Imputing Missing Values in Mammography Mass Dataset: Will it Increase Classification Performance of Machine Learning Algorithms?

Zahriah Sahri, Fahmi Arif, Sharifah Sakinah Syed Ahmad, Rabiah Ahmad, and Rubiyah Yusof

Abstract—Mammography is one of the most effective methods for breast cancer screening and the resulting images are normally reported using the BI-RADS standard. Missing values are found in this BI-RADS dataset which can reduce the classification performance of any machine learning algorithm. This study applies a few established imputation methods that estimate and replace the missing values found in a mammogram mass dataset. Then, a few machine learning algorithms learnt from these imputed datasets to classify between benign and malignant masses. Using classification accuracy as the performance metric, the experimental results showed an increase in accuracy for majority of the combination of machine learning algorithms algorithm and imputation methods.

Keyword—Imputation, machine learning, mammography, missing values.

I. Introduction

Mammography is one of the most effective methods for breast cancer screening. The suspicious mass lesions found in the mammographic images are normally described using the Breast Imaging Reporting and Data Systems (BI-RADS) standard. To help physicians in interpreting the mammographic data more effectively, a few studies [1-4] have proposed the use of machine learning (ML) algorithms in classifying the severity (benign or malignant) of a lesion using the BI-RADS attributes (mass shape, mass margin, mass density, and BI-RADS assessment), the ground truth value, as well other relevant characteristics of a patient .

Unfortunately, due to various reasons, some of these BI-RADS attributes fail to be captured during the mammography process. These manifest into missing values (MV) as can be seen in this mammography dataset [5]. It is documented in [6-7] that the presence of MV in a dataset is found to reduce the classification performance of ML algorithms. Henceforth, many studies across various problem domains (as reviewed by [8]) have applied imputation before classification takes place. Imputation is the act of filling in the MV found in a dataset with estimated values using observed data. It is found out that in general, imputation does increase the classification accuracy of ML algorithms [9-10].

Despite the proven benefits of imputation, however, as of this writing, very few studies [11] have performed imputation for mammographic dataset. The lack of interest could stem from the low percentage of MV, which amount to 3.3% only. Besides, according to [7], 1-5% missing rate is manageable for subsequent interpretations. However, researchers in [11] have reported better classification accuracy when the MV were imputed with their proposed method than otherwise. Thus, this paper applies several established imputation methods, ranging from statistical to machine learning, to impute the MV in [5]. Unlike [11], this paper applies different combinations of imputation methods with different ML algorithms. This paper hypothesizes that despite the small amount of MV, imputation can influence positively the classification task of ML algorithms.

II. Related Work

A missing value indicates a lack of response from scientific experiments [12]. ”NULL”, empty cells and ”?” are some of the possible codes for missing values in a dataset. It is a common fact that missing values persistently appear in most real-world data sources. However, why worry about missing values? Two major negative effects are reported in [13]. First, missing values reduce statistical power. Second, missing values could result in biased statistical estimates in several ways. In addition, most of the data mining algorithms cannot work with datasets having missing values as well as reducing ML predictive performance for pattern classification [6-7].

Litwise deletion, pairwise deletion, regression, mean-mode imputation, expectation-maximization, and multiple imputation are some statistical tools available for imputing missing values. Litwise deletion and pairwise deletion are deficient in several aspects, despite their simplicity [14]. Litwise deletion, in particular, can discard an enormous amount of potentially useful data. Mean-mode imputation, because of its simplicity, is commonly used in the social sciences as a fast alternative to litwise deletion [15]. Also, it is often used as a base for other proposed imputation methods such as in [16]. Expectation-maximization and multiple imputation, currently represent the state of the art, have been applied to various problem domains. Researchers in [17] estimated the missing values of leaf area index, a

Manuscript received August 18, 2014. This work was supported in part by the Universiti Teknikal Malaysia Melaka under Grant PJP/2015/FTMK(1B)/S01396.

Zahriah Sahri is with Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia (e-mail: szahriah@utem.edu.my).

Fahmi Arif is with Institut Teknologi Nasional Bandung, Jalan PHH Mustafa 23, Bandung, Indonesia (e-mail: fahmi.arif@itenas.ac.id).

Sharifah Sakinah Syed Ahmad is with Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia (e-mail: sakinah@utem.edu.my).

Rabiah Ahmad is with Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia (e-mail: rabiah@utem.edu.my).

Rubiyah Yusof is with Universiti Teknology Malaysia, 54100 Kuala Lumpur, Malaysia (e-mail: rubiyah.kl@utm.my).

https://doi.org/10.17758/EIRAI.IAE1017305

biophysical variable, using Expectation-maximization and helped reduce the root mean square error of the Gaussian Bayes Network output. In the medical field [18], multiple imputation was found to preserve observed and real data better than complete-case and dropping a particular variable approaches when predicting for deep venous thrombosis in patients.

Imputation methods based on machine learning are sophisticated procedures that generally consist of creating a predictive model to estimate values that will substitute the missing items. These approaches model the missing data estimation based on information available in the dataset. There are several options varying from imputation with k-nearest neighbour (kNN) to imputation procedures based on auto-associative neural networks [8]. Researcher in [19] proposed Probabilistic Neural Network preceded by mode for imputing the missing categorical data. Using four benchmark datasets in [5] as the experimented datasets, their proposed method outperformed 3 statistical and 3 machine learning methods. For DNA microarray datasets, the researchers in [20] employed a genetic algorithm optimized *k*NN to impute the missing values. The comparative experiments with the mean-mode and the standard *k*NN imputation methods showed that the proposed evolutionary *k*NN was the most effective especially for datasets with higher missing rates. Meanwhile, Fuzzy-rough sets were used by [21] to impute missing data found in 27 benchmarks datasets in [5]. It was found out that these rough sets methods exhibited better excellent performance against 11 state-of-the-art imputation methods.

III. Methodology of Study

This study evaluated a few established imputation methods to estimate and replace the missing values found in mammogram dataset. Mean imputation, class-conditional mean imputation and multiple imputation are the statistical methods used in this study. For the ML-based methods; *k* nearest neighbour imputation, neural network and support vector regression are applied. This study applied a few ML algorithms for classification task using the imputed datasets. They are Naïve-Bayes (NB), C4.5, Decision Stump (DS), Randorm Tree (RT), Random Forest (RF) and Support Vector Machine (SVM). Figure 1 shows the methodology of this study.

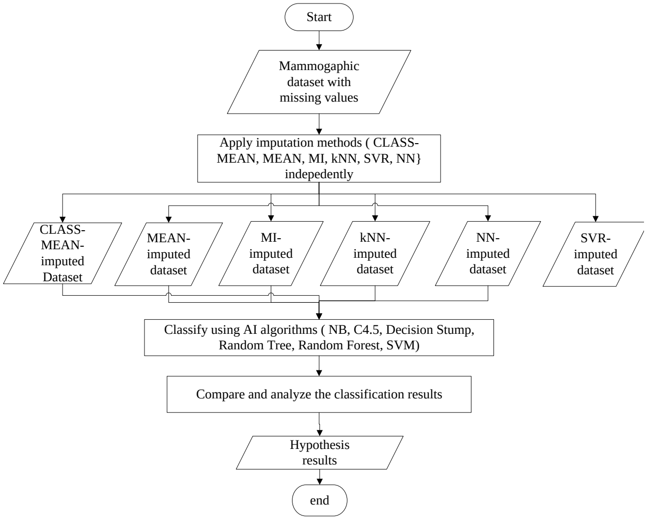
Start
Mammogaphic dataset with missing values
Apply imputation methods (CLASS-MEAN, MEAN, MI, kNN, SVR, NN} indepedently
CLASS-MEAN-imputed Dataset
MEAN-imputed dataset
MI-imputed dataset
kNN-imputed dataset
NN-imputed dataset
SVR-imputed dataset
Classify using AI algorithms (NB, C4.5, Decision Stump, Random Tree, Random Forest, SVM)
Compare and analyze the classification results
Hypothesis results
end

Fig. 1 Proposed methodology

A. Mean Imputation (MEAN)

This method consists of replacing the missing data for a given attribute by the mean (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute. The algorithm imputes missing values for each attribute separately. Consider that there are missing values in an attribute F_j, they are replaced by the mean value $\widetilde{x_j}$,

$$\widetilde{x_j} = \frac{1}{N_{obs,j}} \sum_{i=1}^{N} x_{i,j} \quad (1)$$

where $N_{obs,j}$ is the number of observed values in F_j; $x_{i,j}$ is an observed value in F_j, N is the total number of instances.

B. Class-conditional Imputation (CLASS-MEAN)

An improvement of the MEAN method by restricting the estimation of missing values to the same class labels. For each symbolic attribute, the value that occurs the most often within a class is used to substitute all the missing values within the same class in this attribute. For each continuous attribute, the mean value of all the available values within a class is used to substitute all the missing values within the same class in this attribute [8]. Let us consider that the value $x_{i,j}$ of an attribute $F_{j,}$ of the k-th class, C_k, is missing then it will be replaced by

$$\widetilde{x_j} = \frac{1}{N_{obs,j}} \sum_{i:\, x_{i,j} \in C_k} x_{i,j} \quad (2)$$

where $N_{obs,j}$ is the number of observed values in F_j of the k-th class; $x_{i,j}$ is an observed value in F_j of the k-th class, C_k.

C. Multiple Imputation

This method was proposed by Rubin [22] in 1987. The whole MI procedure is made of three steps. They are imputation, analysis, and pooling processes. This study applies only the first step, i.e. imputation process, because the main interest is to fill in the MV. This simulation method replaces each missing value with M > 1 plausible values, which are drawn randomly from their predictive distribution. M is the number of repetition. Imputing a missing value with m simulated values produces M apparently completed datasets and then the mean of M imputed values replaced the missing values.

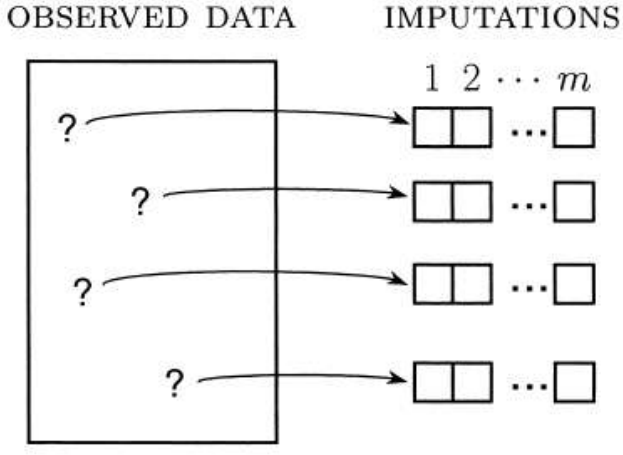

Fig. 2 Schematic representation of Multiple Imputation [12]

D. k-Nearest Neighbour Imputation (kNN)

In this method, for each incomplete instance - instance with MV- the most similar and nearest complete instances - instances that have no MV- are selected. The similarity between two instances is determined using a distance function. Then, the imputed values for each missing datum in an attribute is calculated from the mean or median of the respective attribute in the complete instances. Mean is used with continuous attributes, whilst median is suitable for discrete attributes. For *k*NN, two parameters need to be determined: the number of nearest neighbor (*k)* and the distance function. This study uses Euclidean as the distant function and *k*=5.

Generally, the steps of *k*-NN imputation are as follows:

1) choose *k*, the number of nearest neighbors to be selected.
2) calculate the distance between the sample with the to-be-imputed missing value with an another sample using a distance metric. Let $X_i = \{x_{i1}, x_{i2}, \cdots, x_{im}\}$ denotes the instance that contains the to-be-imputed missing values and $X_q = \left\{x_{q1}, x_{q2}, \cdots, x_{qm}\right\}$ be the other sample. The Euclidean distance between X_i and X_q is:
3) $d\left(X_i, X_q\right) = \sum_{j=1}^{m} \left|x_{ij} - x_{qj}\right|$ (3)
4) where *m* is the number of features in X_i and X_q, and x_{ij} is the *j*th feature of sample X_i and x_{qj} is the *j*th feature of sample X_q.
5) Repeat step 2 to compute the distance between X_i with each remaining sample in the dataset
6) sort in ascending order (based on the calculated distance values) all X_q excluding X_i.
7) select the top *k* samples from the sorted list as the *k*-nearest neighbours to X_i. These *k*-nearest neighbours are $X_{kNN} = \{X_1, X_2, \cdots, X_k\}$.
8) Let x_{ij} be the to-be-imputed missing value in X_i. Then the estimated value is obtained from

$$x_{ij} = \frac{\sum_{l=1}^{k} x_{lj}}{k} \qquad (4)$$

where *k* is the number of nearest neighbours, x_{lj} is the *j*th feature of sample x_{lj}, and $X_l \in X_{kNN}$.

E. Muliple Layer Perceptron (MLP)

A basic MLP imputation approach consists of training an MLP using only the complete cases as regression model: given *d* input features, each incomplete attribute is learned (it is used as output) by means of the remaining complete attributes given as inputs. The MLP imputation scheme can be described as follows:

1) From dataset *X*, select an attribute F_j which contains missing values.
2) Divide the whole dataset *X* into two subsets X_{obs} and X_{miss}. X_{obs} are instances of *X* which F_j contains no missing values; whilst X_{miss} are the opposite of X_{obs}.
3) Using F_j as the target attribute, train MLP with the remaining attributes of X_{obs}.
4) Using the trained MLP, predict the missing values of F_j in $X_{miss.}$
5) Repeat step 1 to 4 for the other attributes of *X* with missing values.

F. Support Vector Regression

SVR - used for regression analysis - is a version of Support Vector Machine. Similar to the MLP imputation, the SVR model is obtained from the complete instances only. The imputation steps are as follow:-

1) From a dataset *X*, select an attribute F_j which contains missing values.
2) Divide the whole dataset *X* into two subsets X_{obs} and X_{miss}. X_{obs} are instances of X which F_j contains no missing values; whilst X_{miss} are the opposite of X_{obs}.
3) Using F_j as the target attribute, train SVR with the remaining attributes of X_{obs}.
4) Using the trained SVR, predict the missing values of F_j in X_{miss}.
5) Repeat step 1 to 4 for the other attributes of *X* with missing values.

IV. Experiments and Results

A. Mammography Mass Dataset

The mammography data set from the UCI machine learning repository [5] was taken for this investigation. This data contains 445 malignant and 516 benign cases. There are five attributes in this dataset corresponding to each mammogram along with the ground truth. The dataset contains 162 missing values amongst individual attributes for which suitable replacements have been performed using the compared imputation methods in this study. Table 1 and Table 2 show the characteristics of the Mammogram Mass datasets and missing values information.

TABLE 1:
MISSING VALUES INFORMATION OF THE MAMMOGRAM MASS DATASET

#Instances	#Attributes	# Instances having missing values	#Attributes having missing values	% Missing data
916	6	131	5	3.37

TABLE 2:
MAMMOGRAM MASS DATASET

Attributes	Types	Values	Label	# Missing Values
BI_RADS assessment	Ordinal	0	Assessment incomplete	2
		1	Negative Benign findings	
		2	Probably benign	
		3	Suspicious	
		4	abnormality	
		5	Highly suggestive of malignancy	
Age	Integer		Age of patients	5
Mass Shape	Nominal	1	Round	31
		2	Oval	
		3	Lobular	
		4	Irregular	
Mass Margin	nominal	1	circumscribed	48
		2	microlobulated	
		3	obscured	
		4	ill-defined	
		5	spiculated	
Mass Density	Ordinal	1	High	76
		2	iso	
		3	low	
		4	fat-containing	
Severity	Binomial	0	Benign	0
		1	Malignant	

B. Experimental Design

This study used MATLAB to impute the missing values using kNN, and MLP methods. SPSS was used to impute missing values using CLASS-MEAN, MEAN, and MI method. Meanwhile, LIBSVM was used to impute the missing values using SVR. RAPIDMINER was used for all of the machine learning algorithms used in this study.

C. Results

This study used classification accuracy as the evaluation criteria. For each machine learning algorithm used in this study, the classification accuracy was obtained from 10-crossvalidation accuracies. Table 3 shows the tabulated results and Figure 3 illustrates in bar graph.

From the results, NB and SVM were two of the ML algorithms that clearly improved or maintained their classification performance across all imputation methods applied in this study. For RT and RF, their classification task was negatively affected by the imputation data, obtained through CLASS-MEAN and MI methods respectively. C4.5 performed worse on MEAN and MI imputed datasets than the non-imputed dataset. DS exhibited similar observation with C4.5 but on different imputation methods – NN and SVR.

For the imputation methods, only kNN consistently help increased the learning performance of the ML algorithms under study. The other four, CLASS-MEAN, MEAN, NN, and SVR, reduced the performance of one ML algorithm as aforementioned above. Only MI method affected negatively the learning performance of two ML algorithms, namely, C4.5 and RF. There are altogether 36 combinations of ML algorithms and imputation methods in this study. 30 out 36 of these combinations produced higher or equal classification accuracies than classification using non-imputed Mammogram Mass dataset. It can be said that majority of the machine learning algorithms benefit from the imputation process before learning takes place.

TABLE 3:
CLASSIFICATION ACCURACIES OF ML ALGORITHMS ON MAMMOGRAM MASS DATASET THAT ARE OBTAINED USING DIFFERENT IMPUTATION METHODS

	NO IMPUTE	CLASS-MODE	MEAN	MI	kNN	NN	SVR
NB	82.42	84.40	82.42	82.42	84.29	83.15	83.04
C4.5	75.45	75.96	74.61	74.61	76.17	79.29	81.48
DS	81.79	82.00	82.00	82.00	82.00	81.69	81.69
RT	78.98	78.26	79.92	79.92	79.30	81.69	80.44
RF	82.20	82.73	82.73	82.10	82.63	82.11	82.63
SVM	82.56	82.94	82.94	83.20	83.05	83.03	83.77

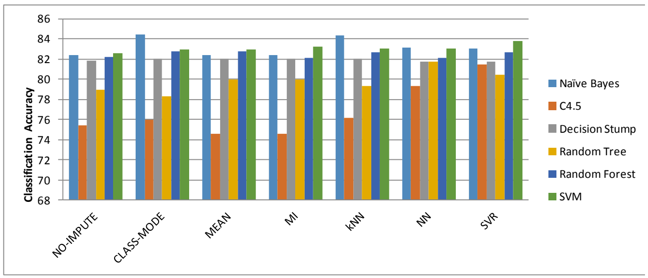

Fig. 3 Classification Accuracy: Different ML algorithms against different imputation methods

V. Conclusion

This study hypothesizes that despite the small amount of missing values in a dataset, ML algorithms will benefit from the imputation process before learning takes place. Using Mammogram mass dataset with only 3.3% of missing values, 3 imputation methods from statistical and 3 imputation methods from ML were applied to impute the missing values in this dataset. To prove the hypothesis, 6 different ML algorithms with different ways in classifying data were used to classify the imputed data from various imputation methods into benign and malignant. From the experimental results, it is found that majority of the combination of ML algorithms and imputation methods outperformed the combination of ML algorithms with the non-imputed Mammogram dataset. It can be concluded that this study hypothesis is proven positive.

References

[1] F. Narváez, G. Diaz, C. Poveda, and E. Romero, “An automatic BI-RADS description of mammographic masses by fusing multiresolution features,” *Expert Systems with Applications*, vol. 74, pp. 82-95, May 2017.

[2] D. Kaushik and K. Kaur, “Application of Data Mining for high accuracy prediction of breast tissue biopsy results,” in *Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), 2016 Third International Conference*, 2016, pp. 40-45.

[3] S. Chaurasia, P. Chakrabarti, and N. Chourasia “Prediction of Breast Cancer Biopsy Outcomes - an Approach using Machine Leaning Perspectives,” *International Journal of Computer Applications,* vol. 100, Aug 2014, pp. 29-32.

[4] J. H. Miao, K. H. Miao, and G. J. Miao, “Breast Cancer Biopsy Predictions Based on Mammographic Diagnosis Using Support Vector Machine Learning,” *Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Bioinformatics*, vol. *5*, pp.1-9.

[5] A. Asuncion and D. Newman, “UCI machine learning repository,” ed, 2007.

[6] L. Peng, L. Lei, and N. Wu, “A Quantitative Study of the Effect of Missing Data in Classifiers,” in *Computer and Information Technology, 2005. CIT 2005. The Fifth International Conference*, 2005, pp. 28-33.

[7] E. Acuña and C. Rodriguez, “The Treatment of Missing Values and its Effect on Classifier Accuracy,” in *Classification, Clustering, and Data Mining Applications*, D. Banks, *et al.*, Eds., ed: Springer Berlin Heidelberg, 2004, pp. 639-647.

[8] P. J. García-Laencina, J. L. Sancho-Gómez, and A.R. Figueiras-Vidal, “Pattern classification with missing data: a review,” *Neural Computing and Applications*, vol. 19, pp. 263-282, Mar 2010.

[9] Q. Song, M. Shepperd, X. Chen, and J. Liu, “Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation,” *Journal of Systems and Software,* vol. 81, pp. 2361-2370, Dec 2008.

[10] A. Farhangfar, L. Kurgan, and J. Dy, “Impact of imputation of missing values on classification error for discrete data,” *Pattern Recognition,* vol. 41, pp. 3692-3705, Dec 2008.

[11] T. T. Nguyen and Y. Tsoy, “A kernel PLS based classification method with missing data handling,” *Statistical Papers,* vol. 58, pp. 211-225, Mar 2017.

[12] J. L. Schafer and J. W. Graham, “Missing data: our view of the state of the art,” *Psychological Methods,* vol. 7, p. 147, Jun 2002.

[13] N. Tsikriktsis, "A review of techniques for treating missing data in OM survey research,” *Journal of Operations Management*, vol. 24, pp. 53-62, Dec 2005.

[14] M. Glasser, “Linear regression analysis with missing observations among the independent variables,” *Journal of the American Statistical Association,* vol. 59, pp. 834-844, Sep 1964.

[15] J. M. Jerez, *et al.*, “Missing data imputation using statistical and machine learning methods in a real breast cancer problem,” *Artificial Intelligence in Medicine,* vol. 50, pp. 105-115, Sep 2010.

[16] A. Farhangfar, L. A. Kurgan, W. Pedrycz, “A Novel Framework for Imputation of Missing Values in Databases,” *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on,* vol. 37, pp. 692-709, Sep 2007.

[17] Y. T. Mustafa, V. A. Tolpekin, and A. Stein, “Application of the Expectation Maximization Algorithm to Estimate Missing Values in Gaussian Bayesian Network Modeling for Forest Growth,” *Geoscience and Remote Sensing, IEEE Transactions on,* vol. 50, pp. 1821-1831, May 2012.

[18] K. J. M. Janssen, A. R. T. Donders, F. E. Harrell, Y. Vergouwe, Q. Chen, D. E. Grobbee, and K. G. Moons, “Missing covariate data in

https://doi.org/10.17758/EIRAI.IAE1017305

medical research: To impute is better than to ignore," *Journal of Clinical Epidemiology,* vol. 63, pp. 721-727, Jul 2010.

[19] K. J. Nishanth and V. Ravi, "Probabilistic neural network based categorical data imputation," *Neurocomputing,* vol. 218, pp. 17-25, Dec 2016.

[20] H. M. de Silva and A. S. Perera, "Evolutionary k-Nearest Neighbor Imputation Algorithm for Gene Expression Data," *ICTer,* vol. 10, Jun 2017.

[21] M. Amiri and R. Jensen, "Missing data imputation using fuzzy-rough methods," *Neurocomputing,* vol. 205, pp. 152-164, Sep 2016.

[22] D. B. Rubin and N. Schenker, "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse," *Journal of the American Statistical Association,* vol. 81, pp. 366-374, Jun 1986.

Zahriah Sahri received the Diploma degree in computer science from Universiti Teknologi Mara, Shah Alam, Malaysia in 1995; the Bachelor's degree in computer science (Hons.) from Universiti Teknologi Malaysia, Johor Bahru, Malaysia, in 2003; the Master's degree in computer science from Universiti Putra Malaysia, Selangor, Malaysia, in 2006; and the Ph.D. degree at the Universiti Teknologi Malaysia, Johor Bahru, Malaysia, in 2016. She worked as an IT application developer at different type of industries. She is currently a Senior Lecturer in the Department of Intelligent Computing and Analytics, Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka. Her research interests include power transformer fault diagnostic, missing data imputation, feature selection, and meta-heuristic algorithms.

Fahmi Arif has received his Bachelor of Engineering in mechanical engineering (Bandung, Indonesia), Master of Engineering in industrial engineering (Bandung, Indonesia), and PhD in industrial computing (Melaka, Malaysia).

After several years working oversea in the field of education and consultancy as well, currently he is working at his hometown as a Senior Lecturer in the Department of Industrial Engineering at Institut Teknologi Nasional Bandung. Along his career, he has published various articles in some international conferences and journals. His research interest is in data mining, artificial intelligent, and machine learning especially for its application in industrial automation. Currently he is working in some research projects in the field of Cyber Physical System for Industry 4.0.

Dr Fahmi also serves as a reviewer in some local and international journal such as Journal of Information and Organizational Science (ISSN: 1846-9418) and editorial board for Journal of Intelligent System (ISSN: 2356-3974).

Sharifah Sakinah received her Bachelors and Masters degrees of applied mathematics in School of Mathematics from University Science Malaysia. Following this, she received her Ph.D from the University Of Alberta, Canada in 2012 in Intelligent System.

She is currently a Senior Lecturer in the Department of Intelligent Computing and Analytics, Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). Her research in graduate school focused on the granular computing and fuzzy modeling. Her current research interests including evolutionary optimization, fuzzy system, granular computing, evolutionary method, and data analitics

Rabiah Ahmad received her Diploma and Masters degrees in computer science from Universiti Teknologi Malaysia, Johor Bahru, Malaysia, in 1995 and 1997, respectively; the the Master's degree in information security from the Royal Holloway University of London, UK, in 1998; and the Ph.D. degree in information studies (Health Informatics) from the University of Sheffield, UK, in 2006.

She started her career as lecturer at Universiti Teknologi Malaysia in 1997 and in 2010 moved to Universiti Teknikal Malaysia Melaka. She is currently a Professor in Information Security and Privacy since 2014. Her research interests include Information Security Management, Security Architecture, Healthcare system security and Cyber Physical System Security. She was a project leader with the amount of RM 700K for 3 projects funded by Ministry of Science Technology and Innovation, 3 projects funded by Ministry of Education. She also a project member for RM 2.6 Million Project funded under the Long Term Research Grant Scheme (LRGS). She was awarded University (UTeM) Research Award in 2011. Throughout her career, she has published various articles in the area of health informatics, information security and cyber physical at national and international conferences. In addition, her publications appeared in local and international journal of the same area.

Prof. Rabiah Ahmad was appointed as Editorial Board for the International Journal of Cryptography Research (ISBN 1985-5753). Besides that, she also serves as a reviewer for various conferences and journals i.e., International Journal of Medical Informatics.

Rubiyah Yusof received the B.Sc. (Hons.) degree in electrical and electronics engineering from the University of Loughborough, Loughborough, U.K., in 1983; the Master's degree in control systems from the Cranfield Institute of Technology, Cranfield, U.K., in 1986; and the Ph.D. degree in control systems from the University of Tokushima, Tokushima, Japan, in 1994.

She is currently the Dean of the Malaysia–Japan International Institute of Technology (MJIIT) and a Professor of Faculty of Electrical Engineering, Universiti Teknologi Malaysia (UTM), Kuala Lumpur, Malaysia. Throughout her career as Professor and Researcher at UTM, she has been acknowledged for her many contributions in artificial intelligence, process control, and instrumentation design. She is recognized for her work in biometrics systems, such as Syariah Compliant Automated Chicken Processing System (Sycut), KenalMuka (Face Recognition System), and the Signature Verification System, which won both national and international awards. She is the author of the book *Neuro-Control and Its Applications* (Springer Verlag, 1995), which was translated to Russian in 2001.

Prof. Rubiyah Yusof is a Member of the AI Society of Malaysia, the Instrumentation and Control Society of Malaysia, and the Institute of Electrical and Electronics Engineers of Malaysia.

.

https://doi.org/10.17758/EIRAI.IAE1017305