American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN: 2320-0936
Volume-9, Issue-1, pp-189-196
www.ajer.org
Research Paper
Open Access

The Use of Supervised Learning Neural Network to Approximate Missing Data in Database

A.O Anibasa[1], A.Muhammed[2], A.R Ladodo[3]
Department of Electrical/Electronic Engineering Faculty of Engineering, Nigerian Defence Academy, Kaduna
Department of Electrical/Electronic Engineering Kogi State Polytechnic, Itakpe
Department of Electrical/Electronic Engineering Faculty of Engineering, Nigerian Defence Academy, Kaduna
Corresponding Author: A.O Anibasa

ABSTRACT: *Missing data creates various problems in analyzing and processing of data in databases. Due to this reason missing data has been an area of research in various disciplines for a quite long time. This report introduces a new method aimed at approximating mission data in a database using supervised learning neural networks. the research focus lies on the investigation of using the proposed method in approximating missing data with great accuracy as the number of missing cases within a single record increases. The research also investigates the impact of using different neural network architecture in training the Neural Network the approximation found to the missing values. It is observed that approximation of missing data obtained using the proposed model to be highly accurate with 95persent correlation coefficient between the actual missing values using the proposed model. It is found that result obtained using RBF are better than MLP. Results found using the combination of both MLP and RBF and found to be better than those obtained using either MPL or RBF. It is also observed that there is no significant reduction in accuracy of result as the number of missing cases in a single record increases. Approximation found for missing data are also found to depend on the particular neural network architecture employed in training the data set.*

Date of Submission: 15-01-2020 Date of acceptance: 31-01-2020

I. INTRODUCTION

A neural network is an information processing paradigm that is inspired by the way biological nervous systems, like the brain process information [20]; [11]. It is a machine that is designed to model the way in which the brain performs a particular task or function of interest [11].

A neural network consists of four main parts [11]. These are the processing units U_j, where each U_j has a certain activation level $a_j(t)$ at any point in time, weighted interconnection between the various processing units which determine how the activation of one unit leads to input for another unit, an activation rule which acts on the set of input signals at a unit to produce a new output signal , and a learning rule that specified how to adjust the weights for a given input/output pair. One of the main important features of neural networks is its ability to adapt to new environment. Hence learning algorithms are critical to neural networks. Due to their ability to derive meaning from complicated data, neural networks are used to extract patterns and detect trends that are two complexes to be noticed by many other computer techniques [10]. A trained neural network can be considered as an expert in the category of information it has been given to analyze [20]. This expert can be then be used to provide predictions given new situations. Because of their ability to adapt to a non-linear data neural networks are also being used to model various non-linear applications [11]; [10]. Neural networks have many advantages as machine learning technique and some of them are outlined as follows.

- Adaptive learning: an ability to learn how to do tasks based on the data given for training or initial experience [11].
- Non linearity: an artificial neuron can be linear or non linear. A neural network made up of an interconnection of nonlinear neurons is nonlinear. [11] Points that, nonlinearity is a highly important

property, especially if the underlying physical mechanism responded for generation of the input signal is inherently nonlinear.

- Adaptively: neural network have a built in capability to adapt their synaptic weights to changes in the surrounding environment. In particular, a neural network trained to operate in a specific environment can easily be retained to deal with minor changes in the operating environment conditions [11].
- Fault tolerance: partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retrained even with major network damage [11]; [6]. For each input pattern, the value of the desired output is specified.

Supervised learning is based in the system trying to predict outcomes for known examples and is a commonly used training method [9]. It compares its predictions to the target answer and adjusts the weights accordingly. The data starts as an input to the input layer neurons. The neurons pass the inputs along to the next nodes.

The process of feeding errors backwards through the network is called back propagation [15]. Both the multi-layer perceptron and the radial basis function are supervised learning techniques. The multi-layer perceptron uses the back-propagation while the radial basis function uses a feed-forward approach which trains on a single pass [Nabney 2001]. Learning without a teacher is called unsupervised learning. Unsupervised learning does not involve the use of target data. Instead of learning an input-output mapping, the aim is to discover or model the probability of input data [9]. Neural network which use unsupervised learning are most effective for clustering or detection of similarities on unlabeled patterns of a given training [10]; [13].
Single layer network implement the well known statistical techniques of regression and Generalized Linear Models (GLMs) [15].

II. METHODOLOGY

2.1 INTRODUCTION

The Neural Network was trained to recall to itself or predict its input vector (auto – associative neural network).

The result of a neural network depends on the particular network architecture and parameters (number of hidden layers, hidden units, activation and optimization functions) used in training the neural network, to select the best architectures that gives best results, different network architectures and parameters were selected and the network trained using the parameters. The architecture and parameters that gave best result were selected to be used in this research. Below are MLP and RBF architectures and parameters used in this research?

2.1.1 MLP architecture used in the experiment.

Each neural in one layer is directly connected to the neurons of the subsequent layer. A NETLAB tool box that runs in MATLAB discussed in [15] was used to implement the MLP neural network. A two layered MLP architecture was used because of it resulted in better result and due to the universal approximation theorem, which states that a two layered architecture is adequate for MLP [15]

Figure 1 depicts the architecture of the MLP used in the research. The MLP network contains 14 inputs, 2 hidden layers with 10 neurons and 14 output unit. The optimization technique used for training this architecture was the scaled conjugate gradient (SCG) method.

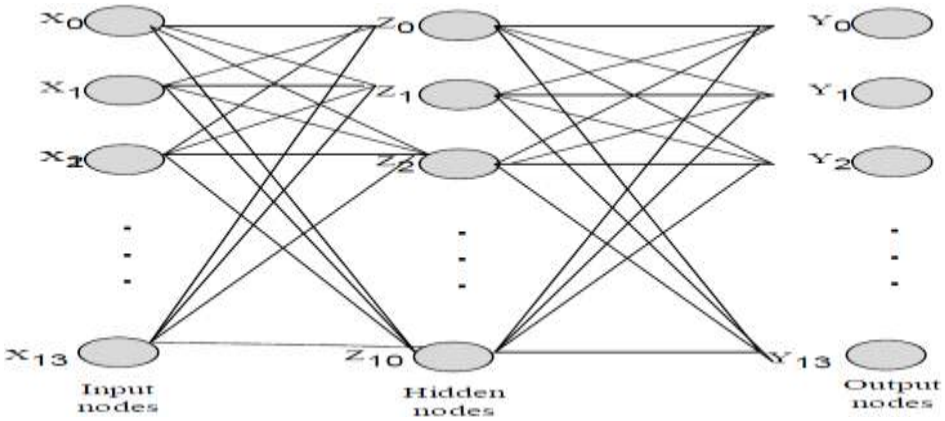

Figure 1: MLP Architecture used in the Research

SCG method was used because it gave better result and has been found to solve the optimization problem encountered when training an MLP network more efficiently than the gradient descent and conjugate gradient method [6]

2.1.2 RBF Architecture used in the research

A fully connected two layered RBF architecture was used in the experiment. Each neuron in one layer is directly connected to the neurons of the subsequent layer. Like the MLP a NETLAB toolbox that runs in MAT LAB discussed in [15] was used to implement the RBF architecture. Like the MLP, the network has 14 input, 10 neurons and 14 output units. The thin plate splines function was used as hidden unit activation function and the SCG was used as network optimization method. The RBF network used in this research is depicted in figure 2.

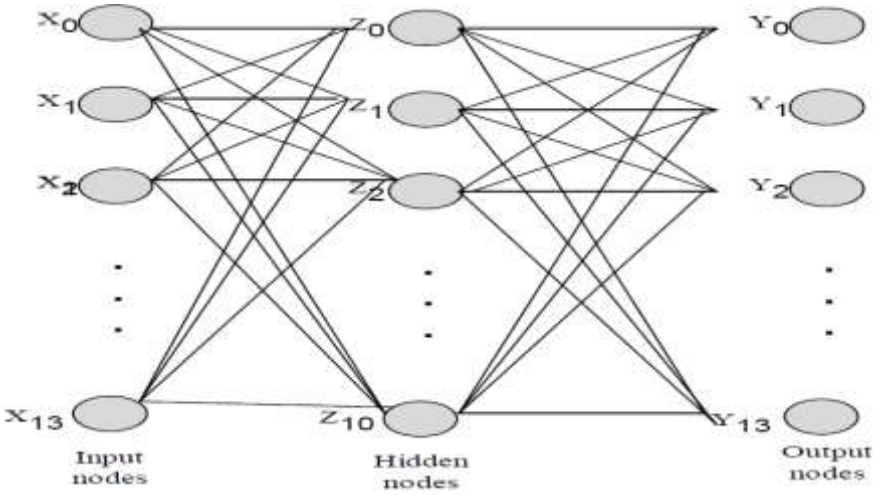

Figure 2: RBF Architecture used in the Research

2.2 Experiment Data

The input data (real data base) used in the experiment was obtained from the Nigeria breweries (NB). The database used has 14 variables. To examine the distribution of the database and its represent ability in the investigation, key statistical summary of the database are given in table 1.Where N represents the number of observation in the variable.

Looking at the distribution of statistical summaries of the database, it can be observed that the database has ideal measures of central tendency and variation.

Table 1: Statistical summary of input data

Variables	N	Mean	Min	Max	Standard Deviation
x_1	198	4.248	3.900	4.540	0.117
x_2	198	6.925	5.600	8.800	0.549
x_3	198	9.442	0.000	39.800	7.355
x_4	198	21.230	12.900	25.000	1.720
x_5	198	63.652	34.000	98.000	13.968
x_6	198	0.042	0.010	0.180	0.022
x_7	198	161.419	72.000	286.000	82.501
x_8	198	2.090	1.000	3.200	0.239
x_9	198	0.342	0.100	0.800	0.248
x_{10}	198	0.160	0.000	0.300	0.057
x_{11}	198	35.384	13.000	64.000	8.005
x_{12}	198	5.824	1.700	13,600	2.130
x_{13}	198	20.325	8.000	38.000	4.714
x_{14}	198	1.849	1,000	4.300	0.620

The estimates used to measure the modeling quality are:

Correlation coefficient (r): the correlation coefficient measure the liner relationship between two variables. For a given data $x_1, x_2, \ldots\ldots\ldots\ldots x_n$ and corresponding approximated values $\hat{x}_1, \hat{x}_2 \ldots\ldots\ldots\ldots\ldots \hat{x}_n$ the correlation coefficient is computed as

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x_i}\right)\left(\hat{x}_i - \hat{x}_i\right)}{\left[\sum_{i=1}^{n}\left(x_i - \overline{x_i}\right)^2 \sum_{i=1}^{n}\left(\hat{x}_i - \hat{x}_i\right)^2\right]^{1/2}} \quad (1)$$

Stranded Error (Se): The standard error represents average deviation between actual and predicted observations [Draper and Smith 1998] for a given data $x_1, x_2, \ldots\ldots\ldots\ldots x_n$ and corresponding approximated values $\hat{x}_1, \hat{x}_2 \ldots\ldots\ldots\ldots\ldots \hat{x}_n$ the standard error (Se) is computer as

$$Se = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2}{n}} \quad (2)$$

The higher the value of standard error, the less the accuracy and vice versa.

III. DESIGN AND SIMULATION RESULTS

In this section, all design scenarios for different parameters for simulation and the parameters when using MLP and RBF process are shown in figures and tables below. The output results for Scenario are obtained using Mat lab.

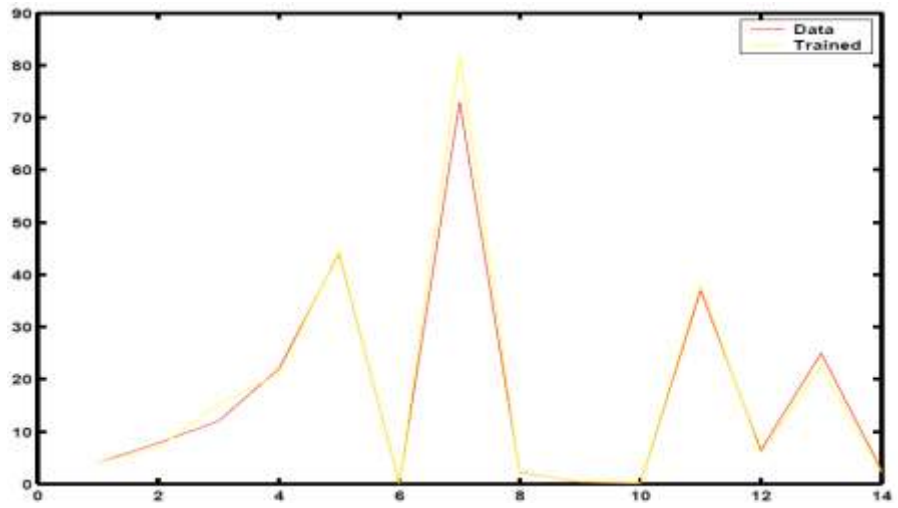

Figure 3: Data versus trained using MLP

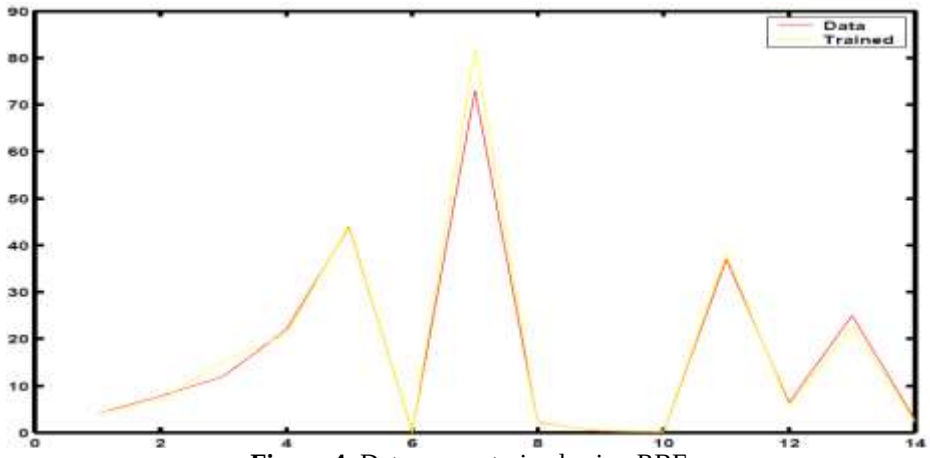

Figure 4: Data versus trained using RBF

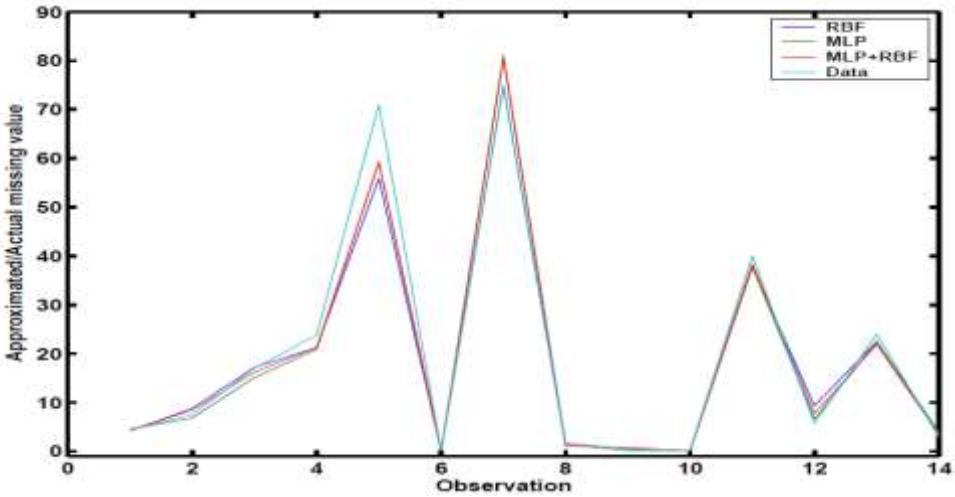

Figure 5: one missing case: using MLP, RBF, and MLP + RBF

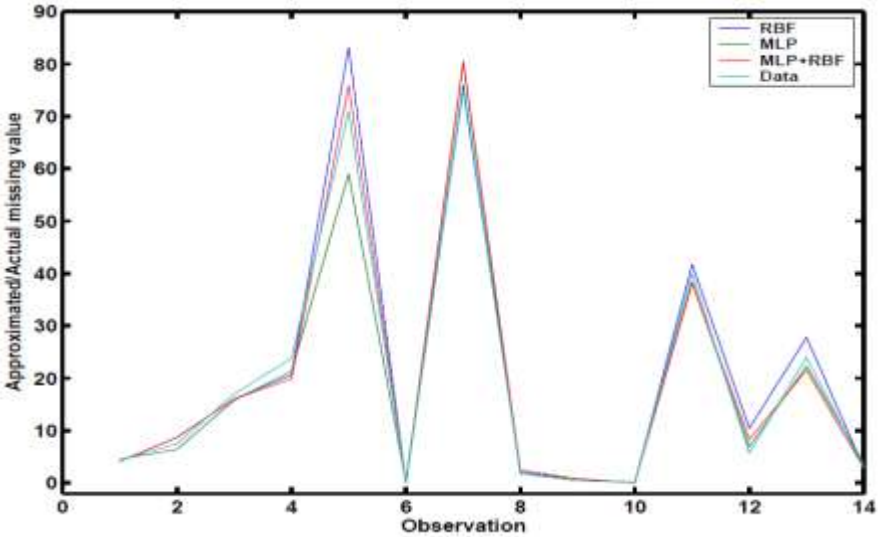

Figure 6: Two missing case: actual versus approximated values using MLP, RBF, and MLP + RBF

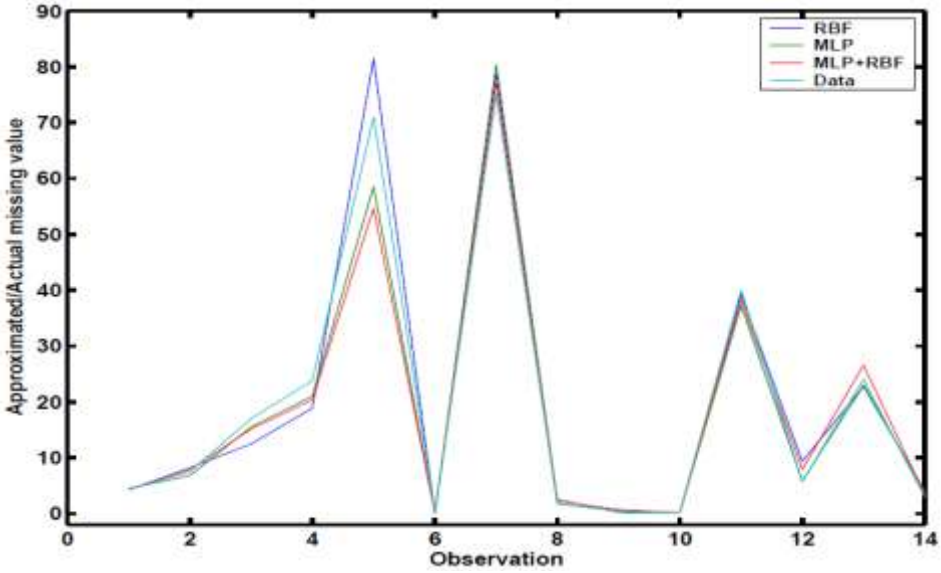

Figure 7: Three missing case: using MLP, RBF and MLP + RBF

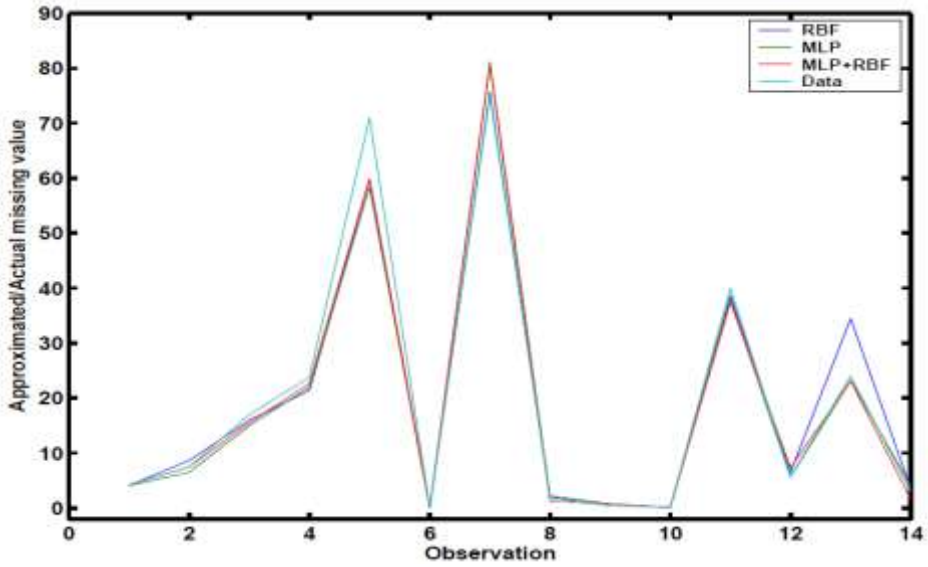

Figure 8: Four missing case: using MLP, RBL, and MLP + RBF

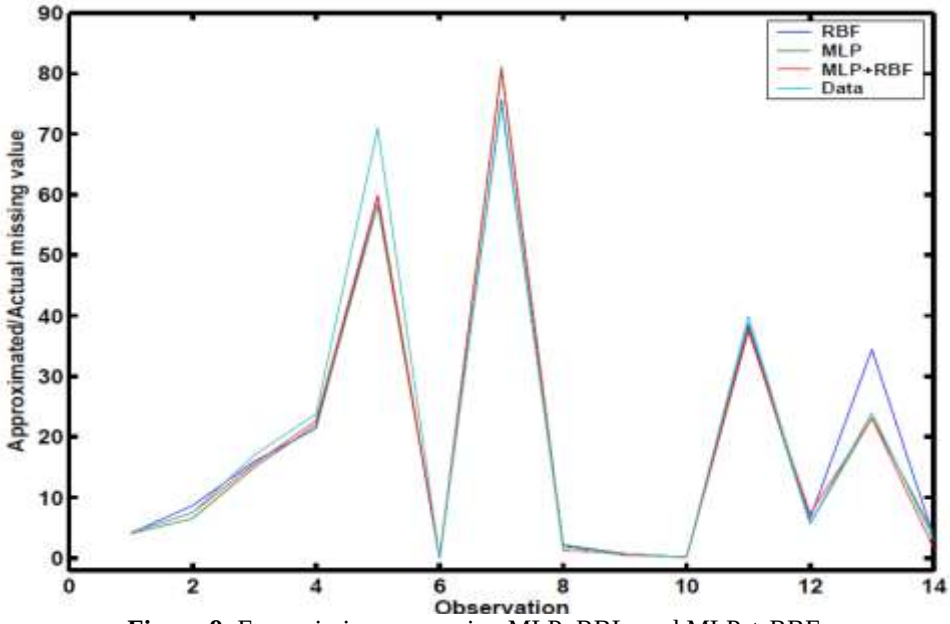

Figure 9: Four missing case: using MLP, RBL, and MLP + RBF

Table 2: Actual and approximated values using MLP

Data `	Number of missing cases in a record				
	1	2	3	4	5
4.28	4.54	4.54	4.53	4.47	4.407
7.5	6.86	6.29	6.41	6.80	6.52
17	15.50	15.10	15.8	15.5	15.0
23.8	21.20	20.90	21.3	21.0	22.0
71	59.20	59.20	59.0	58.5	58.4
0.1	0.18	0.17	0.17	0.05	0.02
75	79.90	81.1	80.3	80.3	81.2
1.8	2.48	2.41	1.81	2.04	2.21
0.4	0.10	0.104	0.72	0.22	0.72
0.2	0.58	0.06	0.02	0.11	0.159
4.0	38.10	37.8	38.4	37.2	38.0
5.7	6.64	6.66	6.96	5.82	5.67
24	22.10	22.4	22.3	23.0	23.0
2.9	3.23	3.86	3.74	3.83	3.97

Table 3: Actual and approximated values using RBF

Data `	Number of missing cases in a record				
	1	2	3	4	5
4.28	4.21	4.20	4.12	4.25	4.13
7.5	7.89	8.29	8.71	8.21	8.65
17	16.96	17.16	16.04	12.48	15.95
23.8	20.74	21.25	20.60	18.88	21.43
71	68.11	55.83	83.21	81.46	59.78
0.1	0.06	0.04	0.05	0.05	0.08
75	83.92	74.84	75.96	78.79	25.70
1.8	1.00	1.14	2.15	1.73	2.01
0.4	0.70	0.71	0.76	0.55	0.71
0.2	0.10	0.10	0.09	0.16	0.11
4.0	56.45	57.73	61.73	62.16	62.65
5.7	9.79	9.30	10.43	9.33	6.54
24	22.40	22.52	27.81	36.79	34.45
2.9	3.31	3.48	2.87	3.98	3.50

Table 4: Actual and approximated values using MLP + RBF

Data `	Number of missing cases in a record				
	1	2	3	4	5
4.28	4.50	4.20	4.012	4.24	4.20
7.5	7.40	8.42	8.71	7.85	7.47
17	14.46	16.17	16.07	15.13	15.53
23.8	21.63	21.24	19.88	20.38	22.67
71	65.38	59.16	75.92	54.66	59.89
0.1	0.01	0.04	0.06	0.07	0.03
75	92.09	80.37	80.56	77.16	80.44
1.8	2.26	1.20	2.50	2.21	1.30
0.4	0.20	0.71	0.75	0.72	0.67
0.2	0.30	0.11	0.11	0.21	0.10
40	37.66	38.36	38.13	38.23	37.43
5.7	7.40	7.28	8.40	7.87	7.34
24	27.91	21.87	21.51	26.55	23.09
2.9	2.56	3.35	2.81	3.33	1.61

IV. DISCUSSION OF THE RESULT

Figures 3 and 4 illustrate the data and trained network of one record for both MLP and RBF respectively. It can be observed that trained values using the network are similar to the actual data for both MLP and RBF networks. A sample of the actual missing data and its approximated values using the model for the 14 variables used in the model are presented in tables 2, 3, and 4 and figures 5, 6, 7, 8, and 9 respectively. Figures 5, 6, 7, 8 and 9 illustrate the actual missing values and the corresponding approximated values using the model for 1,2,3,4 and 5 missing cases respectively. Figures, shows that the approximated values in all the missing cases are more accurate. It can also be observed that there is no significant difference in approximated values for the 1, 2, 3, 4, and5 missing cases. It can also be observed that approximations found using the combined is better than MLP and RBF. Approximation found using RBF is relatively higher than MLP. For further illustration the values depicted in Figure 5, 6, 7 and 8 are also presented in Table 2, 3 and 4.

V. CONCLUSION AND RECOMMENDATION

Conclusion

It is observed that result found using the combination of both RBF and MLP trained networks are superior than the one found using either RBF or MLP. Result found using RBF was found to be far better than MLP.

Recommendation

As a future work, the results presented in this report are quite promising and reliable, a number of avenues for future work exist that may improve the effectiveness of this approach or which can utilize the approach proposed in this research to tackle other problems/applications.

REFERENCES

[1]. Alfonseca, M. (1991), genetic algorithms. In proceedings of the international conference on April, pages 1-6. AMC press.
[2]. Allison, P.D. (2000), multiple imputations for missing data: A cautionary tale. In sociological methods and research, volume 28, pages 301-309.
[3]. Allusion, p. (2002), missing data, thousand oaks', CA:sage.
[4]. Back, J.,Hoffmeiner, F, and Schwefel, H. (1992), applications of evolutionary algorithms.
[5]. Banzhat, W., Nordan, P., keller, R., and Francone, F., (1998), genetic programming and introduction: on the automatic evolution of computer programs and its applications. Mrgan Kaufmann publisher, California, fifth edition.
[6]. Bishop, C.M. (1995), neural networks for pattern recognition. Oxford University Press, Oxford.
[7]. Draper, N., Smith, H. (1998), applied regression analysis New York, third edition.
[8]. Forrest, S. (1996), genetic algorithms. ACM compute. Sur., 28(1): 77-80
[9]. Freeman, J., and Scapula, D. (1991), neural networks algorithms, applications and programming techniques
[10]. Hassoun, M.H. (1995), fundamentals of artificial neural network. MIT press, Cambridge, Massachusetts.
[11]. Haykin, S. (1999), neural networks Prentices hall, New Jersey, second edition.
[12]. Christopher, R., H., Jeffery, A., J., and Michael, G K. (1995), A genetic algorithm for function optimization: a mat lab implementation.
[13]. Kolarik, T., and Rudorfer, G (1994), Time series forecasting using neural networks. In processing of the international conference on APL: the language and its application, page 86-94 ACM press.
[14]. Little, R., and rubin, D. (1987), statistical analysis with missing data john Wiley and sons, New York, first edition.
[15]. Nabney, T. (2001), Netlab: algorithms for pattern recognition. Springer-verlag, United Kingdom.
[16]. Roth, P. (1994), missing data: a conceptual overview for applied psychologist. In personal psychology, volume 47, pages 537-560
[17]. Scheffer, J. (2000), Dealing with missing data, I.I.M.S Quad A, Massey University, Auk land. http://www. Massey.ac. nz/ wwiims/research /letters.
[18]. Yuan, Y. (2000), multiple imputations for missing data: concept and new development. In SUGI paper 266-25.
[19]. Abdella, M.,and Marwala, T. (2005), Treatment of Missing data using neural networks and genetic algorithms. International joint conference on neural networks, montreal, Canada.
[20]. Yoon, Y., and Peterson, L. (1990), artificial neural networks: an emerging new technique. In proceedings of the ACM SIGBDP Conference on trends and directions in expert systems, pages 417- 422.

A.O Anibasa, et.al. "The Use of Supervised Learning Neural Network to Approximate Missing Data in Database". *American Journal of Engineering Research (AJER)*, vol. 9(01), 2020, pp 189-196.