Fuzzy based Techniques for Handling Missing Values

Malak El-Bakry[1], Ayman El-Kilany[3], Sherif Mazen[4]
Information Systems Department
Faculty of Computers and Information
Cairo University, Cairo, Egypt

Farid Ali[2]
Information Technology Department
Faculty of Computers and Information
Beni-suef University, Egypt, Cairo, Egypt

***Abstract*—Usually, time series data suffers from high percentage of missing values which is related to its nature and its collection process. This paper proposes a data imputation technique for imputing the missing values in time series data. The Fuzzy Gaussian membership function and the Fuzzy Triangular membership function are proposed in a data imputation algorithm in order to identify the best imputation for the missing values where the membership functions were used to calculate weights for the data values of the nearest neighbor's before using them during imputation process. The evaluation results show that the proposed technique outperforms traditional data imputation techniques where the triangular fuzzy membership function has shown higher accuracy than the gaussian membership function during evaluation.**

Keywords*—Time series data; fuzzy logic; membership functions; machine learning; missing values

I. Introduction

In computer science field, the data quality problem began to rise in the 1990s with arise of the data warehouse systems where the failure of a database project was returned to its poor data quality. [1] There is a lot of definitions for the word "data quality" but as mentioned in [2] there is a well-known definition used by a lot of researchers which is "fitness for use". Data quality can be mainly summarized in how the system fits into the reality, or how users really utilize the data in the system. [2].

Data quality can be assessed in terms of data quality dimensions. These data quality dimensions consist of timelines to ensure that the value is new, consistency to ensure that representation of the data is unchanging in all cases, completeness to ensure that the data is completed with no missing values, and accuracy to ensure that the recorded value is identical with the actual value. [1].

Incompleteness of data is a natural phenomenon as the data is usually generated, entered, or collected with missing values. Missing data can be defined as the values that are not stored for a variable in the observation of interest. There are three types of missingness of the data. First, the missing completely at random (MCAR): the variable is missing completely at random where the probability of missingness is the same for all missing variables. Second, the Missing at random (MAR): Variable is missing at random where the probability of missingness is depending only on an available information. This type can also be named as missing conditionally which means missing with a condition; for an example if gender is male, they will leave questions related to women in the survey empty. Third, the Not Missing at Random (NMAR) data where the missingness probability is not random and it depends on the variable itself and can't be predicted from another variable in the dataset. [3].

Missing data occurs in many types of the data sets but in specific it occurs with a very high percentage in the time series data. Time series data is a type of data that usually have incompleteness given to its nature. Time series data exist in nearly every scientific field, where data are measured, recorded, and monitored over time. Consequently, it is understandable that missing values may occur. Also, most of the time series data are collected by sensors and machines which is another reason for the occurrence of the missing values. [4].

This paper aims to ensure the data quality of time series data. More specifically, it aims to ensure the completeness dimensions of the time series data that suffers from missing value. Towards this aim, two novel techniques for imputing the missing values in time series data are proposed and compared with traditional techniques. The two proposed techniques impute the missing value by calculating the k-nearest neighbour between the missing value and the other values. Then it calculates a weight for each value in the nearest neighbours using fuzzy membership functions. Two fuzzy membership functions are used which are: the gaussian membership function and the triangular membership function. After calculating the weights, the data values and their weights are used in the weighted mean function to calculate the imputed value. The accuracy of the proposed techniques is evaluated by using three traditional classifiers: Neural Network, Naïve Bayes, and Decision Tree. Evaluation Results shows that the two proposed techniques have higher accuracy than the traditional data imputing techniques. In addition, it also shows that the triangular membership function yields higher accuracy rather than the gaussian membership function.

The rest of this paper is organized as follows: Section 2 presents the related work and some techniques used in imputing the missing values. Section 3 and 4 includes the summarization of the proposed techniques and the results. Finally, the paper is concluded in section 5.

II. Related Work

A lot of methods with different techniques have been proposed in the literature to solve the missing data problem. The management of missing data can be divided into three categories; deletion and ignoring methods; imputations methods and model-based methods. These categories will be discussed below with more details.

A. Deletion and Ignoring Methods

Deletion/Ignorance of missing values is recognized as the simplest way in handle missing values. Authors in [5] proposed the traditional techniques for dealing with missing data. The listwise deletion algorithm was proposed where an entire record is excluded from the data set if any value is missing. The pairwise deletion method was also proposed where the method computes the correlation between missing and complete data to pair the correlated values and it only delete the un-correlated values. Listwise deletion would result in removing more data than the pairwise method. The drawback of this method is that it may be very risky in case of the missingness is a large portion of the data as it may interrupt the results of the analysis.

B. Imputation Methods

The imputation methods work by substituting each of the missing values by an estimate value. The hot and cold deck imputation is one of the best methods used in missing data imputation. In [6], they used the cold deck imputation for variables where it uses external sources such as a value from a previous survey. It imputes missing values called as recipients using similar reported values from previous survey. Cold deck imputation was performed through probabilistic record linkage techniques in order to find the best matching records from different data sources containing the same set of entities.

Another imputation technique was proposed by authors in [7] to generate an estimate value for the missing values. In [7], the authors proposed a technique that considers multiple imputations for imputing missing values. This technique works by imputing missing values n-times to correspond to the uncertainty of all the possible values that can be imputed. Then the values are analyzed in order to get a combined single estimate. As an example, you can choose two different techniques and use them together so you can take advantages of both techniques and avoid the disadvantages of these techniques.

C. Model-Based Methods

The model-based methods are the methods which imputes the missing values by using a predictive technique. These methods are mainly machine learning techniques that needs learning phase to be able to estimate missing values.

In [8], the authors work on the weather data for environmental factors and found out that this data set contains a lot of missing values. They calculated the percentage of missingness in the data to found out that 19% of the weather data for 2017 are missed. This percentage is big in these types of data and can cause misleading during the analysis that will be done on it. Four missing data imputation was applied on this data set. They divided the data sets into training and testing to measure the quality of the four imputation algorithms. The k-nearest neighbor (KNN) method results were the best results, and its results was so close to the original data with no missing values and the prediction model's performance is stable even when the missing data rate increases.

In [9], authors implemented a new approach that is based on vector autoregressive (VAR) model by combining prediction error minimization (PEM) method with expectation and minimization (EM) algorithm. They called this algorithm a vector autoregressive imputation method (VAR-IM). Their proposed system is applied on a real-world data set involving electrocardiogram (ECG) data. They used linear regression substitution and list wise detection as a traditional method to be compared with their proposed method VAR-IM. They concluded that the proposed method VAR-IM produced a large improvement of the imputation tasks as compared to the traditional techniques. This technique has three limitations, the first one is it only deal with data that is missing completely at random. The second limitation is the validity of the approach requires that the time series should be stationary. The third limitation is the percentage of missing data has significant impact on most missing data analysis methods, the proposed technique does not have the priority to be used if the percentage of missing data is quite low (say less 10%). Despite these limitations, the proposed technique provides an important alternative to existing methods for handling missing data in multivariate time series.

In [10] ,the authors propose a genetic algorithm (GA) based technique to estimate the missing values in datasets. GA is introduced to generate optimal sets of missing values and information gain (IG) is used as the fitness function to measure the performance of an individual solution. Their goal is to impute missing values in a dataset for better classification results. This technique works even better when there is a higher rate of missing values or incomplete information along with a greater number of distinct values in attributes/features having missing values. They compared their proposed technique with single imputation techniques and multiple imputations (MI) statistically based approaches on various benchmark classification techniques on different performance measures. They show that the proposed methods outperform when compared with another state-of-the-art missing data imputation techniques.

In [11], the authors used the gene expression data that are recognized as a common data source which contains missing expression values. In this paper, they present a genetic algorithm optimized k- Nearest neighbour algorithm (Evolutionary KNN Imputation) for missing data imputation. They focused on local approach where the proposed Evolutionary k- Nearest Neighbour Imputation Algorithm falls in. The Evolutionary k- Nearest Neighbour Imputation Algorithm is an extension of the common k- nearest Neighbour Imputation Algorithm which the genetic algorithm is used to optimize some parameters of k- Nearest Neighbour Algorithm. The selection of similarity matrix and the selection of the parameter value k can be identified as the optimization problem. They compared the proposed Evolutionary k- Nearest Neighbour Imputation algorithm with k- Nearest Neighbour Imputation algorithm and mean imputation method. Results show that Evolutionary KNN Imputation outperforms KNN Imputation and mean imputation while showing the importance of using a supervised learning algorithm in missing data estimation. Even though mean imputation happened to show low mean error for a very few missing rates, supervised learning algorithms became effective when it comes to higher missing rates in datasets which is the most common situation among datasets.

III. PROPOSED TECHNIQUE

In this paper, two techniques are proposed for imputing missing values in time series data. The two proposed techniques start by finding the K nearest neighbor data points for each data point containing a missing value for a certain feature. Then, the values of the missing feature in the nearest neighbor's data points are weighted using one of the two fuzzy membership functions: triangular fuzzy membership function and gaussian membership function. The missing feature value is then obtained using the weighted mean of the feature in nearest neighbors. Fig.1 Show the steps of the proposed technique.

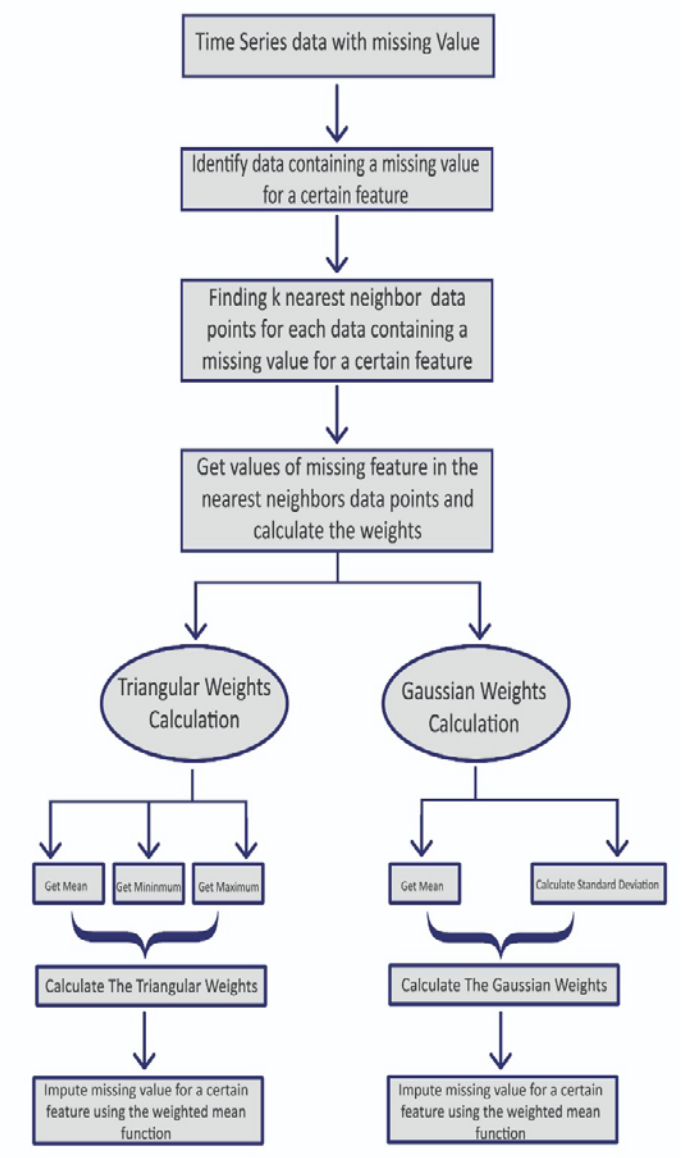

Fig. 1. Proposed Technique Block Diagram.

Two weighted functions are used to get the weight of each one of the nearest neighbors' data points for a certain missing feature before using them to impute the missing value. The triangular and the gaussian membership functions. The triangular membership weight function works by calculating the minimum, the maximum and the average of the nearest neighbors' values of the missing feature. Then, it calculates the weight for each value by using the triangular fuzzy membership function. Finally, the values and their weights are used in the weighted mean function to get the value of the missing data. Algorithm 1 show the exact details of Triangular fuzzy membership function.

Algorithm 1: Triangular fuzzy membership Function

1: Function Triangular fuzzy membership weights (Nearest Neighbors Values for the missing features)
Input: Nearest Neighbors Values for the missing features
Output: Missing feature value
2: Minimum= Minimum value of (Nearest Neighbors Values for the missing features)
3: Maximum= Maximum value of (Nearest Neighbors Values for the missing features)
4: Mean= Mean value of (Nearest Neighbors Values for the missing features)
5: Get weight for each Nearest Neighbors Values for the missing features using Triangular fuzzy membership function
Triangular function is defined by a minimum value a, a maximum value b, and a mean value m, where a < m < b.

$$\mu(x) = \begin{cases} 0 & x \le a \\ \dfrac{x-a}{m-a} & a < x \le m \\ \dfrac{b-x}{b-m} & m < x < b \\ 0 & x \ge b \end{cases}$$

6: Missing feature value = Calculate weighted mean using Nearest Neighbors Values for the missing features and weights for each one

$$\frac{\sum_{a \epsilon A} \text{Nearest Neighbours Values}(a) Triangular\ weight\ (a)}{\sum_{a \epsilon A} Triangular\ weight\ (a)}$$

7: End

The Gaussian membership weigh function works by calculating the mean, and the standard deviation of the nearest neighbors' values of the missing feature. Then, it calculates the weight for value by using the Gaussian fuzzy membership function. Finally, the values and their weights are used in the weighted mean function to get the value of the missing data. Algorithm 2 show the exact details of Gaussian fuzzy membership function.

Algorithm 2: Gaussian fuzzy membership function

1: Function Gaussian fuzzy membership weights (Nearest Neighbors Values for the missing features)
Input: Nearest Neighbors Values for the missing features
Output: Missing feature value

2: Standard Deviation = Standard deviation of (Nearest Neighbors Values for the missing features)
3: Mean= Mean value of (Nearest Neighbors Values for the missing features)
4: Get weight for each data value using Gaussian fuzzy membership function

$$f(x) = e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

5: Missing feature value = Calculate weighted mean using data value and weights for each one

$$\frac{\sum_{a \epsilon A} \text{Nearest Neighbours Values } (a) Gaussian\ weight\ (a)}{\sum_{a \epsilon A} Gaussian\ weight\ (a)}$$

6: End

IV. Performance Evaluation and Discussion

The objective of performance evaluation is to prove the effectiveness of the proposed technique against standard imputation techniques. Towards this objective, the proposed techniques were evaluated on six datasets with different percentages of missing values. The proposed methods were evaluated against traditional imputation techniques. All algorithms were evaluated using accuracy measure after considering a classification scenario on the data after imputation to find out the quality of the imputed data where three different classifiers were used. [12].

Six different Time series data sets were used in this paper. The datasets are chosen with missing values due to machine malfunctions, and simple human errors. The data available to build time series models are often characterized by missing values, due to various causes such as sensor faults, problems of not reacting experiments, not recovering work situations, transferring data to digital systems. Table 1 shows the details of each dataset.

Each data set was divided into training and testing sets. The training set are 75% of the whole dataset while the remaining 25% is considered as the testing set. Accuracy is used as an evaluation metric where the accuracy is obtained after using three well-known classification methods on Different 6 data sets. The classifiers are the Decision tree, Naive Bayes and artificial neural network classifiers. The artificial neural networks architecture is; 3 hidden layers and 200 epochs. Four imputation methods are used, the two proposed methods (Gaussian weighted mean and Triangular weighted mean) and two traditional methods (Average and weighted mean) [15, 16]. The accuracy between the 4 methods is computed. Results of the proposed and traditional techniques over the 6 datasets are summarized in the figures respectively.

TABLE I. Data Sets

Data Set	Name	No of Samples	No of Attributes	No of class	Percentage of missingness
Data set 1	Ozone Level Detection [13]	2536	73	2	1.28%
Data set 2	Data for Software Engineering Teamwork Assessment in Education Setting [13]	74	102	2	15.9%
Data set 3	Hybrid Indoor Positioning Dataset from WiFi RSSI, Bluetooth and magnetometer Data Set [13]	1540	65	2	27.3%
Data set 4	India COVID-19 data [14]	4838	7	70	2%
Data set 5	Us COVID-19 data [14]	8500	6	31	1.60%
Data set 6	HPI master [14]	4236	8	3	37.1%

As shown in Fig [2] to Fig [7], the two proposed techniques using fuzzy algorithms [17][18] gives higher accuracy than the traditional techniques. Fuzzy logic performs better than the non-fuzzy since fuzzy logic has the advantage of being grey not black nor white. As fuzzy logic uses membership functions, it can answer the uncertainties generated from non-fuzzy logics where you must choose between two options. Membership functions gives each value a membership value in each class rather than a binary decision "belongs to or not belong to". Fuzzy logic has multiple membership functions (Gaussian, triangular, Trapezium, …etc). Membership functions are equally good in performance but usually Gaussian and triangular MFs are found to be closely performing well and better than other types of membership functions. The choice of which of the functions to use depends entirely on the size, problem type and data distribution.

The evaluation results show that the proposed triangular weighted mean technique performs better in terms of accuracy than that of the proposed gaussian weighted mean technique. Triangular MF has many advantages over the gaussian MF as; simple to implement, more convenient, response quickly, and fast computation [19]. Also, the datasets used in this paper were found not to be normally distributed where triangular MF usually works better with these types of data. The normality of the six datasets were tested using Kolmogorov-Smirnov test [20] and it is found that they are not following the normal distribution. The Kolmogorov-Smirnov test, which is also known as KS test returns a test decision for the null hypothesis that the data in vector x comes from a standard normal distribution, against the alternative that it does not come from such a distribution. In addition, it was found that the Triangular MF gives higher weights to the values near to

the mean value and gives less weights to the values far from the mean until it reaches zero weight at the farthest two values from the mean. This would result higher weights to more representative values and consequently better imputations for the missing values.

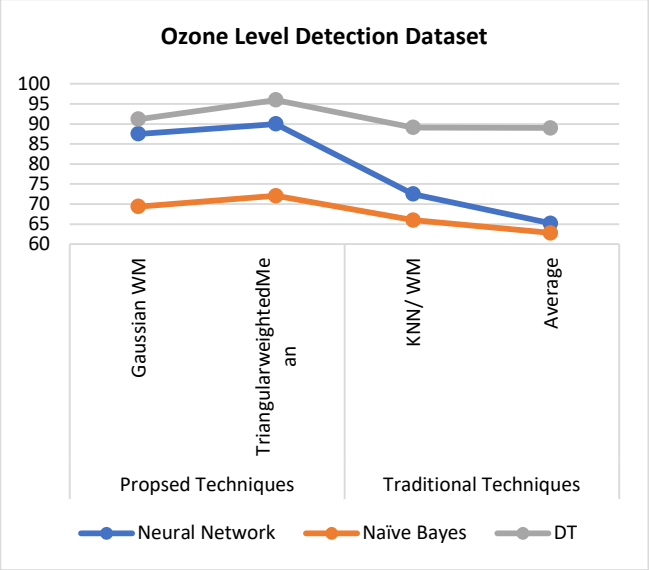

Fig. 2. Results for Ozone Detection Dataset.

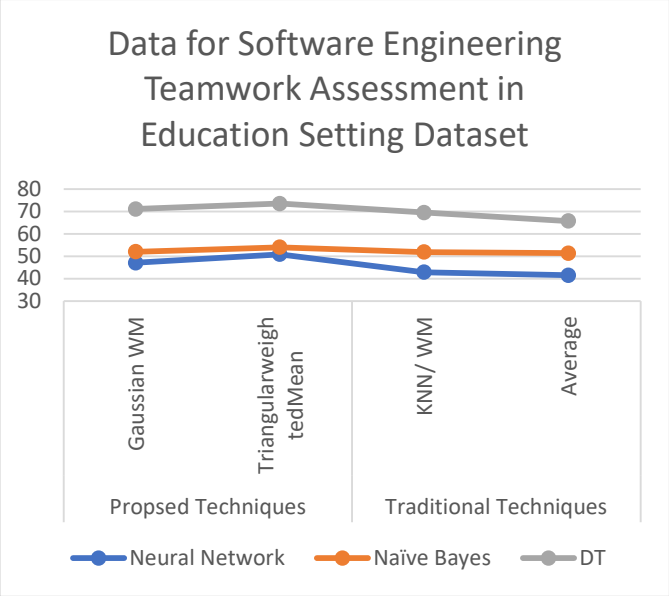

Fig. 3. Results for Data for Software Engineering Teamwork Assessment in Education Setting Dataset.

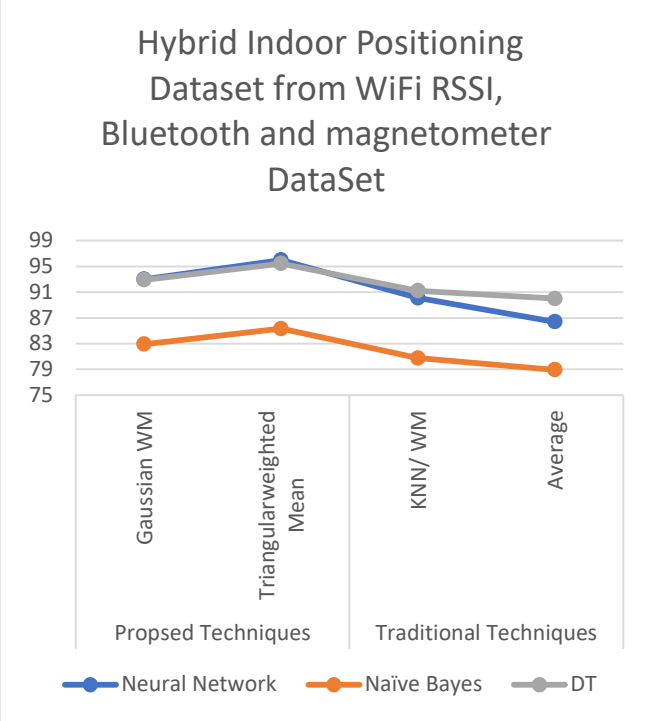

Fig. 4. Results for Hybrid Indoor Positioning Dataset from WiFi RSSI, Bluetooth and Magnetometer Dataset.

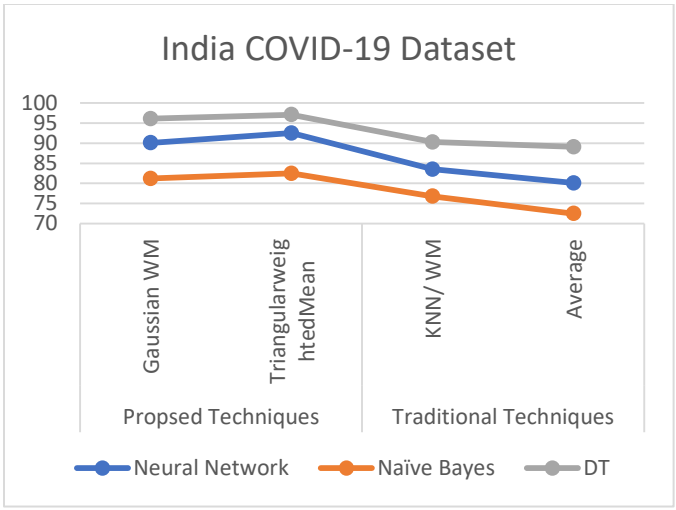

Fig. 5. Results for India COVID-19 Dataset.

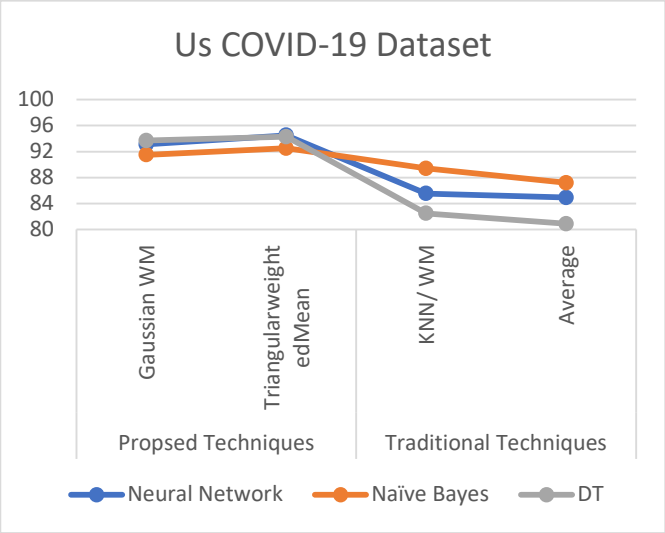

Fig. 6. Results for Us COVID-19 Dataset.

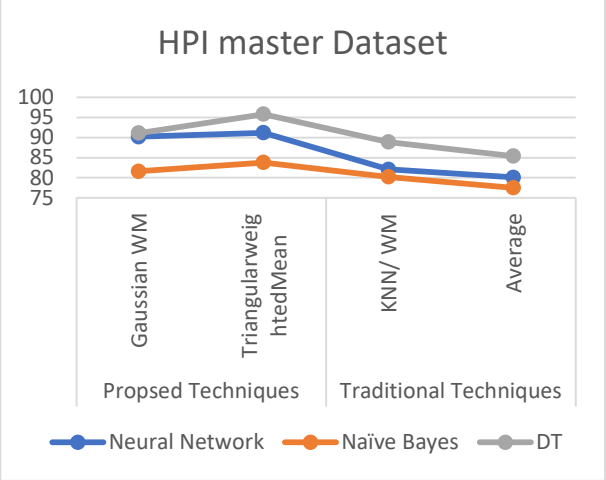

Fig. 7. Results for HPI Master Dataset.

V. Conclusion

The paper introduced two proposed techniques based on the fuzzy logic while imputing missing values in time series data. The first proposed technique is the Gaussian weighted mean technique. This technique uses the KNN first to find the nearest neighbours then it gives to each neighbour a weight using the gaussian membership function, these weights is sent to the weighted mean function to calculate the imputed value. The second proposed technique is the Triangular weighted mean technique. This technique uses the KNN first to find the nearest neighbours then it gives to each neighbour a weight using the triangular membership function, these weights is sent to the weighted mean function to calculate the imputed value. The results of the two proposed techniques were compared with other two traditional techniques. The results output is that the two proposed techniques have higher accuracy than the traditional imputation techniques. Based on the experiments conducted in this paper it can be concluded that fuzzy membership functions can have better accuracy, and this is due to its behaviour in dealing with the data as it gives a membership value for each point. Also, the results of the two proposed techniques were compared to find out that the triangular fuzzy membership function has higher accuracy than the gaussian membership function. Many different tests, and experiments have been left for the future due to lack of time. Future work concerns deeper analysis for the data, new proposals to try different methods. We also can start forecasting the new data in the future as we now have a complete data set.

References

[1] Kumar, p.v., p. Scholar, and m.v. gopalachari, a review on prediction of missing data in multivariable time series.

[2] Pratama, I., et al. A review of missing values handling methods on time-series data. in 2016 International Conference on Information Technology Systems and Innovation (ICITSI). 2016. IEEE.

[3] Tong, G., F. Li, and A.S. Allen, Missing data. Principles and practice of clinical trials, 2020: p. 1-21.

[4] Rantou, K., Missing Data in Time Series and Imputation Methods. University of the Aegean, Samos, 2017.

[5] Williams, R., Missing data part 1: Overview, traditional methods. University of Notre Dame, 2015: p. 1-11.

[6] Jayamanne, I.T., Cold Deck Imputation for Survey Non-response Through Record Linkage, in International Statistical Conference 2017 IASSL. 2017.

[7] Rubin, D.B., Multiple imputation after 18+ years. Journal of the American statistical Association, 1996. 91(434): p. 473-489.

[8] Kim, T., W. Ko, and J. Kim, Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. Applied Sciences, 2019. 9(1): p. 204.

[9] Bashir, F. and H.-L. Wei, Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm. Neurocomputing, 2018. 276: p. 23-30.

[10] Shahzad, W., Q. Rehman, and E. Ahmed, Missing data imputation using genetic algorithm for supervised learning. International Journal of Advanced Computer Science and Applications, 2017. 8(3): p. 438-445.

[11] De Silva, H.M. and A.S. Perera, Evolutionary k-nearest neighbor imputation algorithm for gene expression data. ICTer, 2017. 10(1).

[12] Flach, P. Performance evaluation in machine learning: The good, the bad, the ugly, and the way forward. in Proceedings of the AAAI Conference on Artificial Intelligence. 2019.

[13] https://archive.ics.uci.edu/ml/datasets.php.

[14] https://www.kaggle.com/.

[15] Meng, Z., Ground Ozone Level Prediction Using Machine Learning. Journal of Software Engineering and Applications, 2019. 12(10): p. 423-431.

[16] Petkovic, D., et al. Using the random forest classifier to assess and predict student learning of software engineering teamwork. in 2016 IEEE Frontiers in Education Conference (FIE). 2016. IEEE.

[17] Kreinovich, Vladik, Olga Kosheleva, and Shahnaz N. Shahbazova. "Why triangular and trapezoid membership functions: A simple explanation." Recent Developments in Fuzzy Logic and Fuzzy Sets. Springer, Cham, 2020. 25-31.

[18] Radhakrishna, Vangipuram, et al. "Design and analysis of a novel temporal dissimilarity measure using Gaussian membership function." 2017 international conference on engineering & MIS (ICEMIS). IEEE, 2017.

[19] Sadollah, A., Introductory chapter: which membership function is appropriate in fuzzy system?, in Fuzzy logic based in optimization methods and control systems and its applications. 2018, IntechOpen.

[20] Godina, R. and J.C. Matias. Improvement of the statistical process control through an enhanced test of normality. in 2018 7th International Conference on Industrial Technology and Management (ICITM). 2018. IEEE.