Journal of Applied Informatics and Computing (JAIC)
Vol.7, No.1, Juli 2023, pp. 50~56
e-ISSN: 2548-6861

Predicting Missing Value Data on IEC TC10 Datasets for Dissolved Gas Analysis using Tertius Algorithm

Noper Ardi[1]*, Supardianto[2]*, Ahmadi Irmansyah Lubis[3]*
* Teknik Informatika, Politeknik Negeri Batam
noperardi@polibatam.ac.id [1], supardianto@polibatam.ac.id [2], ahmadi@polibatam.ac.id [3]

Article Info

Article history:

Received 2023-05-03
Revised 2023-07-21
Accepted 2023-07-26

Keyword:
Tertius Algorithm,
Prediction,
J48,
Random Forest,
IEC TC10

ABSTRACT

IEC TC10 is the most widely used Dissolved Gas Analysis (DGA) measurement dataset nowadays. Many DGA-based studies have been carried out using conventional methods and methods based on Artificial Intelligence Techniques (AITs). DGA is a diagnostic test performed on power transformers to detect and diagnose potential faults. The test involves analyzing the gases that are dissolved in the transformer oil, which can provide important information about the condition of the transformer. DGA is a widely used technique for transformer monitoring and maintenance in the power industry. However, this dataset is not perfect. There are still many problems in this dataset, one of which is the problem of missing value data. This problem will be significant if not appropriately handled. More reliable data from DGA measurement results is an in-dispensable reference in diagnosing faults in power transformers. This study focuses on dealing with the problem of missing value data using the Tertius algorithm, then testing the results using the J48 and Random Forest algorithms. The results obtained are pretty significant. Of the total 56 missing data, 36 could be predicted perfectly. And received the results of measuring accuracy using the J48 method of 62.73% and the Random Forest method of 70.71%. This result shows that the approach we applied is relatively good for handling missing values in IEC TC10 datasets.

I. INTRODUCTION

The transformer is one of the key components of a power plant. In electricity systems, the power transformer converts high voltage electricity to low voltage electricity and vice versa. Working for a very long time, transformers often rarely get intensive attention. Yet, with the intensity of work and functionality, transformers are very vulnerable to damage that can cause problems in electricity distribution to consumers. Transformer components that are vulnerable to damage include liquid insulation material in hydrocarbon oil. After being used for a long time, this material will experience degradation in quality.

An engine fault in a power transformer will happen when the engine is used for a long time. The formation of gases, such as ethane (C2H6), methane (CH4), hydrogen (H2), acetylene (C2H2), ethene (C2H4), carbon dioxide (CO2), and carbon monoxide (CO) that occurs in oil Power transformers as a result of decomposition of transformer oil occur due to thermal faults in more extreme cases such as overheating or electrical faults[1]. This process is also followed by the erosion of the insulation paper on the walls of the power transformer.

Previous studies have assessed that some of the gaseous compounds produced can diagnose the condition of power transformers. A. Siddique[2] researched the transformer fault analysis approach using the MultiLayer Perceptron Neural Network (MLPNN) combined with the conventional Roger and Doernenburg ratio approaches. By combining this method, it is proven to produce better accuracy results. H.Malik investigated the diagnosis of DGA based on the most influential parameters using Extreme Machine Learning (ELM)[2]. Furthermore, research conducted by Setiawan et al. regarding the diagnosis of power transformer faults based on DGA uses rough set theory using the dataset in the IEC 60599[3][4].

Research conducted by A. Pramono et al. in diagnosing power transformers using DGA analysis and the J48 Algorithm[5]. The results obtained are quite good compared to the previous conventional method. Furthermore, Y. Benhammed et al. diagnosed the condition of power transformer oil using KNN and Naïve Bayes Classifier.[6][7]

Dissolved Gas Analysis (DGA) is a method used to measure the condition of a power transformer based on the type and amount of gas dissolved in transformer oil caused by the decomposition process of insulator oil. Gas contaminant particles are formed in the form of, Ethane (C_2H_6), Methane (CH_4), Hydrogen (H_2), Acetylene (C_2H_2), Ethene (C_2H_4), Carbon Dioxide (CO_2), and Carbon Monoxide (CO)[3].

Many DGA evaluation approaches are developed, both conventional techniques and Artificial Intelligence Techniques (AITs) based on ANN, SVM, Fuzzy Logic, and Type-2 Fuzzy Logic. However, these methods did not get significant results[8].

Based on data obtained from the IEEE guide for interpreting Gases Generated in Oil Immersed Transformers, ANSI/IEEE std. C57.104, 1991 – rev 2008. Many approaches have been developed in DGA analysis, both conventional approaches, and approaches based on Artificial Intelligence Techniques (AITs). Conventional approaches include the Doernenburg Ratio Method, Roger Method, Key Gas Method, and Duval Triangle [1].

Transformers as high-voltage equipment cannot be separated from the possibility of experiencing abnormal conditions triggered by internal or external factors. These unnatural conditions, in general, can be in the form of overheating, corona, and arcing, which can cause disturbances to their performance. One method to find out whether there is an abnormality in a transformer is to know the impact of the anomaly of the transformer itself. To determine the effect of eccentricity on the transformer, used Dissolved Gas Analysis (DGA) method was used. DGA is a method used to measure the condition of a power transformer based on the type and amount of gas dissolved in the transformer oil caused by the decomposition process of petroleum and insulator[9].

When an abnormality occurs in the transformer, the insulating oil as a hydrocarbon chain will decompose due to a large amount of energy. It will form hydrocarbon gases that dissolve in the insulating oil. DGA is a process to calculate the levels of hydrocarbon gases that are formed due to disturbances.

There are two causes of changes in gas composition in the transformer, namely disturbances due to heat and electrical disturbances. Decomposition of gases due to heat occurs due to oil and solid materials from the insulation in the transformer. The gas formation process generally occurs due to the ionic bombardment process. There is little heat generated, associated with low energy and partial energy dissipation.

Another approach that can be used is to assume that all the gaseous hydrocarbons in the oil decompose into the same substance and each product of the resulting importance is the same as one another. In thermodynamic models, it is possible to calculate the pressure of each part produced gas as a function of temperature, using the equilibrium constant equations for the relevant decomposition reactions.

IEC TC10 is a widely used dataset as a reference in dealing with DGA problems. However, this dataset is not perfect. There are many problems related to the data contained in this dataset. One of these problems is associated with the number of missing or empty data values. This problem is also known as missing value data. In general, the problem of lost value data is not a significant problem if the quantity of available information is vast. However, this problem will significantly affect performance if the available data is small, as is the case in the case of this IEC TC10 dataset[10].

Therefore, this research will focus on handling the problem of missing value data. In this study, the Tertius algorithm will be used to deal with the issue of missing value data, and the J48 algorithm and Random forest will be used to measure the accuracy of the measurement.

II. METHODOLOGY

A. Experimental Design

Several stages will be conducted in this research. The Stage mainly consists of two main stages: the data preparation stage, which includes the fixed of missing data value, and the data testing stage. The rest of the scene can be seen in the Figure below.

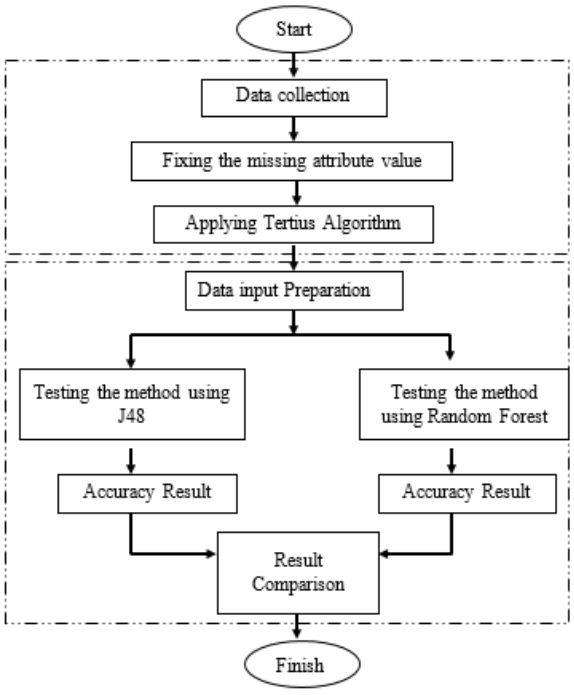

Figure 1. Research Flowchart

The stages of research to be carried out are divided into two main stages, namely the data preprocessing step and the data testing stage[11][12].

III. RESULT AND DISCUSSION

A. Data collection

This study will obtain 167 DGA data objects from the publication dataset IEC60599 by M.Duval[10][13]. The IEC TC 10 database is one of the methods for using Dissolved Gas Analysis (DGA) presented in the NEW IEC Publication 60599 authored by Michel Duval. The data contained in the IEC TC 10 database is derived from testing approximately 10,000 transformers[1].

Within these datasets, seven types of gas molecules are employed as tools to identify faults. These gas molecules are Hydrogen (H_2), Methane (CH_4), Ethane ($C_2 H_6$), Ethylene ($C_2 H_4$), Acetylene (C_2H_2), Carbon Monoxide (CO), and Carbon Dioxide (CO_2). The types of faults occurring in the transformers vary based on the concentrations of these gases present in their insulating oil[3]. From 167, the data is divided into sections, namely 9 PD condition data, 26 data for D1 condition, 48 data for D2 state, 16 data for T1 and T2 conditions, 18 data for T3 state, and 50 data for the normal condition from IEC TC10 database.

B. Fixing the missing value

Missing data is a very crucial problem in the case of DGA. In this research Tertius algorithm will be used. The Tertius Algorithm is one type of association algorithm that can be employed to explore relationships among data within a dataset. This algorithm was first introduced in the work of Peter A. Flach and Nicholas Lachiche in 2001[14][15].

Association rule mining aims to find interesting associations or correlations among items in a dataset. These associations are commonly expressed in the form of rules, often referred to as "if-then" rules. For example, "If item A and item B are present, then item C is likely to be present as well." The Tertius algorithm operates on a dataset, searching for frequent itemsets and generating association rules based on those itemsets. The process can be summarized as follows[16]:

1. Frequent Itemset Generation: The algorithm starts by identifying all frequent itemsets in the dataset. A frequent itemset is a set of items that occur together in the data with a frequency above a specified minimum support threshold. The support threshold is a user-defined parameter that determines the minimum frequency required for an itemset to be considered "frequent."
2. Rule Generation: From the frequent item sets, the algorithm generates association rules. An association rule consists of an antecedent (the "if" part) and a consequent (the "then" part). The antecedent and consequent are subsets of items from a frequent itemset. The algorithm generates rules with various combinations of antecedent and consequent to capture different associations.
3. Rule Evaluation: The generated rules are evaluated based on a measure such as "confidence" or "confirmation." Confidence measures the likelihood of the consequent being present given the antecedent. Higher confidence values indicate stronger associations. On the other hand, confirmation is used in the Tertius algorithm and measures the goodness of the rule based on statistical significance.
4. Rule Selection: After evaluating the rules, the algorithm selects the most interesting and significant rules based on a predefined threshold or user-defined criteria. These selected rules are considered to be meaningful associations in the data.

The Tertius algorithm's distinguishing feature lies in its use of "confirmation" to evaluate and prioritize the generated rules. Smaller values of confirmation indicate better quality rules, meaning the algorithm tends to favor rules with less redundancy and higher significance[17].

The selection of the Tertius Algorithm for data preprocessing in this research is justified by its superior performance compared to other association algorithms. In essence, the Tertius Algorithm operates similarly to other association algorithms, which involves searching for relationships among data in the dataset. However, the Tertius Algorithm incorporates the use of confirmation values in selecting the generated rules, where smaller confirmation values lead to better-quality rules[18].

Due to a large amount of missing data, it can cause low performance in calculating the accuracy of data classification. There are 56 missing data in the IEC TC10 dataset, divided into 9 PD, 3 D1, 3 D2, 6 T12, 2 T3, and 33 standard categories. This problem is solved based on the diagnostic class.

Based on IEC TC10 data, 9 attributes are missing in the PD data. Information regarding the lost data can be seen in the following table PD data snippet:

TABLE I
PARTIAL DISCHARGE DATA

Inspection	H_2	CH_4	C_2H_4	C_2H_6
Low energy PD and x-wax formation	32930	2397	-	157
X-wax deposition	92600	10200	-	-
PD inducing displacement of insulation and bolt	8266	1061	-	22
X-wax deposits	33046	619	2	58
Heavy x-wax deposits	26788	18342	27	2111

Based on the TDCG method, the data is transformed based on the code in Table 1. The results can be seen in Table 2 below:

TABLE 2
PD DATA CONVERSION

H_2	CH_4	C_2H_2	C_2H_4	C_2H_6
Z1	Z2	W3	W4	Z5
Z1	Z2	X3	W4	Z5
Z1	Z2	W3	W4	W5
Z1	Z2	W3	W4	W5
Z1	Y2	X3	W4	W5
Z1	Z2	Y3	W4	Z5
Z1	Y2	W3	W4	W5
Z1	Z2	W3	W4	Z5
Z1	Z2	W3	W4	Z5

Columns in gray represent missing data. Based on the TDCG conversion, the missing values are categorized based on the lowest class of each attribute. Therefore, all missing values are assigned a temporary feature, i.e., 'W.' To facilitate the search for lost data, each data code is assigned a serial number, as shown in Table 3 below:

TABLE 3
DATA CONVERSION TO SERIAL NUMBER

H_2	CH_4	C_2H_2	C_2H_4	C_2H_6
1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25
26	27	28	29	30
31	32	33	34	35
36	37	38	39	40
41	42	43	44	45

Columns in gray represent missing data. Furthermore, Table 2 is processed to look for rules to determine the relationships of missing data. The following are the search results for rules from PD data based on the Tertius and Tertius classification algorithms.

```
1. /* 0.541697 0.000000 */ C2H6 = Z5 ==> CH4 = Z2
2. /* 0.509592 0.222222 */ CH4 = Z2 ==> C2H6 = Z5
3. /* 0.458299 0.000000 */ C2H6 = W5 ==> CH4 = Y2 or C2H2 = W3
4. /* 0.416375 0.000000 */ C2H2 = X3 and C2H6 = W5 ==> CH4 = Y2
5. /* 0.285714 0.000000 */ C2H2 = Y3 ==> C2H6 = Z5
6. /* 0.254131 0.111111 */ C2H2 = X3 ==> CH4 = Y2
7. /* 0.254131 0.111111 */ CH4 = Y2 ==> C2H2 = X3
8. /* 0.236393 0.444444 */ C2H6 = Z5 ==> C2H2 = Y3
9. /* 0.186429 0.000000 */ C2H2 = Y3 ==> CH4 = Z2
10. /* 0.176214 0.666667 */ CH4 = Z2 ==> C2H2 = Y3
11. /* 0.156438 0.111111 */ C2H6 = W5 ==> C2H2 = W3
12. /* 0.156438 0.111111 */ C2H2 = W3 ==> CH4 = Z2
```

Figure 2. Tertius Classification Algorithm on PD part 1

```
1. /* 0.000000 0.000000 */ TRUE ==> H2 = Z1
2. /* 0.000000 0.111111 */ C2H2 = Y3 ==> FALSE
3. /* 0.000000 0.222222 */ CH4 = Y2 ==> FALSE
4. /* 0.000000 0.222222 */ C2H2 = X3 ==> FALSE
5. /* 0.000000 0.444444 */ C2H6 = W5 ==> FALSE
6. /* 0.000000 0.555556 */ C2H6 = Z5 ==> FALSE
7. /* 0.000000 0.666667 */ C2H2 = W3 ==> FALSE
8. /* 0.000000 0.777778 */ CH4 = Z2 ==> FALSE
```

Figure 3. Tertius Classification Algorithm on PD part 2

Based on this rule, the prediction results of missing PD data are obtained as shown in Table 4 below:

TABLE 4
MISSING VALUE ON PD RESULT FINDINGS

Missing data code	***Value***	***Tertius Rule***	***Tertius classification Rule***	***Predicted value***
3	W3	10,8	0	Y3
4	W4	0	0	0
13	W3	10,8	0	Y3
14	W4	0	0	0
15	W5	2,5	0	Z5
18	W3	10,8	0	Y3
19	W4	0	0	0
33	W3	7,11	0	X3
43	W3	10,8	0	Y3

Missing data code is a code number for missing data based on Table 3, and Column value is a temporary value based on TDCG conversion. The Tertius Rule and Tertius Classification rule are algorithms used to predict the loss value. The number value in the Tertius Rule column and the Tertius classification rule indicates that the value was successfully expected based on the rule to that number from the Tertius algorithm. While 0 shows that the value is not predictable. The predicted value column is a column to determine the expected value based on TDCG.

Based on Table 4, there are 6 predictable data, namely data with codes 3, 13, 15, 18, 33, and 43. Meanwhile, 3 data that cannot be predicted are codes 4, 14, and 19. The 3 data cannot be expected. This is because there is no related rule that explains the data. The missing data is in a lined column condition, so the reason rule cannot be obtained. To deal with this problem, the remaining data that the Tertius algorithm cannot predict will be searched for its value using the mean value approach. The final results of the prediction of missing PD data can be seen in Table 5.

TABLE 5
THE FINAL RESULTS ON MISSING ATTRIBUTE DATA ON PD

Missing data code	***Predicted value***	***Value***
3	Y3	22,5
4	0	8,167
13	Y3	22,5
14	0	8,167
15	Z5	4630

18	Y3	22,5
19	0	8,167
33	X3	5,5
43	Y3	22,5

For the next step is fixing missing attribute value on Discharge of low energy data (D1) and Discharge of High energy data (D2).

Discharge of low energy (D1) is a spark-shaped disturbance that causes the formation of larger holes in the insulating paper or carbon particles in the oil. Based on IEC TC10 data, there are 3 data attributes missing in D1 data.

Discharge of high energy (D2) is a disturbance of power flowing through and causing widespread carbonization of the insulating material, coalescence of iron, and possibly disconnection of equipment. Based on IEC TC10 data, there are 3 data attributes missing in D2 data.

Searching for missing data attributes in D1 and D2 data is similar to the search for missing data in the previous PD. The search results are as follows:

```
 1. /* 0.704167 0.076923 */ C2H4 = W4 ==> H2 = W1
 2. /* 0.599684 0.000000 */ CH4 = Z2 ==> H2 = Z1
 3. /* 0.522075 0.307692 */ CH4 = W2 ==> H2 = X1
 4. /* 0.517697 0.192308 */ CH4 = W2 and C2H2 = Z3 ==> H2 = X1
 5. /* 0.510734 0.153846 */ C2H4 = Z4 ==> H2 = Z1
 6. /* 0.500226 0.307692 */ CH4 = W2 and C2H6 = W5 ==> H2 = W1
 7. /* 0.468775 0.346154 */ CH4 = W2 ==> H2 = W1
 8. /* 0.435560 0.038462 */ C2H2 = Z3 and C2H4 = W4 ==> H2 = W1
 9. /* 0.368421 0.461538 */ C2H6 = W5 ==> H2 = W1
10. /* 0.342507 0.000000 */ CH4 = Y2 and C2H4 = Z4 ==> H2 = Y1
```

Figure 4. Tertius Classification Algorithm on D1

```
 1. /* 0.755737 0.020833 */ CH4 = Z2 ==> H2 = Z1
 2. /* 0.573245 0.270833 */ C2H4 = Z4 ==> H2 = Z1
 3. /* 0.455765 0.020833 */ CH4 = W2 and C2H2 = Z3 ==> H2 = X1
 4. /* 0.432467 0.000000 */ C2H4 = W4 ==> H2 = W1
 5. /* 0.427976 0.020833 */ C2H6 = Z5 ==> H2 = Z1
 6. /* 0.401689 0.250000 */ C2H2 = Z3 and C2H6 = W5 ==> H2 = X1
 7. /* 0.364612 0.041667 */ C2H4 = Y4 ==> H2 = X1
 8. /* 0.364170 0.333333 */ C2H6 = W5 ==> H2 = X1
 9. /* 0.357816 0.000000 */ CH4 = W2 and C2H4 = Y4 ==> H2 = X1
10. /* 0.312604 0.104167 */ CH4 = W2 ==> H2 = X1
```

Figure 5. Tertius Classification Algorithm on D2

TABLE 5
THE FINAL RESULTS ON MISSING ATTRIBUTE DATA ON D1 AND D2

Missing data code	*Predicted value*	*Value*	Category
40	0	81,791	D1
87	0	555,68	D1
90	0	81,791	D1
20	Y5	125,5	D2
40	0	130,57	D2
50	Y5	125,5	D2

Based on the search results with the Tertius algorithm, missing data cannot be predicted with the Tertius algorithm. These data are data with codes 40, 87, and 90. The 3 unpredictable data is caused by the absence of a related rule that explains the data. Meanwhile, data D2, 2 out of 3 missing data, can be predicted using the Tertius algorithm. The data are data with codes numbered 20 and 50.

Meanwhile, 1 data cannot be predicted, namely data with code 40. Data that cannot be predicted using the Tertius algorithm will find its value using the mean value approach to deal with this problem. The final result of the prediction of missing data D1 can be seen in Table 6 above.

A thermal fault below 300ºC (T1) is a disturbance that causes the color of the paper to turn brown. A thermal spot above 300ºC (T2) is a disturbance that can cause carbonized paper. Based on IEC TC10 data, there are 6 data attributes missing in T1 and T2 data.

Thermal faults above 700ºC (T3) are a type of disturbance that can cause carbonized oil and metal to be discolored or even melted. Based on IEC TC10 data, there are 2 data attributes missing in T3 data.

The process of searching for missing data attributes in T1 and T2 data is similar to the search for missing data in the previous D1 and D2. The search results are as follow:

```
 1. /* 0.688194 0.062500 */ C2H4 = W4 ==> H2 = W1
 2. /* 0.653671 0.000000 */ CH4 = W2 ==> H2 = W1
 3. /* 0.594999 0.125000 */ C2H2 = W3 ==> H2 = W1
 4. /* 0.538511 0.062500 */ C2H6 = W5 ==> H2 = W1
 5. /* 0.414729 0.000000 */ CH4 = X2 ==> H2 = Z1
 6. /* 0.377545 0.062500 */ CH4 = Z2 and C2H2 = X3 ==> H2 = Y1
 7. /* 0.374555 0.375000 */ C2H6 = Z5 ==> H2 = Y1
 8. /* 0.343328 0.437500 */ C2H4 = Z4 ==> H2 = Y1
 9. /* 0.315569 0.125000 */ C2H2 = X3 and C2H6 = Z5 ==> H2 = Y1
10. /* 0.305234 0.000000 */ CH4 = Z2 and C2H2 = W3 ==> H2 = X1
11. /* 0.305234 0.000000 */ C2H2 = W3 and C2H4 = Z4 ==> H2 = X1
12. /* 0.305234 0.000000 */ C2H2 = W3 and C2H6 = Z5 ==> H2 = X1
13. /* 0.305234 0.000000 */ CH4 = Y2 and C2H2 = X3 ==> H2 = X1
```

Figure 6. Tertius Classification Algorithm on T1/T2

```
 1. /* 0.688194 0.062500 */ C2H4 = W4 ==> H2 = W1
 2. /* 0.653671 0.000000 */ CH4 = W2 ==> H2 = W1
 3. /* 0.594999 0.125000 */ C2H2 = W3 ==> H2 = W1
 4. /* 0.538511 0.062500 */ C2H6 = W5 ==> H2 = W1
 5. /* 0.414729 0.000000 */ CH4 = X2 ==> H2 = Z1
 6. /* 0.377545 0.062500 */ CH4 = Z2 and C2H2 = X3 ==> H2 = Y1
 7. /* 0.374555 0.375000 */ C2H6 = Z5 ==> H2 = Y1
 8. /* 0.343328 0.437500 */ C2H4 = Z4 ==> H2 = Y1
 9. /* 0.315569 0.125000 */ C2H2 = X3 and C2H6 = Z5 ==> H2 = Y1
10. /* 0.305234 0.000000 */ CH4 = Z2 and C2H2 = W3 ==> H2 = X1
11. /* 0.305234 0.000000 */ C2H2 = W3 and C2H4 = Z4 ==> H2 = X1
12. /* 0.305234 0.000000 */ C2H2 = W3 and C2H6 = Z5 ==> H2 = X1
13. /* 0.305234 0.000000 */ CH4 = Y2 and C2H2 = X3 ==> H2 = X1
```

Figure 7. Tertius Classification Algorithm on T3

TABLE 7
THE FINAL RESULTS ON MISSING ATTRIBUTE DATA ON T1, T2 AND T3

Missing data code	*Predicted value*	*Value*	*Category*
28	W3	50	T1/T2
33	W3	50	T1/T2
38	W3	50	T1/T2
46	Y1	1250,5	T1/T2
68	W3	50	T1/T2
73	Z3	200	T1/T2
3	Z3	200	T3
30	Z5	4630	T3

3	Z3	200	T3

In T1 and T2 data, Missing data can be predicted with the Tertius algorithm. These data are data with 28, 33, 38, 46, 68, and 73. In T3 data, 2 missing data can be predicted by the Tertius algorithm. These data are data with code numbers 23 and 30. The value of the missing data attributes can be seen in table 7 above.

Based on IEC TC10 data, there are 33 missing data attributes in Normal data. Information regarding lost data and the data transformation results can be seen in the appendix. The following are the rule search results from Normal data based on the Tertius and Tertius classification algorithms.

```
1. /* 0.295396 0.000000 */ C2H2 = W3 and C2H4 = W4 and C2H6 = W5 ==> H2 = W1
2. /* 0.295396 0.000000 */ CH4 = W2 and C2H2 = W3 and C2H6 = W5 ==> H2 = W1
3. /* 0.292919 0.040000 */ C2H4 = Z4 ==> H2 = X1
4. /* 0.274924 0.100000 */ C2H4 = W4 and C2H6 = W5 ==> H2 = W1
5. /* 0.273765 0.120000 */ CH4 = W2 and C2H4 = W4 ==> H2 = W1
6. /* 0.256220 0.000000 */ CH4 = W2 and C2H4 = Z4 ==> H2 = X1
7. /* 0.233272 0.140000 */ C2H4 = W4 ==> H2 = W1
8. /* 0.226254 0.020000 */ CH4 = W2 and C2H2 = W3 ==> H2 = W1
9. /* 0.226254 0.020000 */ C2H2 = W3 and C2H6 = W5 ==> H2 = W1
10. /* 0.214529 0.000000 */ C2H2 = X3 and C2H6 = Z5 ==> H2 = W1
11. /* 0.214529 0.000000 */ C2H4 = Z4 and C2H6 = X5 ==> H2 = X1
12. /* 0.198400 0.200000 */ CH4 = W2 and C2H6 = W5 ==> H2 = W1
13. /* 0.181782 0.020000 */ C2H4 = Y4 and C2H6 = Z5 ==> H2 = W1
14. /* 0.181782 0.020000 */ CH4 = X2 and C2H2 = Z3 ==> H2 = X1
15. /* 0.181782 0.020000 */ CH4 = X2 and C2H4 = Z4 ==> H2 = X1
```

Figure 8. Tertius Classification Algorithm on Normal Data

C. Data Normalization

Gas comparison data still has a range of values that are too large. Therefore the data normalization process needs to be done. In this study, data normalization consists of the process of scale equations using the mapminmax function in Matlab.

D. Data Testing

Data testing is divided into 2 parts: data testing using the comparison method and testing data using AIL. Testing is done by dividing the dataset randomly with a portion of 80% Training data and 20% Testing data. The results of testing the data with the comparison method are as follows

1. J48 Method

Data testing using the J48 method was carried out in 30 trials, from these trials obtained the following results graph.

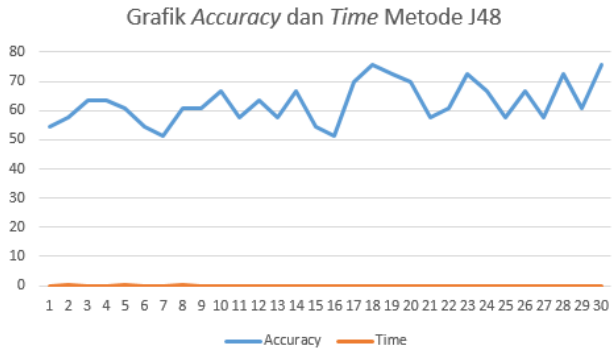

Figure 9. Accuracy and Time Results Chart in J48 Method

The graph in Figure 9 obtained an average measurement result of 62.73%. Whereas the average processing time is 0.0133 seconds

2. Using Random Forest Method

Data testing using the Random Forest method was carried out in 30 trials. From these trials obtained the following results graph.

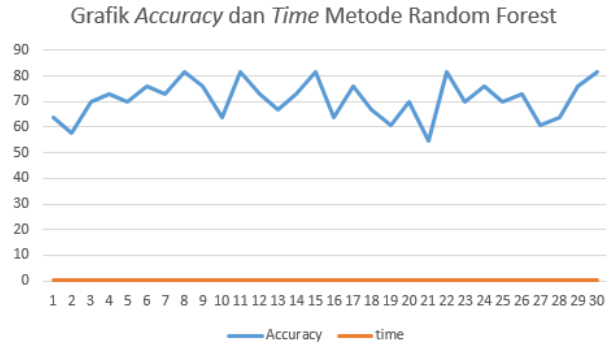

Figure 10. Accuracy and Time Results Chart in Random Forest Method

The graph in Figure 9 obtained an average measurement result of 70.71%. In contrast, the average processing time is 0.0117 seconds.

IV. CONCLUSION

This research was conducted to predict the missing data value attribute in IEC TC10 data which is very important in the DGA analysis process. In this research, the researcher uses the Association rules and the Tertius algorithm to predict the value. The following procedure tests the accuracy percentage using the j48 and Random Forest methods. Searching for missing data using the Tertius algorithm on IEC TC10 data was successfully carried out. Of the total 56 missing data, 36 could be predicted well. The classification accuracy results obtained are 70.71% using Random Forest and 62.73% using J48.

This study has several limitations, and the following are suggestions for further research. It is necessary to explore other missing data search methods because the method used in this study cannot resolve all the lost data in the IEC TC10 dataset. Furthermore, adding the stages of finding the most influential attributes as an effort to get better accuracy results.

REFERENCES

[1] M. H. A. Hamid, M. T. Ishak, M. M. Ariffin, N. I. A. Katim, N. A. M. Amin, and N. Azis, "Dissolved gas analysis (DGA) of vegetable oils under electrical stress," *Int. Conf. High Volt. Eng. Power Syst. ICHVEPS 2017 - Proceeding*, vol. 2017-Janua, pp. 29–34, 2017, doi: 10.1109/ICHVEPS.2017.8225862.

[2] H. Malik, "Extreme Learning Machine Based Fault Diagnosis of Power Transformer Using IEC TC10 and Its Related Data," *Annu. IEEE India Conf.*, pp. 1–5, 2015.

[3] N. Ardi, N. A. Setiawan, and T. Bharata Adji, "Analytical incremental learning for power transformer incipient fault diagnosis based on dissolved gas analysis," *Proc. - 2019 5th Int. Conf. Sci. Technol. ICST 2019*, pp. 3–6, 2019, doi: 10.1109/ICST47872.2019.9166441.

[4] N. A. Setiawan, Sarjiya, and Z. Adhiarga, "Power transformer incipient faults diagnosis using Dissolved Gas Analysis and Rough Set," *Proc. 2012 IEEE Int. Conf. Cond. Monit. Diagnosis, C. 2012*, no. September, pp. 950–953, 2012, doi: 10.1109/CMD.2012.6416311.

[5] A. Pramono, M. Haddin, and D. Nugroho, "Analisis Minyak Transformator Daya Berdasarkan Dissolved Gas Analysis (Dga) Menggunakan Data Mining Dengan Algoritma," *J. Telemat.*, vol. 9, no. 2, pp. 78–91, 2016.

[6] Mukarromah, S. Martha, and Ilhamsyah, "Perbandingan Imputasi Missing Data Menggunakan Metode Mean Dan Metode Algoritma K-Means," *Bul. Ilm. Mat. Stat. dan Ter.*, vol. 04, no. 3, pp. 305–312, 2015.

[7] Y. Benmahamed, Y. Kemari, M. Teguar, and A. Boubakeur, "Diagnosis of Power Transformer Oil Using KNN and Naïve Bayes Classifiers," *2018 IEEE 2nd Int. Conf. Dielectr.*, no. 3, pp. 1–4, 2018.

[8] A. Abu-Siada and S. Islam, "A new approach to identify power transformer criticality and asset management decision based on dissolved gas-in-oil analysis," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 19, no. 3, pp. 1007–1012, 2012, doi: 10.1109/TDEI.2012.6215106.

[9] S. A. I. Alfarozi, N. A. Setiawan, T. B. Adji, K. Woraratpanya, K. Pasupa, and M. Sugimoto, "Analytical incremental learning: Fast constructive learning method for neural network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2016, doi: 10.1007/978-3-319-46672-9_30.

[10] M. Duval and A. DePablo, "Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases," *IEEE Electr. Insul. Mag.*, vol. 17, no. 2, pp. 31–41, 2001, doi: 10.1109/57.917529.

[11] N. Ardi and Isnayanti, "Structural Equation Modelling-Partial Least Square to Determine the Correlation of Factors Affecting Poverty in Indonesian Provinces," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 846, no. 1, pp. 0–13, 2020, doi: 10.1088/1757-899X/846/1/012054.

[12] A. Huda and N. Ardi, "Predictive Analytic on Human Resource Department Data Based on Uncertain Numeric Features Classification," *Int. J. Interact. Mob. Technol.*, vol. 15, no. 8, pp. 172–181, 2021, doi: 10.3991/ijim.v15i08.20907.

[13] E. Society, *IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers*, vol. 2008, no. February. 2009.

[14] J. Kaur and N. Madan, "Association Rule Mining: A Survey," *Int. J. Hybrid Inf. Technol.*, vol. 8, no. 7, pp. 239–242, 2015, doi: 10.14257/ijhit.2015.8.7.22.

[15] P. A. Flach and N. Lachiche, "Confirmation-guided discovery of first-order rules with Tertius," *Mach. Learn.*, vol. 42, no. 1–2, pp. 61–95, 2001, doi: 10.1023/A:1007656703224.

[16] J. Nahar, K. S. Tickle, A. B. M. S. Ali, and Y. P. P. Chen, "Significant cancer prevention factor extraction: An association rule discovery approach," *J. Med. Syst.*, vol. 35, no. 3, pp. 353–367, 2011, doi: 10.1007/s10916-009-9372-8.

[17] S. A. Kumar and V. M.N, "Discerning Learner's Erudition Using Data Mining Techniques," *Int. J. Integr. Technol. Educ.*, vol. 2, no. 1, pp. 9–14, 2013, doi: 10.5121/ijite.2013.2102.

[18] M. Supriyamenon and P. Rajarajeswari, "A review on association rule mining techniques with respect to their privacy preserving capabilities," *Int. J. Appl. Eng. Res.*, vol. 12, no. 24, pp. 15484–15488, 2017.