IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received April 26, 2021, accepted May 16, 2021, date of publication May 19, 2021, date of current version May 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3081699

Dealing With Data Uncertainty for Transformer Insulation System Health Index

RAHMAN AZIS PRASOJO [1,3], (Graduate Student Member, IEEE), SUWARNO [1], (Senior Member, IEEE), AND A. ABU-SIADA [2], (Senior Member, IEEE)

[1]School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung 40132, Indonesia
[2]Discipline of Electrical and Computer Engineering, Curtin University, Perth, WA 6102, Australia
[3]Department of Electrical Engineering, Politeknik Negeri Malang, Malang 65141, Indonesia

Corresponding author: Rahman Azis Prasojo (rahmanazisp@polinema.ac.id)

This work was supported by P2MI Institut Teknologi Bandung under Grant 916/IT1.C12/KP/2021.

ABSTRACT Health index has been widely accepted as a powerful tool for monitoring the condition of power transformer insulation system based on various diagnostic parameters. While this approach has been extensively discussed in the literature, not much attention was given to provide effective solutions to the uncertainty in the used data. According to CIGRE 761, data quality issues may arise due to measurement accuracy as well as incompleteness and unavailability of the required data. Therefore, this article presents the implementation and evaluation of a certainty level model for transformer insulation system health index to deal with data uncertainty. The impact of data unavailability on the health index results is also investigated. Certainty level of the health index is determined by the criticality level of available data, and is reported along with the health index result. A method to handle unavailable data by predicting the oil interfacial tension (IFT) using Random Forest approach is also presented. The proposed certainty level model is designed to accommodate the predicted value of missed data into the health index model while considering its prediction accuracy. The robustness of the developed model is validated through its application in assessing the health condition of six in-service power transformers. The results indicate that by including the proposed certainty level and the prediction approach to eliminate the issue of uncertain and missed diagnostic data, an asset management decision can be taken on operating power transformer fleets with high level of confidence.

INDEX TERMS Power transformers, condition monitoring, health index, data uncertainty, asset management.

I. INTRODUCTION

Power transformers are costly and vital equipment in power system electricity grids. To ensure a power system operates safely and reliably, proper condition monitoring and evaluation schemes for network assets is necessary [1]. This can be done through measuring specific diagnostic parameters for each asset in regular basis to assess the overall health condition of the asset by integrating all diagnostic parameters into one index, which is often called health index. This index can be used to sort the transformers according to their failure probabilities.

CIGRE (International Council on Large Electric Systems) published a Technical Brochure 761 on condition assessment of power transformers [2]. One of the limitations of scoring and indexing reported in this brochure is the quality of the used data. Sources of data quality issues include measurement accuracy as well as incompleteness, and missed data. It is important that the impact of unavailable data on the overall health index to be understood as this may result in false diagnoses.

Generally, there are two main approaches to determine the health index (HI) of power transformers: the conventional HI method based on scoring and weighting, and the non-conventional HI method based on artificial intelligence and machine learning. Scoring and weighting is still the most widely used and straightforward HI approach that has been presented in several studies in the literature [3]–[10]. The conventional method of health indexing based on scoring and weighting relies on personnel expertise. Each HI input

The associate editor coordinating the review of this manuscript and approving it for publication was Gongbo Zhou.

parameter is scored and weighted to provide an overall output index reflecting the overall health condition of the investigated transformer.

A few studies can be found in the literature in regard to uncertainty in the used data. For instance, studies in [11], [12] proposed a Bayesian belief network HI that is able to capture the uncertainties of used data by showing the percentage of each transformer, judged as a certain category. Study in [13] proposed a health estimation approach using Markov chains and evidential reasoning. This model was aimed to overcome problems such as uncertainty, accuracy and confidence in the inspected results. Study in [14] developed a soft computing and probabilistic HI and provided confidence intervals for decision-making under data uncertainty. While these approaches are interesting and going to the right direction, the implementation is not as straightforward as scoring and weighting method currently used by utilities worldwide. Therefore, practical implementation of such techniques is not yet to be seen.

On the other hand, some previous studies presented various prediction approaches for unavailable transformer assessment data. Study in [15] proposed an ensemble classifier to predict oil interfacial tension (IFT). Studies in [16], [17] proposed an adaptive neuro-fuzzy inference system (ANFIS) model to predict paper condition using more widely available oil assessment data. Study in [18] proposed a k-nearest neighbors predictive model to estimate furan level using oil breakdown voltage, acidity, water content, and dissolved gases. Study in [19] employed an artificial neural network to predict several oil parameters, such as IFT, acidity, and breakdown voltage. Study in [20] proposed a support vector machine model to classify power transformer paper condition using oil dielectric characteristics and dissolved gases. Study in [21] presented an exponential relationship between IFT and acidity of the transformer oil in which one measured parameter can be used to estimate the other parameter, if not available.

This article aims to deal with data unavailability in power transformer conventional HI. The proposed certainty levels are reported along with the HI results of the investigated transformers. In addition, the impact of unavailable data on the HI is evaluated, and a guideline to interpret the HI and certainty level is proposed.

II. UNCERTAINTY ON HEALTH INDEX

HI model is a useful diagnostic tool, if developed properly. However, handling too much data required by the model is a challenging task. According to [22], uncertainty of information in the HI model is due to three reasons as follows.

A. DATA ACCURACY

Issue of data accuracy can be raised due to incorrect data entry, or inaccurate measurements. Therefore, validation of the used data is essential before using them in the HI model. All entry and typo errors need to be carefully reviewed and corrected. Questionable measurements need to be retested and remeasured whenever possible.

The use of several measurements that represent the same failure mode may have different sensitivity. This issue can be resolved by using proper weighting factors i.e., more critical and reliable parameter gets a higher weighting.

B. DATA COMPLETENESS

One of the ways for the HI model to handle missed data is by leaving out the unavailable parameters then recalibrating the results which is the most common and straightforward approach. When using this approach, it is necessary to report the level of uncertainty along with the HI value, which is not given much attention in the current literature.

C. DATA TIMELINESS

Ideally, HI calculation should be conducted right after the measurements of the used parameters. However, several transformers do not have complete data at each point of measurements. This issue can be solved with the use of measurement history as will be elaborated below.

III. METHODOLOGY

This section presents the proposed methodology to handle uncertainty in the parameters used by the HI model.

A. MULTI EXPERT PARAMETER PRIORITIZATION

The use of analytic hierarchy process (AHP) to capture the knowledge and experience of experts in power transformer assessment has been reported in [23]–[25]. A multi-expert parameter prioritization using AHP has been developed in [23], [26]. In this study, five experts with vast experience in transformer condition monitoring and diagnostics have taken parts to fill in a designed questioners survey, which has been reported in [26]. The survey is meant to compare the parameters to each other through two main criteria: measurement reliability (MR) and criticality (CR). The MR assesses the reliability of the results due to each parameter, while the CR evaluates the criticality of each parameter toward each factor. Each expert was asked to fill in the comparison matrix (C) given by (1) with balanced scales of pairwise comparison as proposed in [27].

$$C = \begin{vmatrix} C_{11}=1 & C_{12}=1/C_{21} & C_{13}=1/C_{31} & \ldots & C_{1n} \\ C_{21} & C_{22}=1 & C_{23}=1/C_{32} & \ldots & C_{2n} \\ C_{31} & C_{32} & C_{32}=1 & \ldots & C_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ C_{n1} & C_{n2} & C_{n2} & \ldots & C_{nn}=1 \end{vmatrix} \tag{1}$$

$$w_i = \frac{\sqrt[n]{\prod_{m=1}^{n} C_{in}}}{\sum_{i=1}^{n}\left(\sqrt[n]{\prod_{m=1}^{n} C_{in}}\right)} \tag{2}$$

The comparison matrix in (1) is of *nxn* dimension in which each element (C_{ij}) represents the value of each used parameter compared with each other. The main diagonal values are always unity, as each of these elements is compared with itself. Each of the lower and higher triangle value is mutually reversed [23]. The calculation of the experts' answers into weighting factors (w_i) was carried out by AHP using (2). To aggregate the results of multiple experts, consensus analysis was employed. The detailed process of weighting factor determination is explained in [23] and [26].

The results of these processes are the weightages of the HI model used in this paper as shown in Table 1. The table comprises various diagnostic parameters to assess the oil quality, paper condition, and any possible faults. Weighting factors have been given to each parameter (Wp) and each factor (W_f) based on utility personnel expertise.

TABLE 1. Weighting values of the HI model used in this paper.

Factor	Parameters	*Wp*	*Wf*
Oil Quality	Breakdown Voltage (BDV)	0.169	0.216
	Water Content (WATER)	0.108	
	Acidity (ACID)	0.139	
	Interfacial Tension (IFT)	0.124	
	Colour Scale	0.114	
Faults	Dissolved Gas Analysis	1	0.348
Paper Condition	CO&CO2	0.171	0.436
	2FAL	0.234	
	Operating Age	0.214	

Besides using the weighting factor for calculating the HI, this paper proposes the calculation of the certainty level as discussed below.

B. CALCULATION OF CERTAINTY LEVEL

A parameter prioritization analysis has been conducted and the results are listed in Table 1. The Wp values are used to calculate the certainty level (CL) according to data availability based on the model presented in [28].

The more considered and complete parameters, the higher the CL and the more confident decisions can be made. By expressing the proposed certainty level, in addition to the HI value, better asset management scheme can be developed to determine an appropriate action based on the health condition of the power transformer [28].

The CL_i for each parameter (i) in Table 1 is calculated using the corresponding Wp_i and Wf_i. After that, the overall CL can be calculated as per the below equations with results listed in Table 2.

$$CL_i = Wp_i \times Wf_i \tag{3}$$

$$CL = \frac{sum\,(Available\ CL_i)}{Max\ CL} \times 100 \tag{4}$$

The same approach can be replicated using different weighting factors, or even different HI structures. Whenever there is unavailable information, the HI can still be calculated,

TABLE 2. Contribution of each parameter to the overall certainty level.

Parameters	CL_i
BDV	0.0553
Water	0.0346
Acid	0.0460
IFT	0.0413
Colour	0.0387
DGA	0.3480
$CO\&CO_2$	0.0746
2FAL	0.1020
AGE	0.0933

and recalibrated. However, the certainty level needs to be reported along with the HI result as proposed in this paper.

C. HEALTH INDEX CALCULATION

The use of incomplete data for HI calculation can still be done by reporting its certainty level alongside with the HI result. The impact of data unavailability is evaluated in this paper as per the proposed methodology shown in Fig. 1 that is developed based on the assessment of 157 Indonesian power transformers of 150 kV rated voltage. The 157 power transformers encompass complete data as required by the HI. The HI is first calculated for each transformer with all complete data, resulting in HI0. Then, the dataset for each transformer is adjusted to 50 combinations of various scenarios of unavailable data as shown in Appendix 1. For each set of these scenarios, the HI is calculated. Hence, for each transformer, 51 HI values are obtained (HI0 for complete data and HI1 to HI50 for various missed data). The impact of data unavailability is then investigated by comparing the HI1-HI50 to HI0 as the baseline.

One way to calculate the HI when data is incomplete is by ignoring unavailable data then recalibrating the obtained HI. A simplified example of how this approach can be conducted is presented in [22]. For instance, four failure modes, FM_1 to FM_4, are assessed as input parameters to the HI model. In this simplified example, the aggregation of HI is a simple sum of each FM score, as shown in (5). Maximum score for each FM is 25, and hence the maximum HI is 100.

The incomplete data in this example is simulated with unavailable FM_4. With incomplete data, the HI can still be calculated through a recalibration step using (6).

- $FM_1 = 25$ out of 25
- $FM_2 = 20$ out of 25
- $FM_3 = 15$ out of 25
- $FM_4 =$ Unavailable out of 25

$$HI_{max} = (25 + 25 + 25 + 25) = 100 \tag{5}$$

$$HI_{with\ unavailable\ FM4} = (25 + 20 + 15) * 100/75 = 80 \tag{6}$$

This calculation ignores the unavailable data, but retains the maximum value to be 100 by recalibration.

This paper adopts the HI approach using the structure proposed in [29]. Each parameter is assigned a score S_i, then

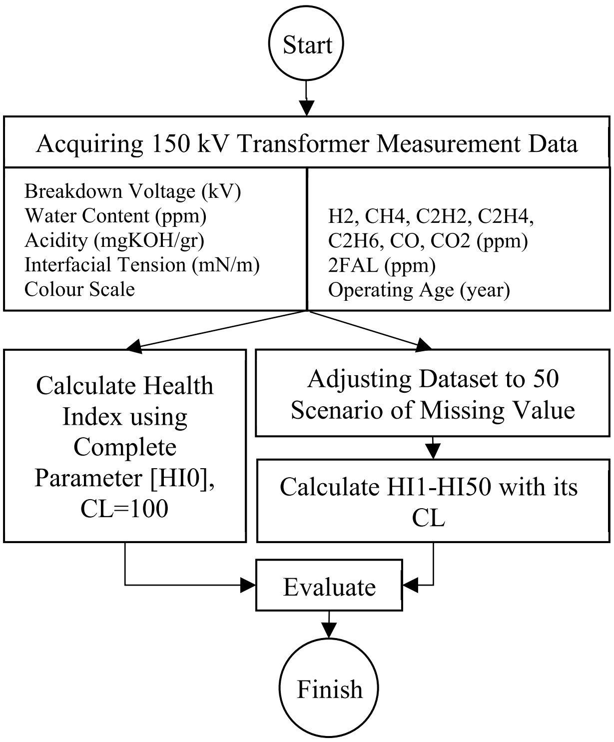

FIGURE 1. Flowchart of the proposed methodology.

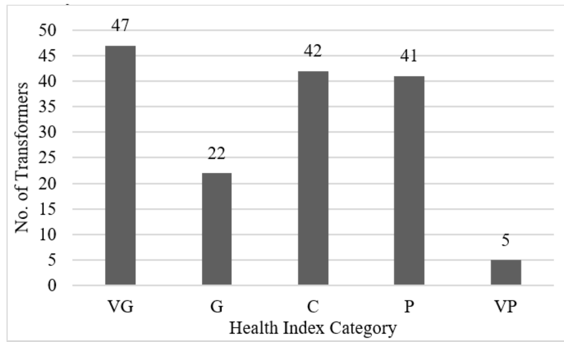

FIGURE 2. Health Index (HI0) results of 157 power transformers using complete data.

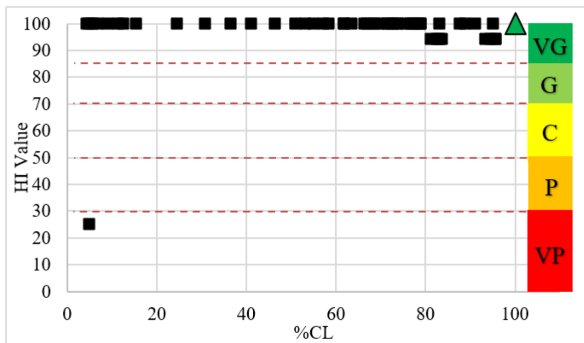

FIGURE 3. Discrepancy of the Health Index results from various data availability scenarios for transformer A.

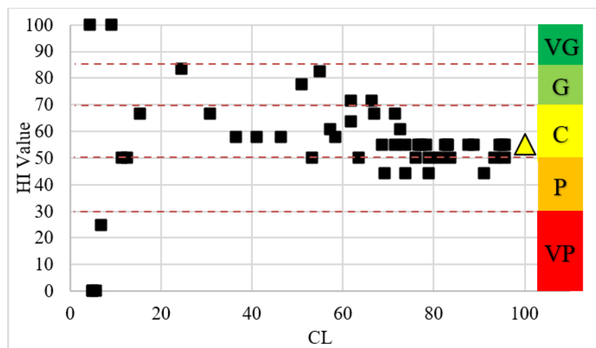

FIGURE 4. Discrepancy of the Health Index results from various data availability scenarios for transformer B.

weighted using parameter weighting factor Wp_i. The HI of each factor is then calculated using (7) which is scored to obtain SF_j. The final HI is then calculated by aggregating the SF_j and Wf_j values using (8). The maximum HI value (100) indicates very good condition, while 0 HI reveals very poor condition. More details about HI calculation can found in [29].

$$HI_{each\ factor} = \frac{\sum_{i=1}^{n} S_i Wp_i}{\sum_{1}^{n} W_i} \tag{7}$$

$$HI_{final} = \frac{\sum_{j=1}^{n} SF_j Wf_j}{\sum_{j=1}^{n} 4W_j} \times 100\% \tag{8}$$

IV. RESULTS AND DISCUSSIONS

Fig. 2 shows the HI of the investigated 157 Power Transformers with complete data (HI0). The results show that 47 transformers are in a very good condition, 22 in a good condition, 42 are in a concerning condition and need attention while 41 transformers are in a poor condition. Only 5 transformers are found in a very poor condition.

A. IMPACT OF MISSED DATA

HI1 to HI50 are calculated based on the assumed missed data scenarios and are compared with the HI0 that is calculated with the availability of complete data. In this way, the impact of data unavailability to the HI calculation can be evaluated.

Based on this analysis, Transformer A is in a very good condition for all assumed scenarios as shown in Fig. 3. In this case, data availability is not really showing observable impact on the HI results. This is because when the transformer is in satisfactory condition, most of the parameters are in a similar healthy state.

Fig. 4 shows that transformer B has a HI0 corresponds to a concerning condition. The discrepancy in HI results is more obvious at CL of about 70% and even more visible at lower CL values. A similar pattern is observed for transformer C as shown in Fig. 5. This transformer is in poor condition when its HI is calculated with complete data (HI0). When less data are used, the HI discrepancy is more visible. Moreover, HI calculation shows that transformer C is in good condition when the CL is around 61%. These results reveal the false

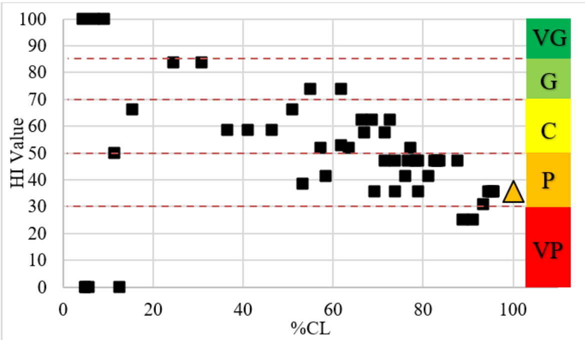

FIGURE 5. Discrepancy of the Health Index results from various data availability scenarios for transformer C.

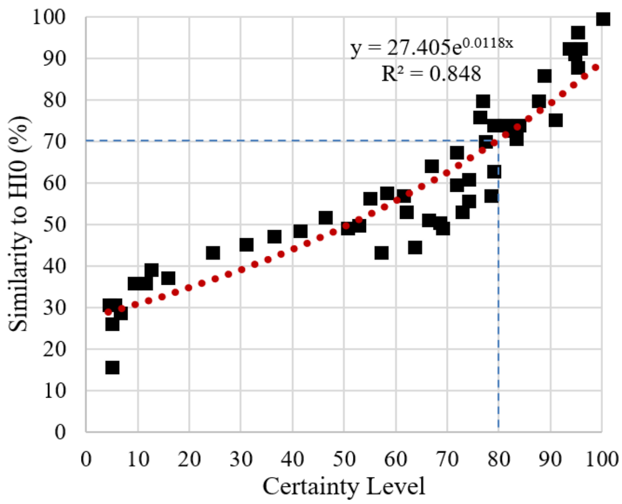

FIGURE 6. Scatterplot of certainty level for Similarity cases of HI1-HI50 with HI0.

health state that may be reported by the HI when some data are missed. One way to overcome this issue is by considering the CL along with the HI.

Fig. 6 shows a scatterplot of certainty level and each similarity to HI0 for all studied HI1-HI50 scenarios. These results reveal the impact of unavailable data on the HI results. The similarity is decreasing when the certainty level is decreasing. An exponential regression line can be obtained with a high coefficient correlation of $R^2 = 0.848$ as shown in Fig. 6. In this study, a similarity percentage of 70% is considered as an acceptable level, then a CL of 80% or higher is necessary to obtain good agreement between HI0 and the HI calculated with missed data. Calculating the HI with incomplete data and with a CL less than 80% is considered inaccurate and calls for the necessity of providing some of these missed data.

B. DATA TIMELINESS IN HEALTH INDEX CALCULATION

Another problem arising on power transformer HI calculation is the data timeliness. Power transformers are frequently observed once or twice a year, and more frequently for critical transformers. However, not all measurements are being available at each inspection time [2], [22]. This can be an issue for utility engineers to use the data for HI calculation.

One of the most common ways to assess the transformer with unavailable recent data is to use historical data. Fig. 7 and 8 show examples of the use of historical data to estimate the transformer current state. However, this may lead to another uncertainty of whether the state of the parameter is still relevant to the transformer current condition or not.

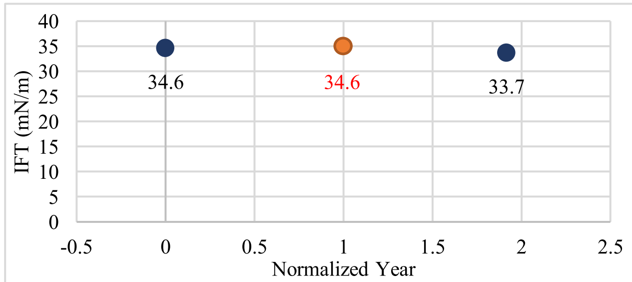

FIGURE 7. Using the previous interfacial tension measurement for unavailable current data.

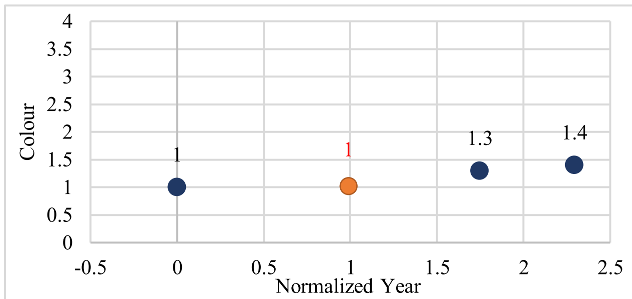

FIGURE 8. Using the previous colour scale observation for current unavailable data.

In this paper, all of the used historical data to assess the transformer current HI were measured in the same year. Therefore, the uncertainty due to data timeliness is disregarded. However, for more accurate analysis about data timeliness, further investigation is required, which is out of the scope of this paper.

C. ANALYSIS OF THE PROPOSED MODEL

After assessing the investigated transformers HI and reporting the CL as suggested, analysis of the obtained results is conducted based on the proposed guideline as shown in Table 3. This guideline is explained through six case studies as listed in Table 4 with the corresponding HI and CL values for each transformer.

In Table 4, transformer 1 (TRF1) shows an HI in the category of very good condition. There are missing values for 2FAL and IFT. The calculated CL with such missing data is 83% which is moderate. According to Table 3, no action is necessary for this transformer. TRF 2 shows similar results but has more complete measurements and higher CL.

The HI assessment of TRF3 shows a good condition. The parameters missing are 2FAL, IFT, and oil colour.

TABLE 3. Guideline for the simultaneous use of HI and CL to assess transformer condition.

		Health Index				
		VG	G	C	P	VP
CL (%)	High (90-100)	No action	No action	Urgent action	Immediate Action	Immediate Action
	Moderate (80-90)	No action	No action	Urgent testing	Urgent action	Immediate Action
	Low (50-80)	No action	More testing	More testing	Urgent Testing	Urgent action
	Very Low (0-50)	More testing	More testing	More testing	More testing	Urgent testing

TABLE 4. Case study of six 150 kV power transformers.

Transformer	TRF1	TRF2	TRF3	TRF4	TRF5	TRF6
Breakdown Voltage (kV)	82	75.9	81.9	100.1	68	63.3
Water (ppm)	9.2	3.5	12.0	3.8	19.3	16.5
Acidity (mgKOH/g)	0.003	0.01	0.01	0.18	0.247	0.02
Interfacial Tension (mN/m)	*NA*	34.6	*NA*	14.6	*NA*	*NA*
Colour	0.5	0.5	*NA*	2.5	*NA*	3.5
H_2 (ppm)	0	3	4.3	0	19	3
CH_4 (ppm)	0	12	66.5	19	157	9
C_2H_2 (ppm)	0	0	0	0	0	0
C_2H_4 (ppm)	0	8	8.1	0	27	186
C_2H_6 (ppm)	0	11	342.1	0	584	11
CO (ppm)	0	957	55	2253	428	987
CO_2 (ppm)	105	2533	1453	1772	2115	6471
2FAL (ppm)	*NA*	*NA*	*NA*	330	*NA*	*NA*
Age	7	5	14	20	21	23
Duval Pentagon Method	x	x	S	O	S	T3H
Oil Quality Factor	A	A	A	D	D	C
Faults Factor	A	A	C	A	C	E
Paper Condition Factor	A	A	A	C	C	D
Health Index	100	100	82.6	62	44.6	21.7
HI Category	**VG**	**VG**	**G**	**C**	**P**	**VP**
CL	83	88	78	100	78	83
CL Category	Mode-rate	Mode-rate	Low	High	Low	Moderate
Required Action	No Action	No Action	More Testing	Urgent Action	Urgent Testing	Immediate Action

NA = Not Available; x = Not Applicable.

The calculated CL (78%) is lower than the acceptable threshold limit. High concentration of C_2H_6 is detected, revealing a thermal fault in the oil or paper [30]. TRF3 has been in service for 14 years only. Based on Table 3, the recommended action for this transformer is to conduct more testing due to the uncertainty of the obtained analysis.

TRF4 results in a concerning HI condition with complete diagnostic parameters. This transformer has been in operation for 20 years. As the parameters are complete, the CL is 100%. The oil quality factor classifies this transformer in poor condition due to high acidity, low IFT, and pretty dark oil colour [31]. The recommendation for this transformer is to provide urgent action such as taking the transformer out of service to rectify this issue before the health condition becomes worse.

TRF5 is in poor HI condition, with the oil quality factor in poor condition (D), while faults factor and paper condition factor need caution (C). The available data as shown in Table 4 show that the acidity is very high. High acidity in this transformer is an indication of oil oxidation that degrades the insulation paper through hydrolysis process [32]. Due to several unavailable data such as 2FAL, IFT, and oil colour, the CL is low. It is recommended that TRF5 needs urgent testing to confirm its current health condition.

Based on the HI of TRF6, its condition is considered very poor. The faults factor of this transformer is critical, and the paper condition is poor. The oil colour is moderately dark with high concentrations of CO and CO_2, which is an indication for solid insulation degradation [31], [33]. The dissolved gas analysis (DGA) interpretation using Duval Pentagon indicated a thermal fault in oil above 700 ^{0}C. Even though the 2FAL and IFT parameters are unavailable for this transformer, the CL is 83% and the recommendation is to take an immediate remedial action.

Results above indicate that complete data availability results in more reliable HI. In case of missing data with a CL of more than 80%, the HI will be close to the HI calculated with complete parameters. In case of CL less than 80%, some of the missed data should be measured or estimated as elaborated in the prediction of the IFT below.

D. IFT PREDICTION MODEL

IFT is a measure of the inter-surface strength between oil and water. IFT can be used as a detection tool for oil-soluble sludge. During the early stages of oil aging, IFT changes rapidly but this change usually stabilizes once a moderate level of damage is achieved [34]. Because IFT is often unavailable, a prediction approach using available measured parameters for most of the Indonesian transformer fleets is useful.

The proposed prediction model is developed using 357 transformers collected data, consisting of twelve input features: water content (Water), voltage breakdown (VBD), colour scale, acidity, carbon monoxide (CO), carbon dioxide (CO_2), hydrogen (H_2), methane (CH_4), acetylene (C_2H_2), ethylene (C_2H_4), ethane (C_2H_6), and CO+CO_2. These features are shown in Table 5, and are used to predict the output oil IFT value. The model is developed by training five different algorithms, namely support vector machine (SVM), decision tree (DT), random forest (RF), linear regression (LR), and artificial neural network (ANN) with the collected data along with the actual IFT value. A total of 18 feature

TABLE 5. List of used features.

Number	1	2	3	4	5	6
Feature	Water	VBD	Color	Acidity	CO	CO_2
Number	7	8	9	10	11	12
Feature	H_2	CH_4	C_2H_2	C_2H_4	C_2H_6	CO+CO_2

TABLE 6. Eighteen combinations of features.

Combination	Features	Combination	Features
1	1,2,3,4,5,6,7,8,9,10,11	10	3,4
2	1,2,3,4,7,8,9,10,11,12	11	4,12
3	1,2,3,4,11,12	12	1,4
4	1,2,3,4,5,6	13	2,4
5	1,2,3,4,12	14	4,5
6	1,2,3,4,5	15	4,6
7	1,2,3,4	16	1,2
8	2,3,12	17	1,4
9	1,2,4	18	2,4

combinations as shown in Table 6 are evaluated, and the optimal model with the highest prediction accuracy is selected.

In this analysis, the used validation method is the 5-fold cross validation. The procedure of the 5-fold cross validation is conducted by dividing the dataset of N samples into five groups of $N/5$ size each. The algorithm is trained using four groups and tested using one group, then performed repeatedly for five times. The mean accuracy is used as the accuracy of the algorithm.

The performance of the proposed model is assessed using symmetric mean absolute percentage error (SMAPE) to measure error percentage, then the percentage accuracy is calculated. Fig. 9 shows the accuracy of various IFT prediction models based on various feature combinations in Table 6. The results show that the median of the random forest model's accuracy in predicting the IFT value is the highest, which is 88.8%.

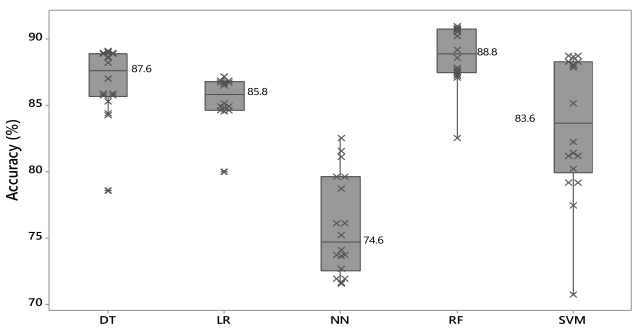

FIGURE 9. Boxplot accuracy of IFT prediction models.

Table 7 shows the percentage accuracy of various IFT prediction models. The maximum accuracy obtained is 90.91% using RF for combination 2, where ten input features were used. The proposed RF model includes 100 number of decision trees. Combinations 8 and 10 show relatively good accuracy with limited measurements. Combination 8 produces an accuracy of 90.66%, which used colour, acidity, and CO + CO_2. Removing CO + CO_2 resulted in 89.14% accuracy.

In [15], a classification accuracy of 87% was achieved by predicting the oil IFT using the soft voting classifier. However, the IFT considered class in this paper was only bad and good classes. Ref. [19] predicted the IFT value with an accuracy of 95%. Ref. [35] also presented an IFT prediction model based on oil spectral response. However, these models require another testing of transformer oil that may not be carried out in Indonesia. The prediction approach presented in this paper uses available transformer data in Indonesia without the need of additional testing.

TABLE 7. Accuracy percentage of various IFT prediction models.

Combination No.	SVM	DT	RF	LR	NN
1	87.81	89.04	90.79	87.15	81.54
2	87.99	89.01	90.91	87.16	81.08
3	88.26	88.82	90.72	86.82	79.62
4	88.59	88.62	90.50	86.45	78.68
5	88.26	88.82	90.78	86.82	79.62
6	88.67	88.93	90.67	86.67	76.07
7	87.97	88.52	90.21	86.70	75.19
8	88.67	88.93	90.66	86.67	76.07
9	85.13	86.96	88.54	84.83	73.62
10	80.20	88.18	89.14	86.60	74.11
11	82.17	85.26	87.31	84.55	71.57
12	79.13	85.87	87.54	84.91	71.96
13	81.14	85.73	87.66	84.62	73.68
14	81.37	84.21	87.03	84.54	72.70
15	70.75	84.34	87.20	85.13	82.52
16	77.46	78.55	82.53	80.01	71.64
17	79.13	85.87	87.44	84.91	71.96
18	81.14	85.73	87.80	84.62	73.68
Average	83.55	86.74	88.75	85.51	75.85

Overall, the random forest-based model outperforms other algorithms and the combinations of features result in high prediction accuracy. Random forest (RF) is a classification and regression method, comprising a combination of multiple decision trees where each tree is generated using a random vector that is sampled independently of the input vector. In regression, tree predictors use numerical values instead of the class labels used by the RF classifier [36]. Several studies have also presented the implementation of RF in the power transformer assessment. In [37] an RF technique for transformer failure discrimination is presented and achieved an accuracy of 98%. The study in [18] compared several machine learning methods to predict furan levels in transformer oil, where the RF model produced a good accuracy. Research in [38], [39] developed a partial discharge recognition model using a variety of machine learning methods, of which RF achieved the highest accuracy.

Another advantage of the RF is that it is easy to measure the relative importance of each feature to the target. Fig. 10 indicates that the feature with the highest relative importance is the oil colour, followed by acidity, then CO and CO_2, and other measurements. The high feature ranking of colour, acidity, CO, and CO_2 for IFT prediction agrees well with several previous studies. In [15], it was found that IFT has a high correlation with acidity, dissipation factor, and colour. An increase in oil acidity followed by a decrease in IFT is shown in [21]. The colour of the transformer oil changes from clear light (for new oil) to dark (for used oil) [19]. IFT is very sensitive to the presence of degradation products dissolved in oil. Meanwhile CO and CO_2 are used as indicators for paper degradation.

```
Feature ranking:
1. Colour (0.350500)
2. Acidity (0.263610)
3. CO (0.078463)
4. CO2 (0.075356)
5. VBD (0.060414)
6. CO+CO2 (0.050506)
7. C2H6 (0.045206)
8. Water (0.026256)
9. CH4 (0.019804)
10. H2 (0.012299)
11. C2H4 (0.012186)
12. C2H2 (0.005401)
```

FIGURE 10. Rank of features based on relative importance using Random Forest.

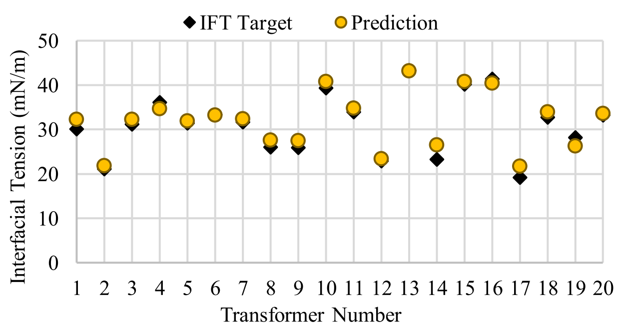

FIGURE 11. Samples of IFT prediction using Random Forest model.

Fig. 11 shows 20 IFT prediction results using the proposed RF model based on the measured oil colour and acidity parameters. Appendix 2 illustrates twenty samples of transformers data in which oil IFT value is estimated using the proposed Random Forest model. The results show the high accuracy of the predicted value compared with the actual measured value (target).

Table 8 shows a case study on Transformers 1 and 6, in which the IFT measurement is not available and is estimated using the developed RF model. Available measurements for oil colour and acidity are used to predict the value of oil IFT.

TABLE 8. Implementation of the IFT prediction model for certainty level.

Transformer	TRF1	TRF6
Acidity (mgKOH/g)	0.003	0.02
IFT Measured (mN/m)	*NA*	*NA*
IFT Value (Predicted)	41.61	29.53
IFT Score (Predicted)	1	2
Colour	0.5	3.5
Age	7	23
Oil Quality Factor	A	C
Faults Factor	A	E
Paper Condition Factor	A	D
Health Index	100	21.7
HI Category	**VG**	**VP**
CL	82→**87**	82→**87**
CL Category	Moderate	Moderate
Required Action	No Action	Immediate Action

TABLE 9. Fifty combinations of data unavailability.

	DGA	Water	VBD	Acidity	IFT	Colour	CO&CO$_2$	2FAL	Age	CL
HI0	1	1	1	1	1	1	1	1	1	100
HI1	1	0	1	1	1	1	1	1	1	96
HI2	1	1	0	1	1	1	1	1	1	93
HI3	1	1	1	0	1	1	1	1	1	94
HI4	1	1	1	1	0	1	1	1	1	95
HI5	1	1	1	1	1	0	1	1	1	95
HI6	1	1	1	1	1	1	0	1	1	91
HI7	1	1	1	1	1	1	1	0	1	88
HI8	1	1	1	1	1	1	1	1	0	89
HI9	0	1	1	1	1	1	1	1	1	58
HI10	1	0	1	1	1	1	1	0	1	84
HI11	1	1	0	1	1	1	1	0	1	81
HI12	1	1	1	0	1	1	1	0	1	82
HI13	1	1	1	1	0	1	1	0	1	83
HI14	1	1	1	1	1	0	1	0	1	83
HI15	1	1	1	1	1	1	0	0	1	79
HI16	0	1	1	1	1	1	1	0	1	46
HI17	1	1	1	1	1	1	1	0	0	77
HI18	1	0	1	1	0	1	1	0	1	79
HI19	1	1	0	1	0	1	1	0	1	76
HI20	1	1	1	0	0	1	1	0	1	77
HI21	1	1	1	1	0	0	1	0	1	78
HI22	1	1	1	1	0	1	0	0	1	74
HI23	1	1	1	1	0	1	1	0	0	72
HI24	0	1	1	1	0	1	1	0	1	41
HI25	1	0	1	1	0	0	1	0	1	74
HI26	1	1	0	1	0	0	1	0	1	72
HI27	1	1	1	0	0	0	1	0	1	73
HI28	1	1	1	1	0	0	0	0	1	69
HI29	1	1	1	1	0	0	1	0	0	67
HI30	0	1	1	1	0	0	1	0	1	36
HI31	1	0	1	0	0	0	1	0	1	68
HI32	1	1	0	0	0	0	1	0	1	66
HI33	1	1	1	0	0	0	0	0	1	64
HI34	1	1	1	0	0	0	1	0	0	62
HI35	0	1	1	0	0	0	1	0	1	31
HI36	1	0	0	0	0	0	1	0	1	62
HI37	1	1	0	0	0	0	0	0	1	57
HI38	1	1	0	0	0	0	1	0	0	55
HI39	0	1	0	0	0	0	1	0	1	24
HI40	1	0	0	0	0	0	0	0	1	53
HI41	1	0	0	0	0	0	1	0	0	51
HI42	0	1	0	0	0	0	0	0	1	15
HI43	0	1	0	0	0	0	0	0	0	4
HI44	0	0	1	0	0	0	0	0	0	7
HI45	0	0	0	1	0	0	0	0	0	6
HI46	0	0	0	0	1	0	0	0	0	5
HI47	0	0	0	0	0	1	0	0	0	5
HI48	0	0	0	0	0	0	1	0	0	9
HI49	0	0	0	0	0	0	0	1	0	12
HI50	0	0	0	0	0	0	0	0	1	11

1= Available data; 0 = Unavailable data

For TRF1, the predicted IFT is 41.61 mN/m, with a score of 1. For TRF6, the acidity is relatively higher, and the oil colour is darker. The prediction model produces an IFT value of 29.53 mN/m, with a score of 2. By performing IFT prediction, the CL of the two transformers is increased from 82 to 87. However, direct measurement for IFT is recommended to increase the confidence level of the resulting transformer health index.

TABLE 10. Twenty samples of transformers data for the proposed IFT prediction using Random Forest model.

No	Water (ppm)	VBD (kV)	Colour	Acidity (mgKOH/g)	CO (ppm)	CO_2 (ppm)	H_2 (ppm)	CH_4 (ppm)	C_2H_2 (ppm)	C_2H_4 (ppm)	C_2H_6 (ppm)	$CO+CO_2$ (ppm)	IFT Target (mN/m)	IFT Pred (mN/m)
1	2.8	65.4	0.5	0.02	480	4112	77	57	0	5	86	4592	30.2	32.3
2	6.7	49.5	5	0.207	139	3165	0	440	2.18	0	191	3304	21.1	21.9
3	3.0	55.6	0.5	0.02	1439	1646	0	0	0	0	23	3085	31.2	32.3
4	7.6	62.1	0.5	0.044	207	1439	0	0	0	0	42	1646	36.1	34.7
5	7.7	53.5	1.7	0.014	91	3087	0	88	0	0	138	3178	31.5	32.0
6	2.8	95.9	0.5	0.03	188	783	0	261	0	0	287	971	33.4	33.3
7	2.5	90	0.5	0.04	155	665	0	294	0	0	343	820	31.7	32.4
8	12.0	79	2	0.053	94	1670	10	2	13	9	7	1764	26.1	27.6
9	11.5	75	2	0.042	217	1994	14	32	0	50	45	2211	26	27.5
10	5.8	59	0.5	0.008	250	470	20	2	0	2	0	720	39.4	40.9
11	7.9	68	1	0.02	72	818	5	21	0	3	77	890	34	34.8
12	6.4	28	3.5	0.151	584	2969	47	24	0	35	9	3553	22.9	23.5
13	3.6	96	1.5	0.006	0	106	4	4	0	3	2	106	43.4	43.3
14	7.1	96	4	0.067	22	240	3	4	0	0	4	262	23.3	26.5
15	5.2	95	1	0.017	255	1783	141	34	2	3	12	2038	40.2	40.8
16	6.5	97	1	0.014	205	1781	12	2	0	3	1	1986	41.4	40.5
17	7.1	96	4	0.157	436	1925	35	6	12	7	2	2361	19.3	21.8
18	16.2	46.4	0.5	0.01	478	1921	6	159	0	29	483	2399	32.8	34.0
19	20.6	41	3.9	0.01	851	6586	32	10	0	12	9	7437	28.2	26.3
20	6.0	88.2	0.5	0.05	702	2830	7	11	0.5	12	18	3532	33.3	33.7

V. CONCLUSION

A method to handle uncertainties due to data unavailability in calculating transformer health index is proposed. In addition, the impact of unavailable data is evaluated and resulted in an exponential regression line with a high correlation coefficient. More unavailable data resulted in less accurate health index. In this case, it is recommended to utilize the certainty level along with the health index to obtain a better judgment of the transformer condition assessment. An interpretation guideline using the health index and certainty level is also proposed.

Interfacial tension prediction model is also developed to handle the unavailable data. Various machine learning algorithms are used to develop the prediction model based on several collected field data. Among the investigated algorithms, random forest-based prediction model achieved the highest accuracy in estimating the oil IFT based on oil colour and acidity. The developed certainty level model can also accommodate the results of the IFT prediction into the health index result.

By using the certainty level together with the HI value, a proper asset management decision can be taken on the transformer even with some missed data.

APPENDIX

See Tables 9 and 10.

REFERENCES

[1] S. Li and J. Li, "Condition monitoring and diagnosis of power equipment: Review and prospective," *High Voltage*, vol. 2, no. 2, pp. 82–91, Jun. 2017.

[2] J. Wetzer, "Transformer health and risk indexing: Transforming asset data into decision-support information," *Transformers Mag.*, vol. 5, no. 2, pp. 74–79, 2018.

[3] A. Jahromi, R. Piercy, S. Cress, J. Service, and W. Fan, "An approach to power transformer asset management using health index," *IEEE Elect. Insul. Mag.*, vol. 25, no. 2, pp. 20–34, Mar. 2009.

[4] A. Naderian, S. Cress, R. Piercy, F. Wang, and J. Service, "An approach to determine the health index of power transformers," in *Proc. IEEE Int. Symp. Electr. Insul.*, Jun. 2008, pp. 192–196.

[5] I. G. N. S. Hernanda, A. C. Mulyana, D. A. Asfani, I. M. Y. Negara, and D. Fahmi, "Application of health index method for transformer condition assessment," in *Proc. IEEE Region Conf. (TENCON)*, Oct. 2014, pp. 1–6.

[6] L. En-Wen and S. Bin, "Transformer health status evaluation model based on multi-feature factors," in *Proc. Int. Conf. Power Syst. Technol. (POWERCON)*, Oct. 2014, pp. 1417–1422.

[7] K. Taengko and P. Damrongkulkamjorn, "Risk assessment for power transformers in PEA substations using health index," in *Proc. 10th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON)*, May 2013, pp. 1–6.

[8] F. Ortiz, I. Fernandez, A. Ortiz, C. J. Renedo, F. Delgado, and C. Fernandez, "Health indexes for power transformers: A case study," *IEEE Elect. Insul. Mag.*, vol. 32, no. 5, pp. 7–17, Sep. 2016.

[9] C. Ranga and A. K. Chandel, "Expert system for health index assessment of power transformers," *Int. J. Electr. Eng. Informat.*, vol. 9, no. 4, pp. 850–865, 2017.

[10] G. Brandtzæg, "Health indexing of Norwegian power transformers," Norwegian Univ. Sci. Technol., Trondheim, Norway, Tech. Rep., 2015. [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2368049

[11] S. Li, H. Ma, T. Saha, and G. Wu, "Bayesian information fusion for probabilistic health index of power transformer," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 2, pp. 279–287, Jan. 2018.

[12] S. Li, G. Wu, H. Dong, L. Yang, and X. Zhen, "Probabilistic health index-based apparent age estimation for power transformers," *IEEE Access*, vol. 8, pp. 9692–9701, 2020.

[13] S. Milosavljevic and A. Janjic, "Integrated transformer health estimation methodology based on Markov chains and evidential reasoning," *Math. Problems Eng.*, vol. 2020, pp. 1–12, May 2020.

[14] J. I. Aizpurua, B. G. Stewart, S. D. J. McArthur, B. Lambert, J. G. Cross, and V. M. Catterson, "Improved power transformer condition monitoring under uncertainty through soft computing and probabilistic health index," *Appl. Soft Comput.*, vol. 85, Dec. 2019, Art. no. 105530.

[15] A. N. Hassan and A. El-Hag, "Two-layer ensemble-based soft voting classifier for transformer oil interfacial tension prediction," *Energies*, vol. 13, no. 7, p. 1735, Apr. 2020.

[16] R. A. Prasojo, K. Diwyacitta, Suwarno, and H. Gumilang, "Transformer paper expected life estimation using ANFIS based on oil characteristics and dissolved gases (case study: Indonesian transformers)," *Energies*, vol. 10, no. 8, p. 1135, Aug. 2017.

[17] R. A. Prasojo, K. Diwyacitta, Suwarno, and H. Gumilang, "Transformer paper condition assessment using adaptive neuro-fuzzy inference system model," in *Proc. Int. Conf. Electr. Eng. Comput. Sci. (ICECOS)*, Aug. 2017, pp. 237–242.

[18] K. B. Shaban, A. H. El-Hag, and K. Benhmed, "Prediction of transformer furan levels," *IEEE Trans. Power Del.*, vol. 31, no. 4, pp. 1778–1779, Aug. 2016.

[19] K. Shaban, A. El-Hag, and A. Matveev, "A cascade of artificial neural networks to predict transformers oil parameters," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 16, no. 2, pp. 516–523, Apr. 2009.

[20] R. A. Prasojo and Suwarno, "Power transformer paper insulation assessment based on oil measurement data using SVM-classifier," *Int. J. Electr. Eng. Informat.*, vol. 10, no. 4, pp. 661–673, 2018.

[21] H. Gumilang, "Unique relationship between interfacial tension test (IFT) and neutral number test (acidity) of transformer insulation oil in PLN P3B JB–Jakarta and banten regional," in *Proc. IEEE 9th Int. Conf. Properties Appl. Dielectr. Mater.*, Jul. 2009, pp. 29–32.

[22] *CIGRE 761: Condition Assessment of Power Transformers*, CIGRE/Int. Council Large Electr. Syst., Mar. 2019. [Online]. Available: https://e-cigre.org/publication/761-condition-assessment-of-power-transformers

[23] R. A. Prasojo, A. Setiawan, Suwarno, N. U. Maulidevi, and B. A. Soedjarno, "Development of analytic hierarchy process technique in determining weighting factor for power transformer health index," in *Proc. 2nd Int. Conf. High Voltage Eng. Power Syst. (ICHVEPS)*, Oct. 2019, pp. 1–5.

[24] F. Tian, Z. Jing, H. Zhao, E. Zhang, and J. Liu, "A synthetic condition assessment model for power transformers using the fuzzy evidence fusion method," *Energies*, vol. 12, no. 5, p. 857, Mar. 2019.
[25] W.-G. Li, Q. Yu, and R.-C. Luo, "Application of fuzzy analytic hierarchy process and neural network in power transformer risk assessment," *J. Central South Univ.*, vol. 19, no. 4, pp. 982–987, Apr. 2012.
[26] R. A. Prasojo, Suwarno, N. U. Maulidevi, and B. A. Soedjarno, "A multiple expert consensus model for transformer assessment index weighting factor determination," in *Proc. 8th Int. Conf. Condition Monitor. Diagnosis (CMD)*, Oct. 2020, pp. 234–237.
[27] T. L. Saaty, "Decision making with the analytic hierarchy process," *Int. J. Services Sci.*, vol. 1, no. 1, pp. 83–98, 2008.
[28] R. A. Prasojo, W. R. Tamma, Suwarno, N. U. Maulidevi, and B. A. Soedjarno, "A method to calculate uncertainty due to unavailable data in transformer assessment index," presented at the Int. Symp. Electr. Insulating Mater. (ISEIM), 2020.
[29] W. R. Tamma, R. A. Prasojo, and Suwarno, "High voltage power transformer condition assessment considering the health index value and its decreasing rate," *High Voltage*, vol. 6, no. 2, pp. 314–327, Apr. 2021.
[30] *IEEE Guide for the Interpretation of Gases Generated in Mineral Oil-Immersed Transformers*, IEEE Standard C57.104-2019, 2019.
[31] Suwarno and R. A. Pasaribu, "Thermal aging of mineral oil-paper composite insulation for high voltage transformer," *Int. J. Electr. Eng. Informat.*, vol. 8, no. 4, pp. 819–834, 2016.
[32] H. Gumilang, "Hydrolisis process in PLN P3BJB transformers as an effect of oil insulation oxidation," in *Proc. IEEE Int. Conf. Condition Monit. Diagnosis*, 2012, pp. 1147–1150.
[33] C. W. Cigre, "Cigre 738—Ageing of liquid impregnated cellulose for power transformers," Tech. Rep., CIGRE/Int. Council Large Electr. Syst., Aug. 2018. [Online]. Available: https://e-cigre.org/publication/738-ageing-of-liquid-impregnated-cellulose-for-power-transformers
[34] K. Assaleh and A. H. El-Hag, "Estimating transformer oil parameters using polynomial networks," in *Proc. Int. Conf. Condition Monitor. Diagnosis (CMD)*, 2008, pp. 1335–1338.
[35] N. A. Bakar and A. Abu-Siada, "A novel method of measuring transformer oil interfacial tension using UV-vis spectroscopy," *IEEE Elect. Insul. Mag.*, vol. 32, no. 1, pp. 7–13, Jan. 2016.
[36] L. Breiman, "Random forests—Random features," Dept. Statist., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. 567, 1999, pp. 1–29. [Online]. Available: https://www.stat.berkeley.edu/~breiman/random-forests.pdf
[37] A. M. Shah and B. R. Bhalja, "Fault discrimination scheme for power transformer using random forest technique," *IET Gener., Transmiss. Distrib.*, vol. 10, no. 6, pp. 1431–1439, Apr. 2016.
[38] J. Guan, M. Guo, and S. Fang, "Partial discharge pattern recognition of transformer based on deep forest algorithm," *J. Phys., Conf. Ser.*, vol. 1437, no. 1, 2020, Art. no. 012083.
[39] I. H. Kartojo, Y.-B. Wang, G.-J. Zhang, and Suwarno, "Partial discharge defect recognition in power transformer using random forest," in *Proc. IEEE 20th Int. Conf. Dielectr. Liquids (ICDL)*, Jun. 2019, pp. 1–4.

RAHMAN AZIS PRASOJO (Graduate Student Member, IEEE) received the B.Sc. degree from the Department of Electrical Engineering, Politeknik Negeri Malang, Malang, Indonesia, in 2015, and the M.Sc. degree from the School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia, in 2017, where he is currently pursuing the Ph.D. degree. Since 2019, he has been an Assistant Professor with the Department of Electrical Engineering, Politeknik Negeri Malang. He has published several conference papers and journal articles in accordance to high voltage power transformer condition monitoring and diagnostics.

SUWARNO (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Electrical Engineering, Institut Teknologi Bandung, Indonesia, in 1988 and 1991, respectively, and the Ph.D. degree from Nagoya University, Japan, in 1996. He is currently a Professor and the Emeritus Dean of the School of Electrical Engineering and Informatics, Institut Teknologi Bandung. He has published over 200 international journal articles or conference papers. His research interests include high voltage insulating materials and technology and diagnostics of HV equipment. He was the General Chairman of several international conferences, such as ICPADM 2006, ICEEI 2007, CMD 2012, ICHVEPS 2017, and ICHVEPS 2019. He also serves as the Head of the Electrical Power Engineering Research Group and the Editor-in-Chief for the *International Journal on Electrical Engineering and Informatics*.

A. ABU-SIADA (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Ain Shams University, Egypt, in 1998, and the Ph.D. degree from Curtin University, Australia, in 2004, all in electrical engineering. He is currently the Head of the Electrical and Computer Engineering, Curtin University. His research interests include power electronics, power system stability, condition monitoring, and power quality. He is the Vice-Chair of the IEEE Computation Intelligence Society, WA Chapter. He is also the Editor-in-Chief of the *International Journal of Electrical and Electronic Engineering* and a regular reviewer for various IEEE transactions.