input value or column, known as a coefficient and defined as B. An extra coefficient is also included providing the line an extra known as intercept or bias coefficient. For instance, in a simpler regression problem, the model can be defined as

$$y = B0 + B1 * x \quad (1)$$

Under high dimension problem, in case of multiple inputs, the line is known as a plane or hyper-plane. B0 and B1 indicate the coefficient values. If the value of coefficient comes to 0, it will efficiently eliminates the impact of the input parameter on the model and consequently from the prediction made from the model (0 * x = 0). It becomes related to the regularization models which modify the learning technique for minimizing the complexity of regression technique by applying pressure on the absolute size of the coefficients.

2.2.2. Learning the Model

Linear regression technique can be learnt through the estimation of the coefficient values employed in the available data representation. Some of the models available to learn the linear regression are provided here.

Simple Linear Regression

It can be used under the presence of one input and statistics can be used for coefficient estimation. It also needs to compute the statistical characteristics from the data namely mean, standard deviation (SD), correlation and covariance. Every data should be present for traversing and calculating the statistics.

OLS

It can be used in case of having multiple inputs to determine the coefficient values. It leads to the minimization of the sum of the squared residuals. It indicates that on giving a regression line using the data, the distance can be calculated from every data point to the regression line, squares it, and totalling every squared error altogether. It considers every data as a matrix and utilizes linear algebra functions for the estimation of the optimum values for the coefficients. It indicates that every data should exist and it requires

adequate storage area to fit the data and carry out the matrix calculations.

Gradient Descent
It is employed in case of having multiple inputs and optimize the coefficient values in an iterative way by the minimization of the error present in the model while training the data.

It is known as Gradient Descent and starts to operate with arbitrary values for every coefficient. The squared errors are summed and determined for every input as well as output values. A learning rate is employed as a scaling factor and the coefficients are updated in the direction with respect to error minimization. This procedure gets iterated till a least squared error is attained. By utilizing this technique, it is needed to choose a learning rate variable which computes the enhancement step size for taking every round.

Gradient descent is mostly trained by the use of linear regression technique due to the fact it is simpler to understand. In practice, it is useful when you have a very large dataset in either the number of rows or the number of columns that may not fit into memory.

2.2.3. Regularization
It can be used to extend the training process of the linear model. It seeks to the minimization of the total of squared error of the model on the training data, however, it also minimizes. A pair of examples of regularization procedure for linear regression is given below.

- **Lasso Regression:** OLS is altered for the minimization of the absolute summation of the coefficients (known as L1 regularization).
- **Ridge Regression:** OLS is altered for the minimization of the squared absolute summation of the coefficients (known as L2 regularization).

They are efficient to utilize in case of co-linearity present in the input values and OLS could overfit the training data.

2.2.4. Making Prediction with Linear Regression
By defining the data representation in a linear equation, predictions are made in a easier way by analyzing the equations for a particular collection of input. Consider a predictive weight (y) from height (x) and the linear regression model represents the problem as

$$y = B0 + B1 * x1 \quad (2)$$

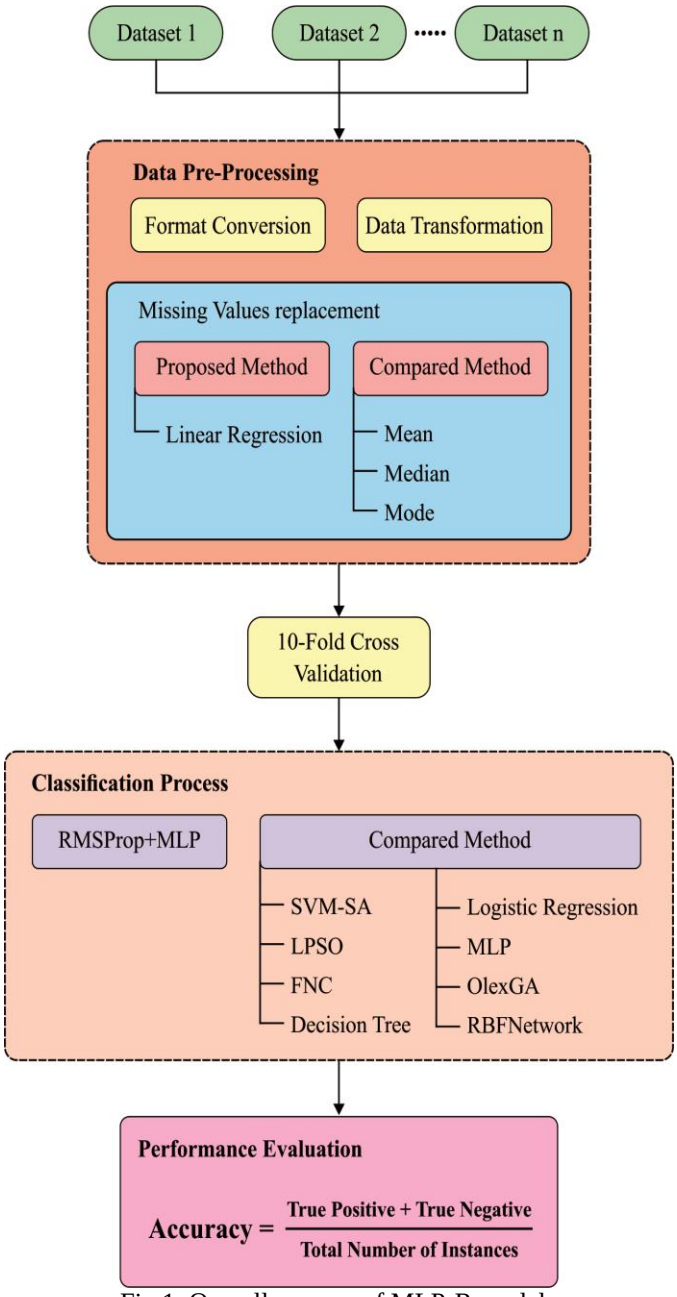

Fig.1. Overall process of MLP-R model

Where $B0$ is the bias coefficient and $B1$ is the coefficient for the height column. A learning approach is applied for determining the proper collection of coefficient values. When it is determined, it can be employed to predict the missing values present in the dataset. By looking into the values present in the dataset and the corresponding class label, the linear regression will predict the missing values by matching the data with the class labels. Using this process, linear regression model can be used for missing value replacement. The software agents will execute the process repeatedly until all the missing values gets replaced.

2.3. 10-fold CV

The missing values are replaced by the linear regression and are provided to the 10-fold CV. Normally, CV is a process employed for assessing the predictive model by partitioning the dataset into training and testing part. By the use of k-fold CV, the actual samples will be randomly partitioned to k identical size portions. From the k subsamples, an individual subsample is kept as the validation data to test the approach and the rest of the k-1 subsamples are utilized to train the model. The CV procedure gets iterated for k times where every k subsamples are utilized only one time as the testing data. This process is illustrated in Fig. 2. The k outcome from every fold will be averaged to attain a whole computation. The benefit of this model is that every sample is applied to train and validate the data. In case of 10 fold CV, the value of k will be ten and the process gets repeated up to ten times. To classify the data, 10-fold CV process is applied where the folds are chosen so that everyone has identical portions of class labels.

Fig. 2. 10-fold CV process

2.4. Classification model

MLP classifier is applied to classify the data. It has one hidden layers are the discriminated function of non-linear.

$$l = f_\theta(m) = \sigma\big(x + B.\sigma(A + A.m)\big) \quad (3)$$

where A, B is two matrices θpoint out two vectors a, x, and vector component wise sigmoid functions σ (...). Diminishing little major function contains j; this element is optimized by stochastic gradient descent (SGD) and actual classification information. In general, function f_θ contains numerous local minima. To train the NN, a local minimum is look at repetitively more trouble like XOR, spiral and parity. Extra attempts are tried to train apart from local minima incident. Owing to the f_θ symmetry, single local minima suggests high local minima in the parameter vector furthermore. At different scales, C (θ) may include minimal hierarchy usually. To NN trained various methods to achieve through these statements. At lesser scales, momentum are subjected to be moulded out the local minima and the plans for rate of learning is maybe to prepared for a minima hierarchy at different scales in a approach similar to imitation annealing. In weight space, because of the MLPs connection examination, different weight settings present equivalent results. Also, it is needed to determine the extremely MLP dimensional space weight. To MLPs, we explore to how SGD optimization technique at this time. We employ a $\{\xi_1, \ldots \xi_N\}$ as test place samples gathering apart from θ parameters evaluating and estimation to relevant vector as specified under.

$$\tau(\theta) = \big(f_\theta(\xi_1), \ldots, f_\theta(\xi_N)\big) \quad (4)$$

It is necessary $\tau(\theta) \approx \tau(\theta')$ where θ and θ' as 2 weight vectors. In the NN-SGD optimization, if the model analyzes accurately, it can be predicted as true, if it exists till some local minima. SGD optimization would follow to be relevant rate of learning until the learning rate is extremely high and search goes over the random walk mostly in the basin on the local minimum. θ parameter space have to be divided into parts to apply rate of learning, each goes to basin on the local minimum particular to rate of learning. In parameter space, it is able to search numerous first points that joined to the like basin beyond definite scale at local minimum.

RMSProp technique is depending upon the average weight of the gradients like different updating parameter through the energy of descent gradients. Assume that the optimizer rate of function tries to hold contours like under and red dot identify the region of local optimum. The points 'A' starts from the initial gradient descent and then iteration ends with point 'B' and another side of the ellipse is revealed in Fig. 3. It is noted that descent gradient ends with point 'C'.

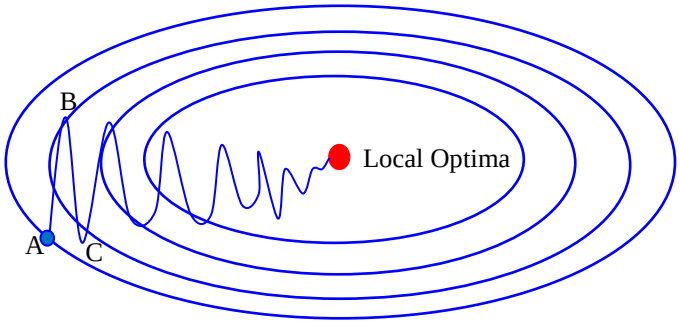

Fig. 3. RMSProp

Under all rounds of gradient descent, it is encouraged to the local optimum with down oscillations. It uses high rate of learning, and to hold the high magnitude in perpendicular oscillation. Then, the perpendicular oscillation diminishes the gradient descent and eliminates the high rate of learning. The biases have reliability for the perpendicular oscillations as 'weight' indicates the motion in the horizontal way. Previously to inform the bias, it diminishes the perpendicular oscillation and if 'weights' is informed through high values. In the backward propagation, dP and dx parameters are employing to inform P and x as specified below.

$$P = P - learningrate * dP \quad (5)$$
$$x = x - learningrate * dx \quad (6)$$

In RMSprop, dP and dx is not based on a every epoch, the exponential average weight of the square of dP and dx is concerned.

$$SdP = \beta * SdP + (1-\beta) * dP^2 \quad (7)$$
$$Sdx = \beta * Sdx + (1-\beta) * dx^2 \quad (8)$$

Where $\beta$is a new hyper parameter which obtains the value of 0 and 1. It considers the weight with the preceding average values and the square of current values for estimating the new average weight. Once the exponentially average weight is computed, it will inform the variables.

$$P = P - learningrate * \frac{dp}{sqrt(SdP)} \quad (9)$$
$$x = x - learningrate * \frac{dP}{sqrt(Sdx)} \quad (10)$$

SdP are relatively low and Sdx are relatively maximum.

III. PERFORMANCE VALIDATION

3.1. Dataset details

For examining the outcome of the projected technique for missing value replacement, three different benchmark dataset namely chronic kidney disease (CKD) [14], hepatitis and market [15] dataset are utilized. The information related to the dataset is available in Table 1. Figs. 4-6 show the distribution of the instances after missing value replacement present in the dataset.

From the table, it is noted that the applied CKD dataset has a total of 400 instances with the presence of 24 attributes. In addition, the instances are classified with respect to two classes absent and present or CKD/non-CKD. Besides, the applied hepatitis dataset has a total of 155 instances with the presence of 19 attributes. Moreover, the instances are classified with respect to two classes absent and present. Then, the applied marketing dataset has a total of 8933 instances with the presence of 9 attributes.

Table 1 Dataset Description

Description	CKD	Hepatitis	Marketing
No. of Instances	400	155	8993
No. of Features	24	19	13
No. of Class	2	2	9

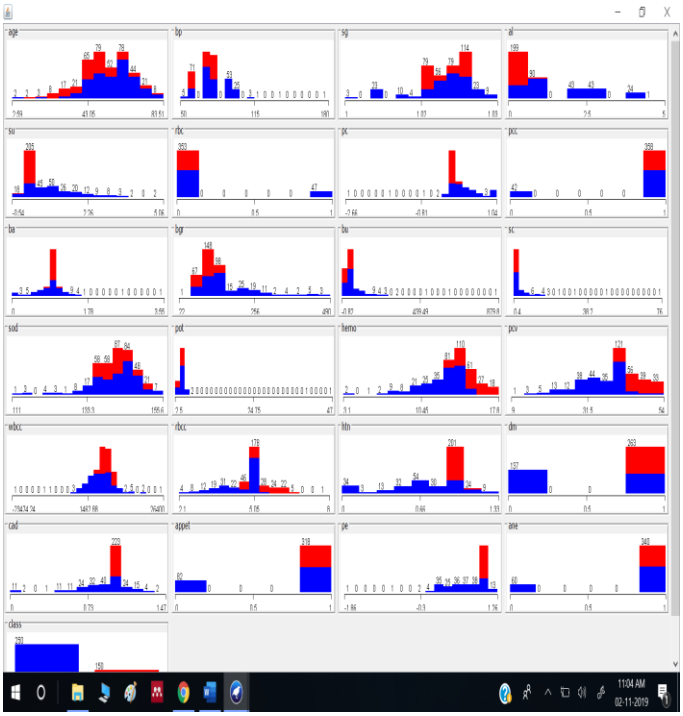

Fig. 4. Frequency Distribution of Attributes for CKD Dataset After Missing Values Replacement

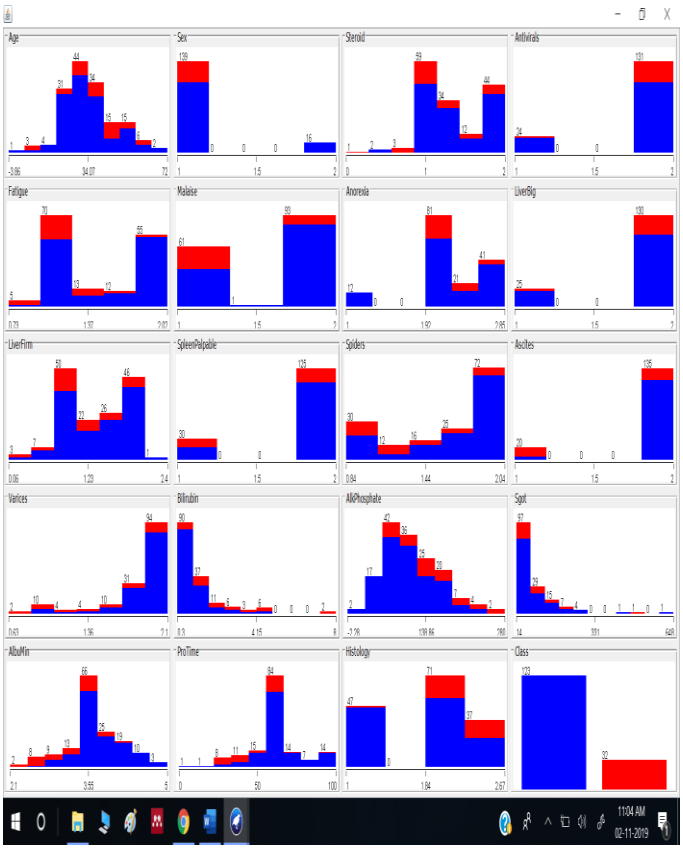

Fig. 5. Frequency Distribution of Attributes for Hepatitis Dataset After Missing Values Replacement

Fig. 6. Frequency Distribution of Attributes for Marketing Dataset After Missing Values Replacement

Table 2-4 shows the statistical analysis of the attributes present in the CKD, Hepatitis and marketing dataset respectively.

Table 2 Statistical Analysis of Features Influence on CKD Dataset

S. No	Features	Features Description	Min.	Max.	Mean	SD
1	Age	Age	2.59	83.50	52.88	13.31
2	bp	Blood Pressure	50	180	76.39	13.48
3	sg	Specific Gravity	1.00	1.03	1.02	0.004
4	al	Albumin	0	5	0.99	1.27
5	su	Sugar	-0.53	5.05	0.63	0.93
6	rbc	Red Blood Cells	0	1	0.16	0.31
7	pc	Pus Cell	-2.66	1.03	0.15	0.35
8	pcc	Pus Cell clumps	0	1	0.89	0.31
9	ba	Bacteria	0	3.55	0.93	0.31
10	bgr	Blood Glucose Random	22	490	146.2	74.96
11	bu	Blood Urea	-0.81	879.80	57.68	69.64
12	sc	Serum Creatinine	0.4	76	3.03	5.62
13	sod	Sodium	111	155.60	141.50	6.39
14	pot	Potassium	2.5	47	4.62	2.82
15	hemo	Haemoglobin	3.1	17.8	13.41	2.21
16	pcv	Packed Cell Volume	9	54	39.42	8.23
17	wbcc	White Blood Cell Count	-23474	26400	6243	4053
18	rbcc	Red Blood Cell Count	2.1	8	4.77	0.85

19	htn	Hypertension	0	1.33	0.79	0.31
20	dm	Diabetes Mellitus	0	1	0.66	0.47
21	cad	Coronary Artery Disease	0	1.46	0.911	0.22
22	appet	Appetite	0	1	0.795	0.404
23	pe	Pedal Edema	-1.859	1.256	0.732	0.38
24	ane	Anemia	0	1	0.85	0.36
25	Class	CKD, Not_CKD	1 - CKD, 2 – Not_CKD			

Table 3 Statistical Analysis of Features Influence on Hepatitis Dataset

S. No	Features	Min.	Max.	Mean	SD
1	Age	3	72	34.90	12.22
2	Sex	1	2	1.1	0.31
3	Steroid	0.02	2	1.38	0.44
4	Antivirals	1	2	1.85	0.36
5	Fatigue	0.73	2.02	1.42	0.43
6	Malaise	1	2	1.6	0.49
7	Anorexia	1	2.85	2.17	0.45
8	LiverBig	1	2	1.82	0.36
9	LiverFirm	0.06	2.39	1.39	0.48
10	Spleen Palpable	1	2	1.8	0.39
11	Spiders	0.83	2.04	1.61	0.38
12	Ascites	1	2	1.86	0.33
13	Varices	0.62	2.09	1.80	0.31
14	Bilirubin	0.3	8	1.42	1.18
15	Alk Phosphate	-2.27	280	113.913	50.70
16	Sgot	14	648	85.79	88.48
17	Alba Min	2.1	5	3.63	0.50
18	Pro Time	0	100	62.13	17.19
19	Histology	1	2.67	1.83	0.57
20	Class	1 and 2			

Table 4 Statistical Analysis of Features Influence on Marketing Dataset

S. No	Features	Min.	Max.	Mean	SD
1	Sex	1	2	1.57	0.44
2	Marital Status	1	5	3.03	1.79
3	Age (Category Mapping)	1	7	3.36	1.49
4	Education	1	6	3.84	1.24
5	Occupation	1	9	3.97	2.31
6	YearInsf	1	5	4.2	1.16
7	Dual Income	0.71	3	1.54	0.70
8	House Child Member	1	9	2.85	1.50
9	Under18	-0.39	9	0.70	1.01

10	Household Status	1	3	1.84	0.74
11	Type of Home	0.50	5	1.90	1
12	Ethnic Class	1	8	5.96	1.75
13	Language	0.94	3	1.13	0.36
14	Income (class)	1-9			

3.2. RESULTS ANALYSIS

3.2.1. Results analysis under CKD dataset

Table 5and Fig. 7 shows the accuracy analysis of different classifier methods under four different missing value replacement techniques namely mean, median, mode and linear regression on the applied CKD dataset. Using the mean based missing value replacement scheme, the existing FNC model shows least classifier outcome with the minimum accuracy of 96.87. Then, the slightly higher and near identical classifier outcome is exhibited by the RBFNetwork and OlexGA models with the accuracy values of 97.56 and 97.90 respectively. Along these lines, the MLP shows moderate classifier outcome with the accuracy value of 98.23. Besides, the DT and LR models offered competitive and near identical results with the accuracy values of 98.97 and 98.67 respectively. On comparing all the other models, the presented MLP-R model offered superior classification outcome with the highest accuracy of 99.10. Using the median based missing value replacement scheme, the existing FNC model shows least classifier outcome with the minimum accuracy of 96.45. Then, the slightly higher outcome is exhibited by the RBFNetwork model with the accuracy values of 97.34 respectively. Along these lines, the OlexGA shows moderate classifier outcome with the accuracy value of 98.12. Besides, the MLP, DT and LR models offered competitive and near identical results with the accuracy values of 98.45, 98.67 and 98.98 respectively. On comparing all the other models, the presented MLP-R model offered superior classification outcome with the highest accuracy of 99.12.

Table 5 Accuracy analysis of different models under different missing value replacement methods on CKD Dataset

Classifiers	Mean	Median	Mode	Linear regression
MLP-R	99.10	99.12	99.17	99.46
FNC	96.87	96.45	97.32	98.34
DT	98.97	98.67	98.39	98.96
LR	98.67	98.98	98.22	99.32
MLP	98.23	98.45	98.34	98.97
RBFNetwork	97.56	97.34	97.87	98.37
OlexGA	97.90	98.12	98.14	99.12

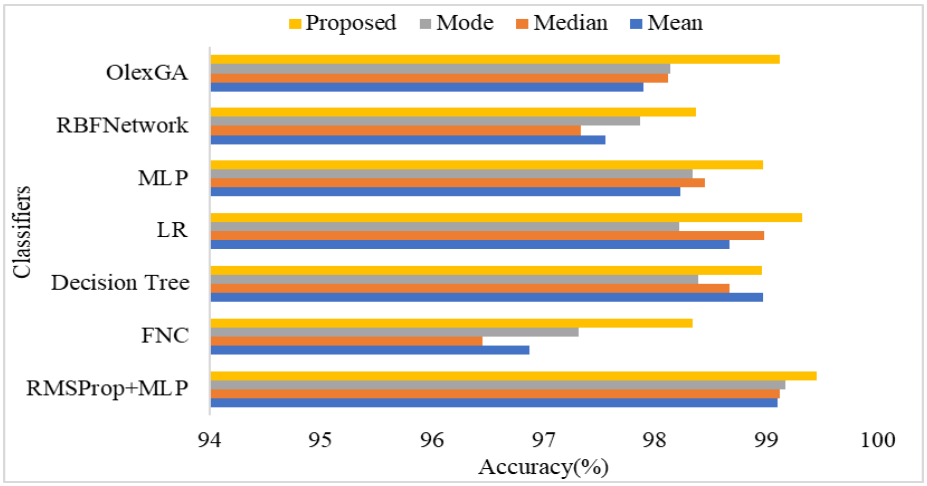

Fig. 7 Accuracy analysis of distinct classifiers on CKD dataset

Using the median based missing value replacement scheme, the existing FNC model shows least classifier outcome with the minimum accuracy of 97.32. Then, the slightly higher outcome is exhibited by the RBFNetwork model with the accuracy values of 97.87 respectively. Along these lines, the OlexGA shows moderate classifier outcome with the accuracy value of 98.14. Besides, the MLP, DT and LR models offered competitive and near identical results with the accuracy values of 98.22, 98.34 and 98.39 respectively. On comparing all the other models, the presented MLP-R model offered superior classification outcome with the highest accuracy of 99.17. Using the linear regression based missing value replacement scheme, the existing FNC model shows least classifier outcome with the minimum accuracy of 98.34. Then, the slightly higher outcome is exhibited by the RBFNetwork model with the accuracy values of 98.37 respectively. Along these lines, the DT and MLP models offered competitive and near identical results with the accuracy values of 98.96 and 98.97 respectively. Besides, the LR and OlexGA models offered with the accuracy values of 99.12 and 99.32 respectively. On comparing all the other models, the presented MLP-R model offered superior classification outcome with the highest accuracy of 99.46.

Table 6 and Fig. 8offered the comparative analysis of different classifier outcome before and after missing value replacement. The table values depicted that the least accuracy of 95.75 is attained by FNC model which is increased to 98.34 accuracy after missing value replacement. Then, the RBFNetwork offered a lower accuracy value of 96.50 which is raised upto 98.37 after missing value replacement.

Table 6 Comparisons of before and after missing values replacement for CKD Dataset

Classifiers	Before Missing Values Replacement	After Missing Values Replacement
MLP-R	98.90	99.46
FNC	95.75	98.34
Decision Tree	98.75	98.96
LR	98.50	99.32
MLP	97.75	98.97
RBFNetwork	96.50	98.37
OlexGA	97.00	99.12

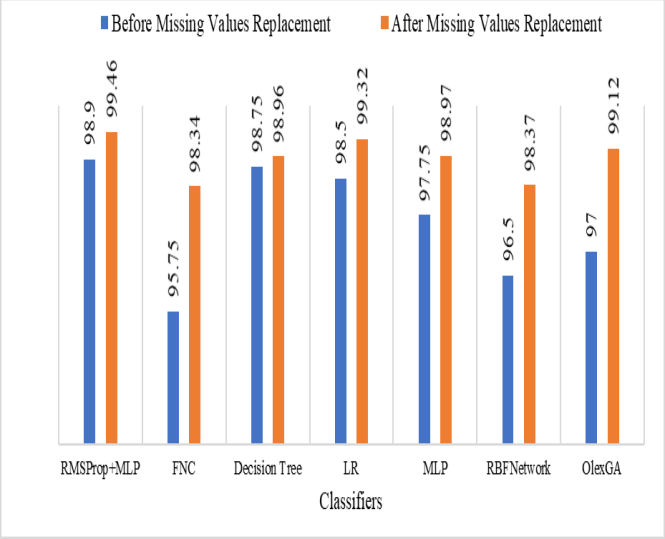

Fig. 8. Comparative analysis under before and after missing values replacement on CKD dataset

Next to that, the OlexGA model offered moderate accuracy of 97 which is improvised to 98.37 after missing value replacement. Simultaneously, the MLP model offered manageable outcome by attaining an accuracy of 97.75 which is increased to 98.97 after missing value replacement. At the same time, the DT model showed closest results to the MLP-R model with the accuracy of 95.75 and is raised to 98.34 accuracy after missing value replacement. Lastly, the MLP-R model offered maximum accuracy of 98.90 and is further increased to the highest accuracy of 99.46 after missing value replacement. It is interesting that the MLP-R model attains maximum outcome which is not even lesser than the results attained by other models after missing value replacement on the applied CKD dataset.

3.2.2. Results analysis under Hepatitis dataset

Table 7and Fig. 9 shows the accuracy analysis of different classifier methods under four different missing value replacement techniques namely mean, median, mode and Proposed on the applied Hepatitis dataset. Using the mean based missing value replacement scheme, the existing

OlexGA model shows least classifier outcome with the minimum accuracy of 27.49. Then, the slighter higher outcome is exhibited by the MLP model with the accuracy value of 84.20 respectively. Along these lines, the Decision Tree and LR offers moderate classifier outcome is nearly equivalent with the accuracy values of 86.45 and 86.49. Besides, the RBFNetwork, LSPO and SVM-SA models offered results with the accuracy values of 89.37and 92.35 and 97.89 respectively. On comparing all the other models, the presented RMSProp+MLP model offered superior classification outcome with the highest accuracy of 98.23.

Table 7 Accuracy Comparisons of proposed missing values replacement method with state of art methods for Hepatitis Dataset

Classifiers	Mean	Median	Mode	Proposed
RMSProp+MLP	98.23	98.56	98.90	99.35
LPSO [16]	92.35	93.47	94.57	94.67
SVM-SA [17]	97.89	98.74	98.98	98.45
Decision Tree	86.45	87.40	88.46	91.36
LR	86.49	89.57	92.47	92.48
MLP	84.20	87.46	90.46	90.46
RBFNetwork	89.37	91.38	93.65	87.46
OlexGA	27.49	45.60	60.46	78.47

[18] [19] Using the median based missing value replacement scheme, the existing OlexGA model shows least classifier outcome with the minimum accuracy of 45.60. Then, the slighter higher and nearly equivalent outcome is exhibited by the Decision Tree and MLP model with the accuracy values of 87.40 and 87.46 respectively. Along these lines, the RBFNetwork and LPSO moderate classifier outcome with the accuracy values of 91.38 and 93.47. Besides, the SVM-SAand LR models offered results with the accuracy values of 89.57and 98.74 respectively. On comparing all the other models, the presented RMSProp+MLP model offered superior classification outcome with the highest accuracy of 98.56. Using the mode based missing value replacement scheme, the existing OlexGA model shows least classifier outcome with the minimum accuracy of 60.46. Then, the slighter higher and nearly equivalent outcome is exhibited by the Decision Tree and MLP model with the accuracy values of 88.46 and 90.46 respectively. Along these lines, the LR and RBFNetwork moderate classifier offered results nearly equivalent outcome with the accuracy values of 92.47 and 93.65. Besides, the LPSO and SVM-SA offered results with the accuracy values of 94.57 and 98.98 respectively. On comparing all the other models, the presented RMSProp+MLP model offered superior classification outcome with the highest accuracy of 98.90.

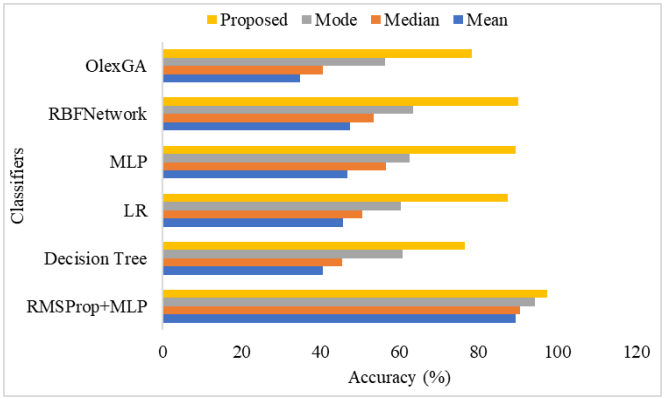

Fig. 9. Accuracy analysis of distinct classifiers on Hepatitis dataset

Table 8 and Fig. 10 offered the comparative analysis of different classifier outcome before and after missing value replacement. The table values depicted that the least accuracy of 20.64 is attained by OlexGA model which is increased to 78.47 accuracy after missing value replacement. Then, the RBFNetwork offered a lower accuracy value of 80.64 which is raised upto87.46 after missing value replacement. Next to that, the MLP model offered moderate accuracy of 81.93 which is improvised to 90.46 after missing value replacement. Simultaneously, the LR model offered manageable outcome by attaining an accuracy of 83.87 which is increased to 92.48 after missing value replacement. At the same time, the Decision Tree model showed equivalent results to the LR model with the accuracy of 83.87 and is raised to 91.36 accuracy after missing value replacement. Concurrently, the LPSO model with the accuracy of 91.11 and is raised to 94.67 accuracy after missing value replacement. Lastly, the SVM-SA model offered maximum accuracy of 97.78 and is further increased to the highest accuracy of 98.45 after missing value replacement. It is interesting that the MLP-R model attains maximum outcome which is not even lesser than the results attained by other models after missing value replacement on the applied Hepatitis dataset.

Table 8 Comparisons of before and after missing values replacement for Hepatitis Dataset

Classifiers	Before Missing Values Replacement	After Missing Values Replacement
RMSProp+MLP	98.10	99.35
LPSO	91.11	94.67
SVM-SA	97.78	98.45
Decision Tree	83.87	91.36
LR	83.87	92.48
MLP	81.93	90.46
RBFNetwork	80.64	87.46
OlexGA	20.64	78.47

Fig. 10. Comparative analysis under before and after missing values replacement on CKD dataset

3.2.3. Results analysis under Marketing dataset

Table 9and Fig. 11 shows the accuracy analysis of different classifier methods under four different missing value replacement techniques namely mean, median, mode and Proposed on the applied Marketing dataset. Using the mean based missing value replacement scheme, the existing OlexGA model shows least classifier outcome with the minimum accuracy of 34.90. Then, the slighter higher outcome is exhibited by the Decision Tree model with the accuracy value of 40.56 respectively. Along these lines, the LR offers moderate classifier outcome with the accuracy values of 45.69. Besides, the RBFNetwork and MLP models offered nearly equivalent results with the accuracy values of 46.79 and 47.36 respectively. On comparing all the other models, the presented RMSProp+MLP model offered superior classification outcome with the highest accuracy of 89.35.

Table 9 Accuracy Comparisons of proposed missing values replacement method with state of art methods for Marketing Dataset

Classifiers	Mean	Median	Mode	Proposed
RMSProp+MLP	89.35	90.45	94.38	97.36
Decision Tree	40.56	45.47	60.87	76.63
LR	45.69	50.47	60.39	87.43
MLP	46.79	56.49	62.64	89.34
RBFNetwork	47.36	53.48	63.49	90.13
OlexGA	34.90	40.57	56.39	78.39

Using the median based missing value replacement scheme, the existing OlexGA model shows least classifier outcome with the minimum accuracy of 40.57. Then, the slighter higher outcome is exhibited by the Decision Tree model with the accuracy value of 45.47 respectively. Along these lines, the LR offers moderate classifier outcome with the accuracy value of 50.47. Besides, the RBFNetwork and MLP models offered nearly equivalent results with the accuracy values of 53.48 and 56.49 respectively. On comparing all the other models, the presented RMSProp+MLP model offered superior classification outcome with the highest accuracy of 90.45.Using the mode based missing value replacement scheme, the existing OlexGA model shows least classifier outcome with the minimum accuracy of56.39. Then, the slighter higher and nearly equivalent outcome is exhibited by the LR and Decision Tree model with the accuracy values of 60.39 and 60.87 respectively. Along these lines, the MLP offers moderate classifier outcome with the accuracy value of 62.64. Besides, the RBFNetwork offered result with the accuracy values of 63.49 respectively. On comparing all the other models, the presented RMSProp+MLP model offered superior classification outcome with the highest accuracy of 94.38.

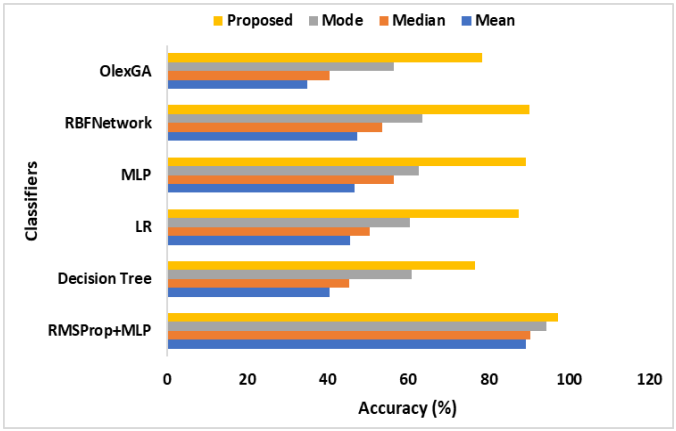

Fig. 11. Accuracy analysis of distinct classifiers on Marketing dataset

Using the Proposed Method based missing value replacement scheme, the existing Decision Tree model shows least classifier outcome with the minimum accuracy of 76.63. Then, the slighter higher outcome is exhibited by the OlexGA model with the accuracy value of 78.39 respectively. Along these lines, the LR offers moderate classifier outcome with the accuracy value of 87.43. Besides, the MLP and RBFNetwork models offered nearly equivalent results with the accuracy values of 89.34 and 90.13 respectively. On comparing all the other models, the presented RMSProp+MLP model offered superior classification outcome with the highest accuracy of 97.36.

Table 10and Fig. 12 offered the comparative analysis of different classifier outcome before and after missing value replacement. The table values depicted that the least accuracy of 17.20 is attained by OlexGA model which is increased to 78.39 accuracy after missing value replacement. Then, the Decision Tree offered a lower accuracy value of 31.86 which is raised upto76.63 after missing value replacement. Next to that, the LR model offered moderate accuracy of 32.59 which is improvised to 87.43 after missing

value replacement. Simultaneously, the MLP model offered manageable outcome by attaining a maximum accuracy of 33.45 which is increased to 89.34 after missing value replacement. Lastly, the RBFNetwork model offered accuracy of 33.42 and is further increased to the highest accuracy of 90.13 after missing value replacement. It is interesting that the MLP-R model attains maximum outcome which is not even lesser than the results attained by other models after missing value replacement on the applied Marketing dataset.

Table 10 Comparisons of before and after missing values replacement for Marketing Dataset

Classifiers	Before Missing Values Replacement	After Missing Values Replacement
Proposed	88.67	97.36
Decision Tree	31.86	76.63
LR	32.59	87.43
MLP	33.45	89.34
RBFNetwork	33.32	90.13
OlexGA	17.20	78.39

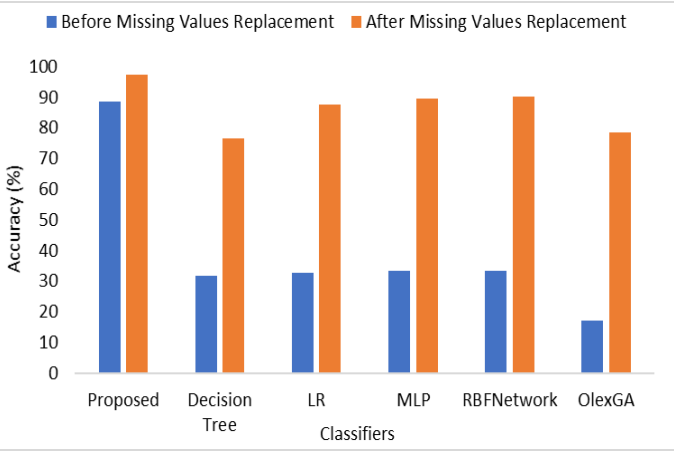

Fig. 12. Comparative analysis under before and after missing values replacement on CKD dataset

3.2.4. Accuracy rate analysis

Fig. 13 (a) demonstrates the investigation of the experimental outcome of the projected model the accuracy analysis of the presented model on the applied CKD dataset under varying number of epochs. The figure stated that the presented model reaches a maximum of 88% accuracy under minimum number of epochs. In addition, the value of accuracy gets increased by increasing the number of epochs. On reaching the epoch count by 1000 on the applied CKD dataset, the presented model attains a maximum of around 99.46% accuracy.

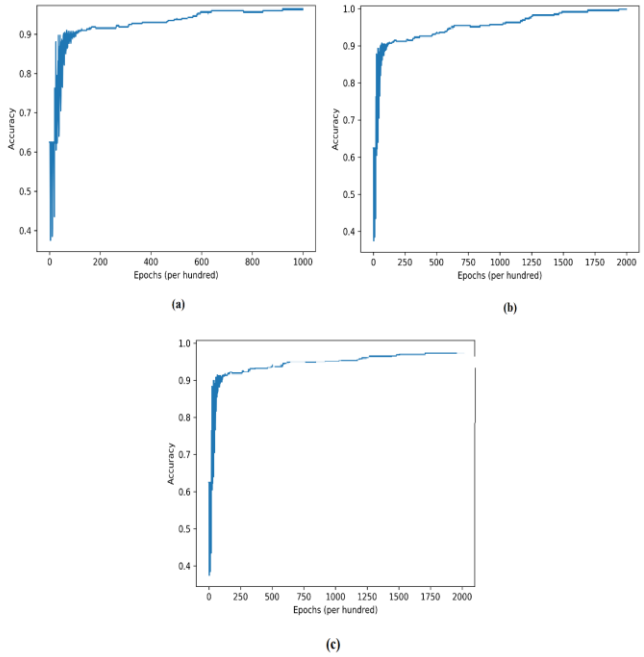

Fig. 13. Accuracy Graph of a) CKD b) Hepatitis c) Marketing Dataset

Similarly, Fig. 13 (b) shows the accuracy analysis of the presented model on the applied Hepatitis dataset under varying number of epochs. The figure stated that the presented model reaches a maximum of 89% accuracy under minimum number of epochs. In addition, the value of accuracy gets increased by increasing the number of epochs. On reaching the epoch count by 1000 on the applied Hepatitis dataset, the presented model attains a maximum of around 99.35% accuracy. In the same way, Fig. 13 (c) shows the accuracy analysis of the presented model on the applied Marketing dataset under varying number of epochs. The figure stated that the presented model reaches a maximum of 90% accuracy under minimum number of epochs. In addition, the value of accuracy gets increased by increasing the number of epochs. On reaching the epoch count by 2000 on the applied Hepatitis dataset, the presented model attains a maximum of around 97.36% accuracy.

3.2.5. Loss rate analysis

Fig. 14 (a) shows the loss rate analysis of the presented model on the applied CKD dataset under varying number of epochs. The figure stated that the presented model reaches a maximum of 3% loss rate under minimum number of epochs. In addition, the value of loss rate gets decreased by increasing the number of epochs. On reaching the epoch count by 1000 on the applied CKD dataset, the presented model attains a minimum of around 0.05% loss rate. Similarly, Fig. 14 (b) shows the loss rate analysis of the presented model on the applied Hepatitis dataset under

varying number of epochs. The figure stated that the presented model reaches a minimum of 3% loss rate under minimum number of epochs. In addition, the value of loss rate gets increased by increasing the number of epochs. On reaching the epoch count by 1000 on the applied Hepatitis dataset, the presented model attains a minimum of around 0.1% loss rate.

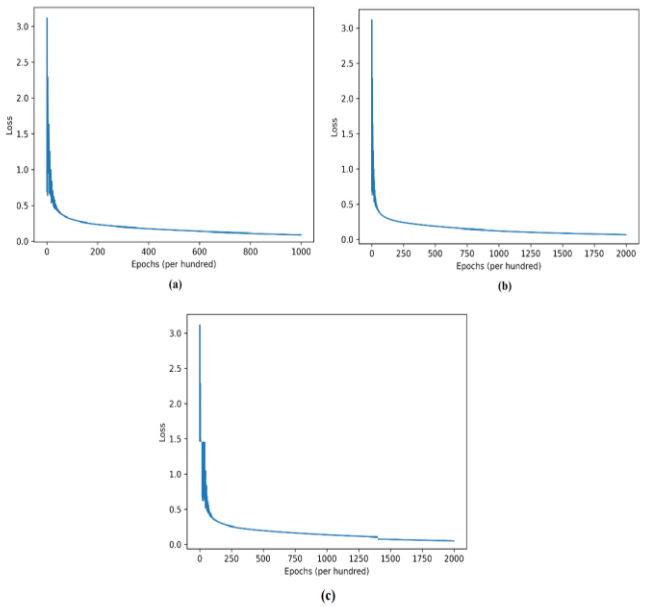

Fig. 14. Loss Graph of a) CKD b) Hepatitis c) Marketing Dataset

In the same way, Fig. 14(c) shows the loss rate analysis of the presented model on the applied Marketing dataset under varying number of epochs. The figure stated that the presented model reaches a minimum of 3% loss rate under minimum number of epochs. In addition, the value of loss rate gets increased by increasing the number of epochs. On reaching the epoch count by 1000 on the applied Hepatitis dataset, the presented model attains a minimum of around 0.1% loss rate. The above-mentioned tables and figures clearly verified the extraordinary classification performance of the projected method. The experimental results showed significant improvement by the use of missing value replacement process.

IV. CONCLUSION

This paper has presented an efficient missing value replacement and classification model to handle missing data. For replacing the missing values, linear regression model is applied to improve the classification process. Then, MLP classifier is applied to classify the data which further tuned by the use of RMSProp model. For examining the outcome of the projected technique for missing value replacement, three different benchmark dataset is applied. The above-mentioned tables and figures clearly verified the extraordinary classification performance of the projected method. The experimental results showed significant improvement by the use of missing value replacement process.

REFERENCES

[1] Horton, N. J., Lipsitz, S. R., Orton, N. J. H., &Ipsitz, S. R. L. (2001). Multiple Imputation in Practice : Comparison of Software Packages for Regression Models With Missing Variables. The American Statistician, 55(3), 244–254.
[2] Batista, G. E. A. P. A., &Monard, M. C. (2002). A Study of K -Nearest Neighbour as an Imputation Method. In Soft Computing Systems: Design, Management and Applications (pp. 251–260).
[3] Malarvizhi, M. R., &Thanamani, A. S. (2012). K-Nearest Neighbor in Missing Data Imputation. International Journal of Engineering Research and Development, 5(1), 5–7.
[4] Soltanveis, F. (2016). Using Parametric Regression and KNN Algorithm With Missing Handling For Software Effort Prediction. In Artificial Intelligence and Robotics (IRANOPEN) (pp. 77–84).
[5] Chen, Q., Cho, M., Kim, M., & Wang, C. (2016). Using link-preserving imputation for logistic partially linear models with missing covariates. Computational Statistics and Data Analysis, 101, 174–185.
[6] Amiri, M., & Jensen, R. (2016). Missing data imputation using fuzzy-rough methods. Neurocomputing, 205, 152–164.
[7] Lee, K. J., & Carlin, J. B. (2010). Multiple Imputation for Missing Data : Fully Conditional Specification Versus Multivariate Normal Imputation. American Journal of Epidemiology, 171(5), 624–632.
[8] Sahri, Z., Yusof, R., &Watada, J. (2014). FINNIM : Iterative Imputation of Missing Values in Dissolved Gas Analysis Dataset. IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, 10(4), 2093–2102.
[9] Belanche, L. A., Kobayashi, V., &Aluja, T. (2014). Handling missing values in kernel methods with application to microbiology data. Neurocomputing, 141, 110–116.
[10] Burgette, L. F., & Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. American Journal of Epidemiology 172(9), 1070-1076.
[11] Kwon, T. Y., & Park, Y. (2015). A new multiple imputation method for bounded missing values. Statistics and Probability Letters, 107, 204–209.
[12] Deb, R., &Liew, A. W. (2016). Missing value imputation for the analysis of incomplete traffic accident data. Information Sciences, 339, 274–289.
[13] Sovilj, D., Eirola, E., Miche, Y., Björk, K., Nian, R., Akusok, A., &Lendasse, A. (2016). Extreme learning machine for missing data using multiple imputations. Neurocomputing, 174, 220–231.
[14] https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease
[15] https://sci2s.ugr.es/keel/missing.php
[16] Tarkhaneh, O. and Shen, H., 2019. Training of feedforward neural networks for data classification using hybrid particle swarm optimization, Mantegna Lévy flight and neighborhood search. Heliyon, 5(4), p.e01275.
[17] Sartakhti, J.S., Zangooei, M.H. and Mozafari, K., 2012. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). Computer methods and programs in biomedicine, 108(2), pp. 570-579.
[18] K.D. Patel, "*Review on Techniques in Natural Language Processing*", International Journal of Scientific Research in Computer Science and Engineering, Vol.7, Issue.5, pp.1-4, 2019.
[19] Amin Rezaeipanah, Zahra Abshirini, Milad Boshkani Zade, "*Solving University Course Timetabling Problem Using Parallel Genetic Algorithm*", International Journal of Scientific Research in Computer Science and Engineering, Vol.7, Issue.5, pp.5-13, 2019.

