Available at https://www.iasj.net/iasj
Iraqi Academic Scientific Journals
Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Research Article

A Breast Cancer Diagnosis Based on Missing Value Imputation and REP Tree Technique

Heba Adnan Raheem
Department of Computer Science, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq

Article Info
Article history:
Received 2 -9-2025
Received in revised form16-9-2025
Accepted 30-9-2025
Available online 30 -9 -2025
Keywords: Breast cancer, missing values, REP tree.

Abstract:

Missing information or worth in a dataset can influence the execution of a classifier, which prompts trouble in separating useful data from datasets. The principal objective of our work is managing missing qualities by utilizing ascription strategies for it, talking about their ease of use, and discussing their relevance in the case of information sets with the goal of being more reasonable for information mining investigation and the arrangement. This paper likewise talks about the "REP Tree" information mining approach that has been used for breast cancer analysis in the wake of a preprocessing step (dealing with missing values), with the ultimate goal of improving the order of information (diagnosis) or conclusion. Experimental results proved that all used algorithms are efficient for dealing with missing values in the data set and diagnosis. We conducted an analysis of the Wisconsin dataset from UCI machine learning with the aim of developing accurate breast cancer prediction models using data mining techniques. The results of the proposed system demonstrated that data distortion can be reduced while classification dataset accuracy remains high. DI imputation was shown to be a superior strategy for impute missing values with accuracy 93%, thus the system successfully achieves its requirements.

Corresponding Author E-mail: hiba.adnan@uokerbala.edu.iq

Peer review under responsibility of Iraqi Academic Scientific Journal and University of

1. Introduction

Missing values generally show up as "NULL" values in the dataset table. For instance, the "?" symbol is used by ARFF papers to indicate missing qualities, which is one example of how different document designs handle missing information. It is possible to distinguish between these kinds of missing attributes. However, lacking attributes can also manifest as out-of-bounds examples or inaccurate data. It is considerably more difficult to find out this information, and it must be removed before anticipated scrutiny [1].

On the basis of the afflicted cell type, around 200 distinct cancers have been identified. The topic of breast tumors is the focus of this article. Breast cancer is the most well-known type of cancer among females worldwide [2]. Approximately 200 distinct cancers have been identified, each of which is categorized according to the cell type that has been compromised.

Predicting the result of an infection is a testing process. Data mining strategies have a tendency to streamline the expectation fragment. Computerized apparatuses have made it possible to gather huge volumes of therapeutic information, which is made accessible to the medical research groups. The outcomes are an expanding popularity of data mining strategies to distinguish examples and knowledge that we need to predict the consequences of sickness using data sets [3]. There are the key limitations of using missing value imputation in breast cancer diagnosis, such as (computational cost, feature importance distortion, false completeness, evaluation difficulty).

2. Related Works on Dealing with Missing Values

In this section, we have presented some of the related research as mentioned in [12].

In [4], well-known techniques for managing missing data are discussed. KNN, C4.5, and MMI are the most commonly utilized techniques for managing missing data nowadays. In summary, MMI will be a superior strategy for nominal data, and KNN will be a superior technique for numeric data. There exist numerous techniques for managing missing information; however, none is completely superior to the others. Different situations call for different configurations.

In [5], this work illustrates that K-Means and KNN techniques give quick and precise methods for assessing missing values.

In [6], in this paper, they try different things with twelve datasets to assess the impact on the misclassification error rate of four techniques for managing missing qualities: "the case deletion method, mean imputation, median imputation, and KNN imputation procedure." Incorporating missing values into a data set can impact how a classifier trained on that data set performs.

In [7], this work illustrates the efficiency of the C4.5 technique to treat missing data and K-means for missing data imputation. The C4.5 method for treating missing data and K-means for imputation of missing data are the subjects of this study's inquiry into the efficacy and conduct of missing data treatment. These techniques are dissected by embedding diverse rates of missing information into various traits of the four usually utilized information sets. For the purpose of imputing missing data, the suggested method exclusively uses numerical properties.

In [8], the precision of classifiers produced by machine learning algorithms degrades in most cases when training data is inadequate, as demonstrated in this study. For instance, mean imputation (MEI) typically does not yield significantly improved classifiers.

In [9], this paper first compares a few distinct strategies—prescient value imputation and classification trees are applied to instances with missing data using distribution-based imputation and reduced models in C4.5.

In [10], meteorological data mining aims to uncover hidden patterns within extensively available meteorological data, enabling the transformation of the retrieved information into actionable knowledge.

In [11], missing values and their associated difficulties are prevalent in the data cleansing process. He investigates the problem of missing values in monotonous data sets. He proposes a straightforward preprocessing strategy, which, when utilized with different methods, helps in disposing of missing values and helps in keeping the data set's monotony. He proposes a heuristic approach to address the practical and complex challenge of cleaning concealed missing data. Notice that a few strategies have been applied in order to deal with missing data in datasets, as mentioned in [4].

3. Overview of the Missing Value Problem

A missing value refers to a value that we aimed to obtain throughout the process of data collection (such as meetings, assessments, or observations). Missing values may arise due to respondents not answering all questions in a survey, errors during manual data entry, inaccurate estimations, flawed analyses, or certain data being modified or unrecorded, among other reasons [1]. Missing data treatment strategies can be separated into the following three classifications [1].

3.1. Reducing the Data Set

The quickest and easiest solution to the missing values imputation problem is to reduce the dataset size and remove all missing values. This should be achievable through the elimination of rows containing tests with missing data [13] or the end of properties (columns) with missing values. Both methodologies can be joined. Disposal of all samples is also called complete case analysis. Only with access to massive data sets is it possible to dispose of all samples, and only in very low test rates do missing values occur. Disposal of characteristics with missing values during analysis is an illogical arrangement in making inductions about these properties. Both systems are wasteful procedures since they typically decrease the data substance of the information [14].

3.2. Treating Missing Attribute Values as Special Values

This technique manages the obscure characteristic values utilizing an entirely different methodology. For attributes that include missing values, we don't try to identify a known value for the attribute; instead, we treat the missing values as another value and handle them as such [15]. Instead of keeping track of the attribute's value, we record the fact that its worth does not exist. Assuming we manage these numbers, this process works fine because they won't affect any subsequent research.

3.3. Imputation Methods

The term "imputation" refers to a set of methods that use previously collected values to fill in gaps. The goal is to fill in missing values by making use of known relationships that are present in the valid data set [14].

4. Wisconsin Dataset

The description of the data set and its types is as follows:

4.1. Breast Cancer

Cancer of the breast develops from cancerous tissue, most commonly in the lobules that supply the milk pipes or the inside of the milk pipes themselves. Bosom malignancy happens in both men and women, despite the fact that the former type is uncommon. Worldwide, it is still the most common kind of tumor found in women. Women still face a significantly higher risk of mortality from this condition due to delayed diagnosis, even with better treatment options. There were 40,000 female fatalities and 232,670 new cases verified in the US in 2014 [16, 17].

4.2. Types of Breast Cancer and Data Description

Breast cancer can manifest in numerous types, depending on the specific region of the breast affected. Breast cancer is classified into two primary categories, as illustrated in Figure 1 [16].

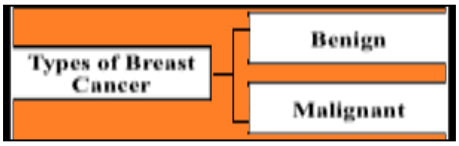

Figure 1: Types of breast cancer

Table 1 presents a concise overview of the dataset under consideration [16].

Table 1: Overview of breast cancer dataset

Dataset	No. of Attributes	No. of Instances	No. of Classes
Wisconsin Breast Cancer (Original)	11	699	2

The dataset's attribute details are presented in Table 2 [16, 18].

Table 2: Wisconsin breast cancer dataset attribute

S. No.	Attribute	Range
1	Sample Code Number	Id number
2	Clump Thickness	1 – 10
3	Cell Size Uniformity	1 – 10
4	Cell Shape Uniformity	1 – 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 – 10
11	Class	2 (Benign) or 4 (Malignant)

5.Methodology

The structure of the proposed system is summarized in Figures 2 and 3.

5.1. Experimental Procedure

This section will describe the procedure of the proposed system as follows:

5.1.1. Input Dataset

Info is a dataset that is put away in a file that contains sensitive data, which is to be protected, such that each point (row) of data is a sequence of real or integer value $X = x1, x2, x3, \cdots, xn$. The dataset contains "m" rows of data. In our case, the Breast Cancer Wisconsin datasets available at the University of California, Irvine (UCI) Repository are used.

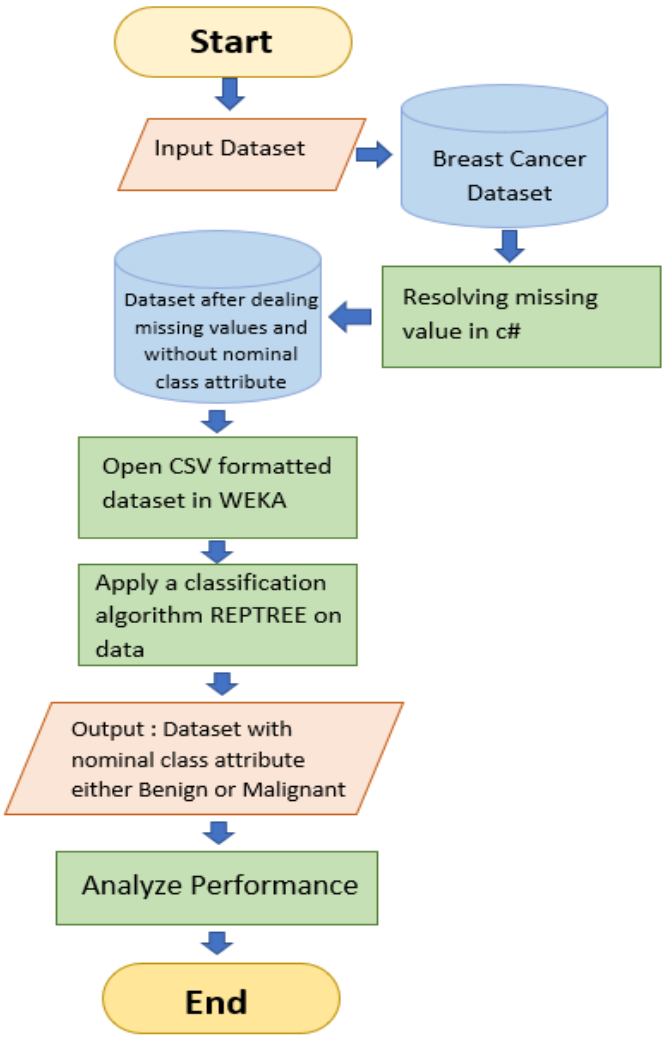

Figure 2: A Flow diagram of the proposed system

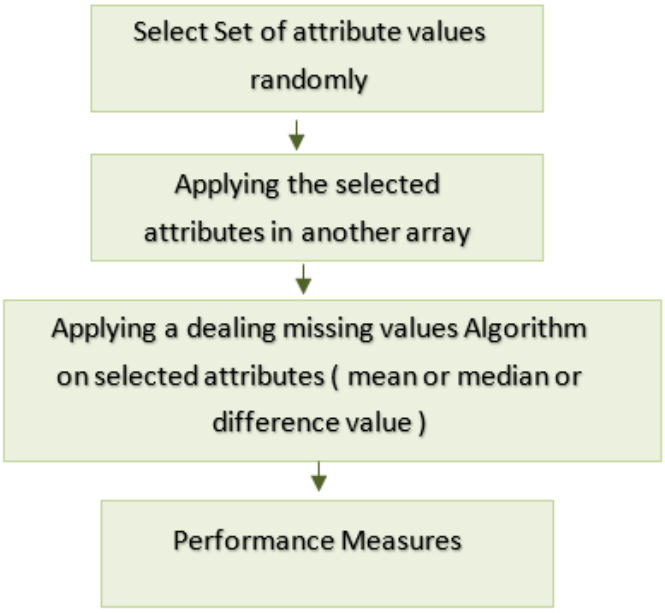

Figure 3: A flow chart of resolving missing values

5.1.2. Resolving Missing Values Stage

The steps of dealing with null values in the dataset are running in C#, and they will be summarized below:

- Select a set of attribute values randomly and apply it in another array: In this step, we are randomly deleting some of the values in our database (which are considered as missing values), and they can be saved in a new matrix value (missing values array), which is the same size as the original data set matrix. Any deleted value can be saved in the missing values array as the corresponding index in the original matrix. And the rest of the elements of the new matrix remain empty because we need this matrix only for comparing results and keeping the deleted values from being lost. We can show this step precisely in the results section.
- Applying a resolving missing values algorithm (mean, median, or substitution imputation) on the selected attributes: In this work, a missing values resolving algorithm is applied to a medical data set. The output is a dataset after dealing with missing values. There are three methods of dealing with the missing values problem that can be used in our work:
  - Mean Imputation (MI): This is the most commonly utilized strategy. It involves replacing the missing data for a specific feature with the average of all known estimates for that feature in the columns where the missing instance is located, as shown in Equation (1):

$$M = \frac{1}{n}\sum_{i=1}^{n} mi \qquad (1)$$

    Where, impute M as the output value for missing data.
  - Median Imputation (MDI): In this scenario, the median value of all known values for a certain attribute is used to replace any missing data in its columns. This strategy is additionally a prescribed decision when the distribution of the estimates of a given element is skewed. It will be replaced by Equation (2):

$$median(x) = \begin{cases} x_{|r+1} & if\ m\ is\ odd, i.e \\ \frac{1}{2}\left(x_{(r)} + x_{|r+1}\right) & if\ m\ is\ even, i.e \end{cases} \qquad (2)$$

o Difference Imputation (DI): This is a new imputation technique that is proposed in this paper to deal with missing values in a dataset. The difference imputation technique is based on substituting each deleted value that is selected arbitrarily in the dataset by the result of computing a difference of the maximum and minimum value over the whole dataset (for a selected attribute), as shown in the following steps:

a) Detect a set of attributes (arbitrary) in some records in the original dataset and save it in a new matrix.

b) Searching the range for each one of the deleted attribute values (min and max values for it) over the whole dataset, and computing the difference value between them.

c) The deleted attribute value is substituted with a difference imputation value that is computed from its column in the original data set, as shown in Equation (3):

$$DI(mi) = Max\ Value(mi) - Min\ Value(mi) \quad (3)$$

Where, impute DI is the output value for missing data, mi is a detected attribute value. Figure 4 illustrates Pseudo code of Difference Imputation (DI). The results of running imputation methods in C# will be shown in Figure 5. Datasets are loaded into the weak tool after being converted to .csv format following imputation.

```
BEGIN
1. INPUT: original_dataset
2. SELECT a set of arbitrary attributes (columns) to be removed
   deleted_attributes = [list of selected attributes]
3. CREATE an empty matrix new_matrix
4. FOR each record in original_dataset DO
      COPY all attributes EXCEPT the deleted_attributes into new_matrix
   END FOR
5. CREATE a dictionary attribute_ranges
6. FOR each attribute in deleted_attributes DO
      min_val = MIN(original_dataset[attribute])
      max_val = MAX(original_dataset[attribute])
      difference = max_val - min_val
      attribute_ranges[attribute] = difference
   END FOR
7. FOR each record in new_matrix DO
      FOR each attribute in deleted_attributes DO
         imputed_value = attribute_ranges[attribute]
         INSERT imputed_value into the corresponding position in the record
      END FOR
   END FOR
```

Figure 4: Pseudo code of Difference Imputation (DI).

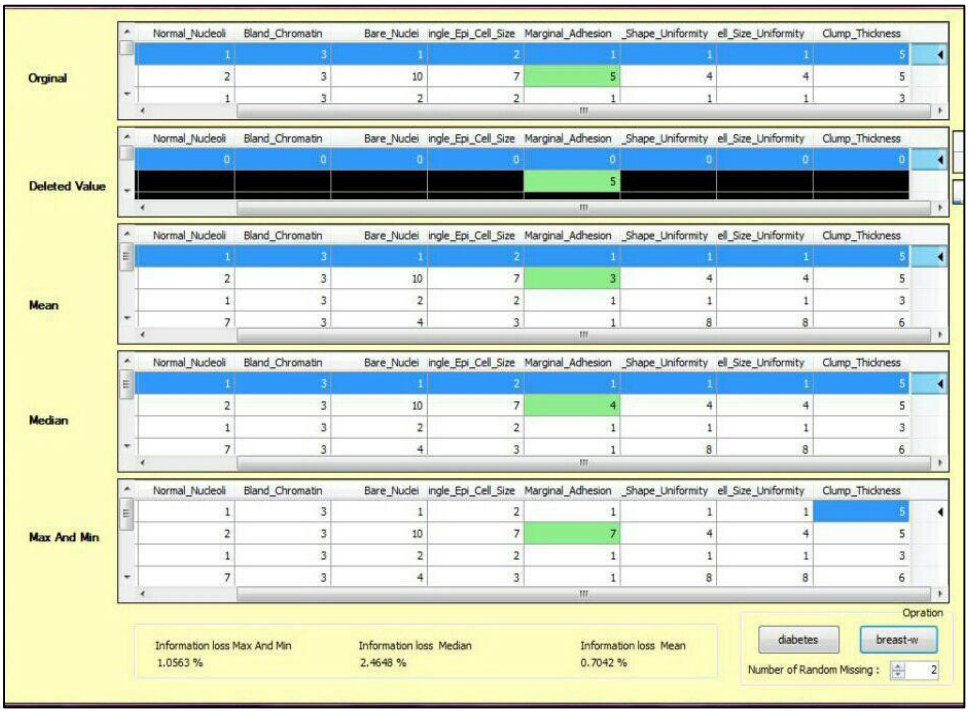

Figure 5: Implementation results for breast cancer Wisconsin dataset after applying missing values imputation techniques

5.1.3. Apply a Classification Algorithm on the Imputed Dataset

In the imputed dataset "REP TREE," the classification algorithm is applied to analyze which is the top technique for dealing with missing values. It learns from chosen trees quickly. Building a decision/regression tree and applying reduced error pruning based on entropy as an impurity metric. No more than a single sorting of numerical property values [19, 20]. Navigate to Explore and pick the "classification" option. Then, hit the "choose" button. Here, we select the "REPTREE" classifier from the "classification mode." Then, use the "cross-validation" option and click on the "start" button to begin the classification process. Below, you can see this procedure illustrated in Figures 6, 7, and 8.

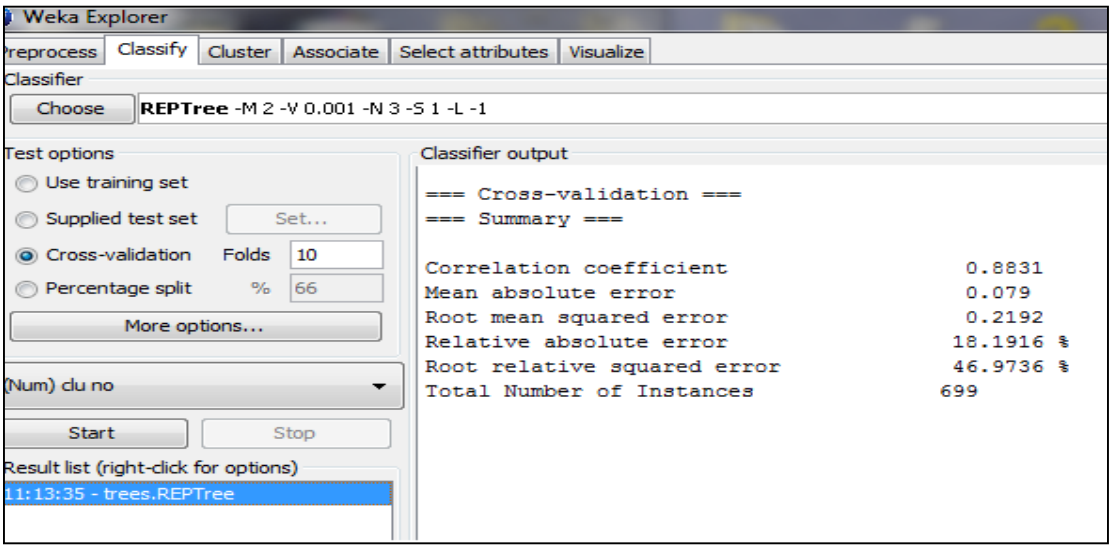

Figure 6. Result of REPTREE classifier for mean imputation

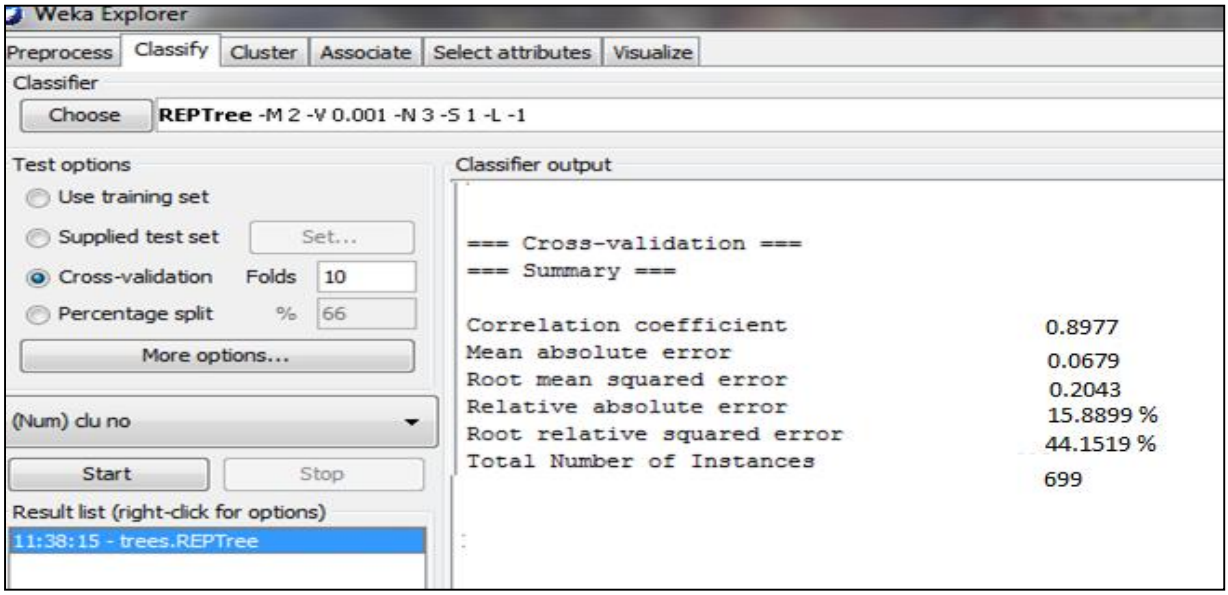

Figure 7: Result of REPTREE classifier for median imputation

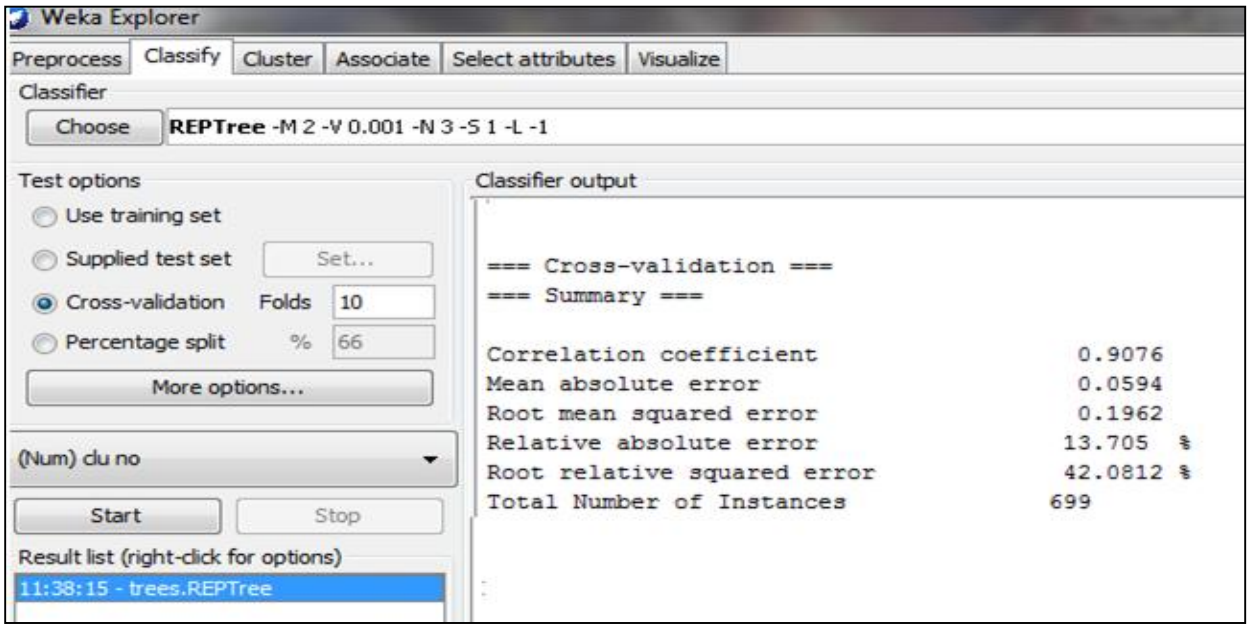

Figure 8: Result of REPTREE classifier for difference imputation

5.2. Experimental Result

This section will describe the performance factors that are used as follows:

- Distortion Ratio: When dealing with missing value procedures (imputation techniques), this performance factor is utilized to quantify the percentage of data set information distortion. It is measured by the percentage of the summation of the difference between the original value and the modified value for the deleted attribute in any record of the dataset, as well as the summation of original values for the dataset. The distortion ratio is calculated through Euclidean distance by Equation 4:

$$DR = \frac{1}{mn}\sum_{i=1}^{m}\left[\sum_{k=1}^{n}[Xik - Yik]^{\frac{1}{2}}\right]^{2} \quad (4)$$

Where m and n are the rows and columns of the dataset, respectively, Table 3 displays the distortion ratio results comparison.
Notice that the distortion of data on imputed data sets has minimal values, and all the proposed methods can provide values that are closest to the original values.

- Analyze Performance: Using the REEPTREE algorithm in the data mining program Weka, we can compare the three full datasets produced by the imputation algorithms and determine which one is the most effective. Table 4 displays a number of statistical analyses of our findings.
The findings illustrate that REPTREE classifiers with feature selection are an unrivaled method that can be used for breast cancer determination. According to experimental results, the accuracy of the DI imputation algorithm is 93.11%, which is higher than the other two techniques. DI is the best way to handle data sets with missing values. Accuracy is characterized by:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Where True Positive (TP) is the number of positive samples correctly predicted, False Positive (FP) is the number of negative samples wrongly predicted as positive, False Negative (FN) is the number of positive samples wrongly predicted, and True Negative (TN) is the number of negative samples correctly predicted.

Table 3: Distortion ratio results comparison			
Data set	**Distortion Ratio**		
Breast cancer	**MI**	**MDI**	**DI**
	0.7042%	2.4648%	1.0563%

Table 4: Comparison of imputation techniques using REEPTREE classification algorithm			
Evaluation Criteria	**Imputation technique**		
	MI	**MDI**	**DI**
Correlation Coefficient	0.8831	0.8977	0.9076
Mean Absolute Error	0.079	0.067	0.059
Root Mean Squared Error	0.219	0.204	0.196
Relative Absolute Error	18.191%	15.889%	13.705%
Root Relative Squared Error	46.973%	44.151%	42.081%
Accuracy	92.91%	92.97%	93.11%

6. Conclusion

It is imperative that missing values be imputed prior to using the dataset, since missing values in the dataset pose a significant problem. For this study, we used a dataset on breast cancer in Wisconsin, where some variables are missing. Three methods, namely the mean, median, and difference, are employed to fill in these blanks. Use these missing methods on this dataset, and you'll get three full or imputed datasets. Experimental results proved that all used algorithms are efficient for dealing with missing values in data sets because the distortion of data has minimal values, and all the imputation methods can provide values that are closest to the original values. The next step is to load the imputed datasets into the weak tool. The REEPTREE classification algorithm is used to assess the accuracy of the three imputation methods on these imputed datasets. The findings show that DI imputation is the most accurate method. Evaluation criteria were used to examine the performance of the REEPTREE technique in relation to the breast cancer diagnosis issue. DI imputation was shown to be a superior strategy for imputing missing values. DI imputation was shown to be a superior strategy for impute missing values with accuracy 93%.

7. Future Scope

For data analysis, innovative methods for imputation of missing data can be applied. Different new issues utilizing missing data analysis can be composed and executed. Another classification algorithm can be utilized to approach missing data imputation in comparative analysis.

8. References

[1] Kabir, M.F., Tomforde, S,"A deep analysis for medical emergency missing value imputation,". In: ICAART (3), pp. 1229–1236 (2024).

[2] Getz, K., Hubbard, R.A., Linn, K.A.: Performance of multiple imputation using modern machine learning methods in electronic health records data. Epidemiology 34(2), 206–215 (2023)

[3] C. A. Pena-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis," *Artificial Intelligence in Medicine,* vol. 17, no. 2, pp. 131-155, 1999.

[4] L. Peng and L. Lei, "A review of missing data treatment methods," *Intelligent Information Management System Technology,* vol. 1, pp. 412-419, 2005.

[5] M. Malarvizhi and A. Thanamani, "K-NN classifier performs better than K-means clustering in missing value imputation," *IOSR Journal of Computer Engineering,* vol. 6, no. 5, pp. 12-15, 2012.

[6] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," in *Classification, Clustering, and Data Mining Applications*, Chicago, 2004: Springer, pp. 639-647.

[7] B. Mehala, P. R. J. Thangaiah, and K. Vivekanandan, "Selecting scalable algorithms to deal with missing values," *International Journal of Recent Trends in Engineering,* vol. 1, no. 2, pp. 80-83, 2009.

[8] X. Su, T. M. Khoshgoftaar, and R. Greiner, "Using imputation techniques to help learn accurate classifiers," in *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, Dayton, OH, USA, 2008: IEEE, pp. 437-444.

[9] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," *Journal of Machine Learning Research*, vol. 8, pp. 1625-1657, 2007.

[10] M. A. Kalyankar and S. Alaspurkar, "Data mining technique to analyze the metrological data," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, no. 2, pp. 114-118, 2013.

[11] B. Doshi, "Handling Missing Values in Data Mining," *Data Cleaning and Preparation Term Paper*, pp. 1-11, 2011.

[12] M. Gimpy, "Missing value imputation in multi attribute data set," *International Journal of Computer Science & Information Technology,* vol. 5, no. 4, pp. 1-7, 2014.

[13] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[14] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of missing data in industrial databases," *Applied Intelligence*, vol. 11, pp. 259-275, 1999.

[15] J. W. Grzymala-Busse and M. Hu, "A comparison of several approaches to missing attribute values in data mining," in *International Conference on Rough Sets and Current Trends in Computing*, 2000: Springer, pp. 378-385.

[16] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, "Diagnosis of breast cancer using decision tree data mining technique," *International Journal of Computer Applications,* vol. 98, no. 10, pp. 16-24, 2014.

[17] Breast Cancer Organization, "Breast Cancer Risk Factors," 2025. [Online]. Available: https://www.breastcancer.org/risk/risk-factors

[18] D. V. Chaurasia and S. Pal, "A novel approach for breast cancer detection using data mining techniques," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 1, p. 17, 2017.

[19] H. W. Ian and F. Eibe, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers, 2005.

[20] K. Wisaeng, "A comparison of decision tree algorithms for UCI repository classification," *International Journal of Engineering Trends and Technology,* vol. 4, no. 8, pp. 3393-3397, 2013.