Colorectal Cancer Histology Image Classification Using Stacked Ensembles

Nuntawut Kaoungku, Kittisak Kerdprasop, Nittaya Kerdprasop, Ratiporn Chanklan, and Keerachart Suksut

***Abstract*—In this research, we propose the image processing and classification using image extraction and data mining with ensemble learning techniques. We apply image extraction to determine the appropriate subset of the initial features to be used later by the ensemble learning for inducing an accurate model to predict cancer from the colorectal cancer histology images. Our proposed ensemble image classification method consists of three main parts: the image pre-processing part to adjust the image contrast to show the clear nucleus that can be recognized as the cause of cancer, the image extraction part to extract only important features, and finally the model creation part that generate the model to be used later as an image-based predictor. The experimental results show that the proposed method can predict the colorectal cancer from the colon images with high accuracy.**

***Index Terms*—Image classification, image pre-processing, image extraction, ensemble learning.**

I. Introduction

Currently, innovative technology has played a prominent role in human daily life in vast aspects including ubiquitous communication, smart transportation, and tele-medicine. The technology of the Internet of Thing (IoT) is constantly evolving to embed in so many machines and appliances that it makes the technology around us. The emergence of IoT also causes new structured and unstructured data to be generated and stored all the time. Such tremendous data urge a more practical and timely research on Machine Learning and Artificial Intelligence to meet the human demand and business needs. At present, there are many unstructured data waiting for an efficient method to utilize them for beneficial purpose.

The unstructured data are data that contain disorderly pattern. Such data are images, text conversation, voice, and other forms of data produced by the current smart applications. These data cannot be applied directly like those in a relational database that are stored with a well-formed schema. Therefore, the unstructured data such as images must be passed through several steps of image processing to convert unstructured data to structured data before applying them with the machine learning technique to turn quite useless data into valuable knowledge.

In medical research, the image data are visually used for early diagnostic to characterize stage of symptom in order to give appropriate advice to patients. For some cases, images can be used to predict the patient's disease. These image utilizations are based on the human expertise. To facilitate doctors with a systematic machine-based approach, many intelligent steps are needed since the image pre-processing up to the machine learning and predicting steps.

One image consists of many features as well as noise that can occur during the data collection process. It is challenging to extract only features of interest from initial features. The two image processing steps, image pre-processing and image extraction, are necessary for image data preparation. Image pre-processing steps such as removing noise and adjusting image color are for preparing images to be ready for future use. Image extraction is for detecting objects or extracting a dominant features with various techniques from pictures resulting in structured or unstructured data depending on the selected technique. The well prepared images can be further used in the machine learning process.

Machine learning technique can learn the statistical model from image data and use that model for future diagnostic prediction. At present, one of the most accurate machine learning techniques is ensemble learning that uses multiple learning algorithms to collaboratively predict data. because each set of data will be effective with some learning algorithm and some researches that are proposed to be used Ensemble techniques with medical image data.

Many researchers had applied ensemble learning in the medical image domain. Gupta and Bhavsar [1] built classifiers from the breast cancer histopathological image data with ensemble learning that consists of four techniques: SVM, decision tree, kNN, and discriminant analysis. Fabio et al. [2] uses six feature extraction techniques: LBP, CLBP, LPQ, GLCM, PFTAS, and ORB to extract color-texture features from images with ensemble technique to classify breast cancer. Chaiyakhan et al. [3] applied image processing on the mammography images to adjust image quality and extract three significant features including texture, shape, and intensity histogram to be further used by SVM to classify breast cancer.

Manuscript received January 30, 2019; revised July 30, 2019.

This work has been supported by grants from Suranaree University of Technology through the funding of Data and Knowledge Engineering Research Units.

Nuntawut Kaoungku and Kittisak Kerdprasop are with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: nuntawut@sut.ac.th, kerdpras@sut.ac.th).

Nittaya Kerdprasop is with the School of Computer Engineering and Data and Knowledge Engineering Research Unit, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: nittaya@sut.ac.th).

Ratiporn Chanklan is with the Data and Knowledge Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology, Thailand. (e-mail: arc_angle@hotmail.com).

Keerachart Suksut is with the Computer Engineering Department, Rajamangala University of Technology Isan, Nakhon Ratchasima 30000, Thailand (e-mail: mikaiterng@gmail.com).

doi: 10.18178/ijfcc.2019.8.3.549

In this work, we propose a technique to classify colon cancer from images prepared with immunohistochemical staining technique to separate the nuclei from the image. We extract significant image features with three techniques: gray-level co-occurrence matrix, local binary pattern, and histogram-based features. We then apply stacked ensemble technique to generate the model for cancer classification.

II. Materials and Methods

A. Immunohistochemical Staining

In general, hematoxylin and eosin are staining techniques used for pathological biopsy diagnoses. But in some cases like infectious disease and cancer, other kind of staining technique such as immunohistochemistry is more effective. Immunohistochemistry (IHC) is the method for detection of antigens on the cell and nucleus and this method has been used extensively to diagnose tumor markers, hormone, enzyme, protein, bacteria, and virus [4].

Ruifrok et al. [5] proposed the color separation method on IHC image from RGB camera by color deconvolution. The color in each of the RGB channels has been converted to the matrix for the combination of Hematoxylin, Eosin, and DAB as shown in Fig. 1. The intensity of light after transmitted the specimen can be computed [5] with equation (1).

$$I_C = I_{0,C} \exp\left(-A \times c_C\right) \tag{1}$$

The optical density (OD) for each of the RGB channels can be computed with equation (2).

$$OD_C = \log_{10}\left(\frac{I_C}{I_{0,C}}\right) = A \times c_C \tag{2}$$

where $I_{0,C}$ is the intensity of light entering the specimen, c is the absorption factor, and A is the amount of stain. Color deconvolution, which can be computed with equation (3), is the separation of the stains with normalization of the *OD* vector.

$$\hat{p}_{i,j} = p_{i,j} \Big/ \sqrt{p_{i,j}^2 + p_{i,j+1}^2 + p_{i,j+2}^2} \tag{3}$$

B. Noise Removal

The median filter is the nonlinear noise reduction which often used as noise removal within the image [6]. It is used in the pre-processing step to adjust the image to be smooth and clear, while retaining image quality such as image clarity and edge contour.

R	G	B	
p_{11}	p_{12}	p_{13}	Hematoxylin
p_{21}	p_{22}	p_{23}	Eosin
p_{31}	p_{32}	p_{33}	DAB

Fig. 1. The matrix for the combination of Hematoxylin, Eosin, and DAB.

C. Contrast Adjustment

Gamma correction or image enhancement is the technique to enhance the intensity of the gray-level of the object in the area of interest within the image to be clearer, which can help improving the image classification. Gamma correction is the nonlinear method to adjust the brightness of the image to be appropriate by adjusting the parameter called gamma [7], which can be computed with equation (4).

$$Corrected = 255 * \left(\frac{Image}{255}\right)^{\left(\frac{1}{\gamma}\right)} \tag{4}$$

where γ is the parameter to convert color intensity in the image. The value of γ that is less than 1 is to convert darkness to brightness, which is called encoding gamma or gamma compression. While γ > 1 is for increasing the intensity of colors in the dark area, which is called decoding gamma or gamma expansion.

D. Grey-Level Co-occurrence Matrix (GLCM)

The pattern in the image is useful for object identification in the interesting area. GLCM [8] is the technique for finding the features of pattern in the image including contrast, correlation, and homogeneity. Contrast is the clarity of the pattern, which can be computed with equation (5). Correlation is the co-occurrence of patterns, which can be computed with equation (6). Homogeneity of the patterns can be computed with equation (7).

$$Contrast = \sum_{i=1}^{m}\sum_{j=1}^{n}(i-j)^2 P(i,j) \tag{5}$$

$$Correlation = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\{i \times j\} \times P(i,j) - \{\mu_x \times \mu_y\}}{\sigma_x \times \sigma_y} \tag{6}$$

$$Homogeneity = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{P(i,j)}{1+|i-j|}\right) \tag{7}$$

where *m* is the width of the GLCM matrix, *n* is the height of the GLCM matrix, *P(i,j)* is the probability at position *(i,j)* of the GLCM matrix, μ_x is the average by rows, μ_y is the average by columns, σ_x is the variance by rows, and σ_y is the variance by columns.

E. Local Binary Pattern (LBP)

LBP is the feature extraction from images to get texture with the value agent of the 3x3 pixel using the center point of the 3x3 pixel group as the reference value for calculating values in the surrounding pixels. The result is in the form of a binary pattern [9]. LBP can be computed with equation (8).

$$LBP(x_x, y_c) = \sum_{p=0}^{P-1} 2^p f(i_p - i_c) \tag{8}$$

where (x_x,y_c) is the pixel point of interesting, i_p and i_c are the pixel points around the pixel point of interesting, *P* is the number of pixel points.

F. Histogram Based Feature

The shape and other properties are often extracted from images as represented as histogram. The histogram normally contains four statistic features (mean, variance, skewness,

and kurtosis) relating to color intensity in the image. Mean is the average of color intensity [10] and can be computed with equation (9).

$$\mu = \sum_{i=1}^{G-1} iP(i) \quad (9)$$

Variance is the change in color intensity around the average (μ) and can be computed with equation (10).

$$\sigma^2 = \sum_{1=1}^{G-1} (i-\mu)^2 P(i) \quad (10)$$

Skewness is for checking the symmetry of the histogram and it can be computed with equation (11). If the histogram is symmetric, then the skewness value is 0.

$$s = \sigma^{-3} \sum_{i=1}^{G-1} (i-\mu)^3 P(i) \quad (11)$$

Kurtosis is for checking whether the maximum and minimum points in a histogram are related to the normal distribution. It can be computed with equation (12).

$$k = \sigma^{-4} \sum_{i=1}^{G-1} (i-\mu)^4 P(i) \quad (12)$$

We can find the population of grayscale intensity occurring in the image with the computation as in equation (13).

$$P(i) = \frac{h(i)}{N \times M} \quad (13)$$

where i is 0, 1, 2, …, G-1, h is number of pixels in each grayscale intensity in the range from 0 to 255, G is level of grayscale, N is number of pixels in the horizontal-oriented image, M is number of pixels in the vertical-oriented image.

G. Ensemble Learning

Machine learning techniques have a variety of intrinsic characteristics that fit different kinds of data. The ensemble learning is a technique that tries to combine the advantages of many algorithms to yield a better predictive performance. Fig. 2 shows the concept of ensemble learning by building multiple models with the same or even different algorithms and then combine the outputs with various techniques such as majority vote or some other measures. Ensemble learning can be divided further into four classes of techniques [11]: voting, bagging, boosting, and stacking.

Voting technique builds models with multiple algorithms to make the model more diverse on the same the training dataset. The outputs are combined based on the majority vote method to predicted class of the new data [11].

Bagging is a technique to build multiple models from sampling data. The number of sampling datasets is based on the required number of nodes in an ensemble model. The predicting outcome is from the voting method [11].

Boosting is based on the concept of building multiple weak models (models with low accuracy) with different weighting values on the training dataset. The combined outputs are using weight of voting to predicted class of the new data [11].

Stacking is similar to the voting method in the process of building model; this process is called learning base-level classifiers. At the outcome combination step, several learning algorithms are used for predicting class of the new data; this process is called learning meta-classifier. The popular meta-learning algorithm is logistic regression [12], [13].

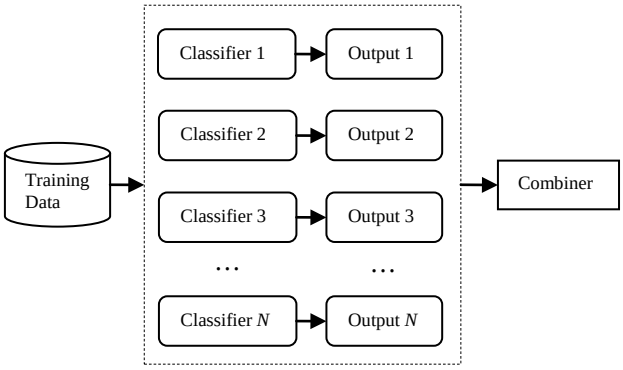

Fig. 2. The concept of ensemble learning.

Algorithm Stacking

1. Input: training data D = {x_i, y_i}, i = 1..m
2. Output: ensemble classifier H
3.
4. *Step 1: learn base-level classifiers*
5. **for** t = 1 to T **do**
6. learn h_t based on D
7. **end for**
8.
9. *Step 2: construct new data set of predictions*
10. **for** i = 1 to m **do**
11. D_h = {x'_t,y_i }, where x'_t = {$h_1(x_i)$, …,$h_T(x_i)$}
12. **end for**
13.
14. *Step 3: learn a meta-classifier*
15. learn H based on D_h
16.
17. return H

Fig. 3. Stacking algorithm [14].

Fig. 3 shows the stacking ensemble algorithm, which consists of three steps that can be explained as follows.

Step 1: from line 4 to 7, is for building multiple models from several learning algorithms.

Step 2: from line 9 to 12, is for generating the new data from predictive results from step 1.

Step 3: from line 14 to 17, is for learning model with the chosen algorithm from data in step 2 and predicting class of the new data.

III. Proposed Work

In this section, we present in detail the proposed process of colorectal cancer histology image classification using stacked ensembles. The idea is that we use the image pre-processing techniques for extracting the cell nucleus and for improving image quality, and then use the image extraction techniques for extracting features with the texture and gray-level color from the image. We finally use this result to create models for data classification.

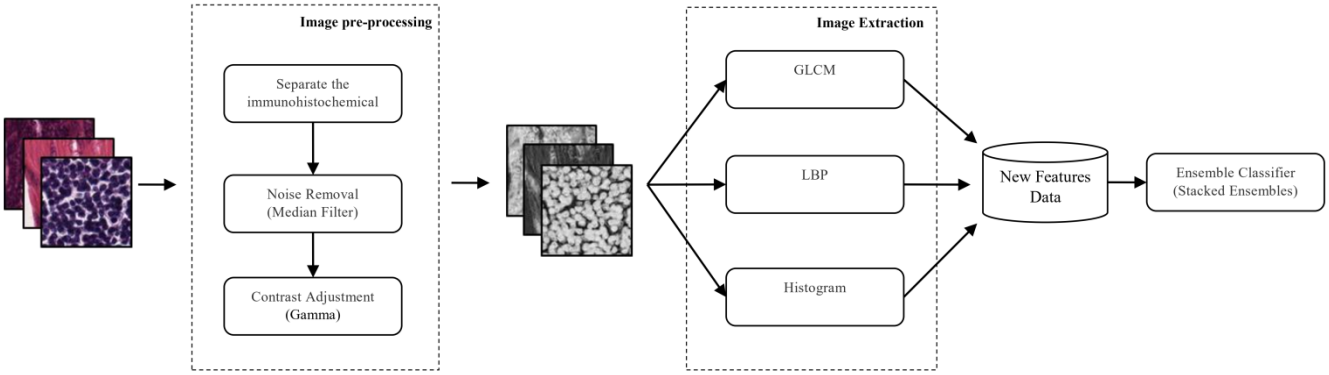

Fig. 4. The concept of colorectal cancer histology image classification using stacked ensembles.

TABLE I: COMPARATIVE RESULTS OF CLASSIFICATION ACCURACY AND ERROR

Classifier	Raw Data		GLCM		LBP		Histogram		Proposed method	
	Accuracy	Error	Accuracy	Error	Accuracy	Error	Accuracy	Error	Accuracy	Error
kNN	0.513	0.487	0.642	0.358	0.668	0.332	0.680	0.320	0.796	0.205
Decision Tree	0.265	0.735	0.603	0.397	0.371	0.629	0.606	0.394	0.365	0.635
Neural Network	0.578	0.423	0.707	0.293	0.691	0.309	0.692	0.308	0.861	0.139
Logistic Regression	0.471	0.529	0.660	0.340	0.660	0.340	0.641	0.359	0.842	0.158
SVM	0.306	0.694	0.694	0.306	0.675	0.325	0.723	0.277	0.884	0.116
Bagging	0.298	0.702	0.604	0.396	0.479	0.521	0.627	0.373	0.670	0.330
Boosting	0.125	0.875	0.603	0.397	0.125	0.875	0.606	0.394	0.125	0.875
Majority Voting	0.518	0.482	0.711	0.289	0.689	0.312	0.705	0.295	0.872	0.128
Stacking	0.306	0.694	0.727	0.273	0.706	0.294	0.724	0.276	**0.887**	**0.113**

Fig. 4 shows the concept of the proposed process which can be divided into three phases: image-preprocessing, image extraction, and building stacked ensembles. The first phase is image pre-processing to be applied for extracting nucleus with the immunohistochemical staining method, then improve image quality with noise removal with median filter technique, and finally adjusting contrast with gamma technique to make the nucleus in the image more clear. Image extraction phase starts with extracting features with three techniques including GLCM, LBP, and histogram. The result is new data features. GLCM uses contrast correlation and homogeneity on four directions (0^o, 45^o, 90^o and 135^o). Histogram contains mean, variance, skewness, and kurtosis.

On the final phase, we build data classification model with the features obtained from the image extraction phases using ensemble method. The base-level in ensemble includes SVM, neural network, and logistic regression algorithms. For the meta-classifier, we use logistic regression algorithm.

IV. EXPERIMENTAL RESULTS

The proposed method has been experimented with 5,000 colorectal cancer histology images containing 8 different classes, including Tumor, Stroma, Complex, Lympho, Debris, Mucosa, Adipose, and Empty, as shown in Fig. 5. On performance evaluation step, we perform the holdout method in which 70% of data (3,500 images) are used for training, whereas the remaining 30% of data (1,500 images) are used for accuracy testing.

Table I shows comparative accuracy and error results of classifiers built from a single-model scheme (kNN, decision tree, neural network, logistic regression, and SVM) versus an ensemble scheme (bagging, boosting, majority voting, and stacking). The training image data are prepared with various image processing techniques: GLCM, LBP, histogram, and the proposed image preparation steps. Raw data are used as a benchmark.

It can be seen that among all single-model classifiers, neural network shows the best performance at accuracy rate 0.578 on raw data. With our proposed image preparation steps that combine features from GLCM, LBP, and histogram techniques, neural network can improve its accuracy rate up to 0.861. The improvement rate is very considerable in SVM in which accuracy rate increases from 0.306 to 0.884. On ensemble scheme comparison, the stacking classifiers built from our proposed image preparation steps show the highest accuracy rate at 0.887.

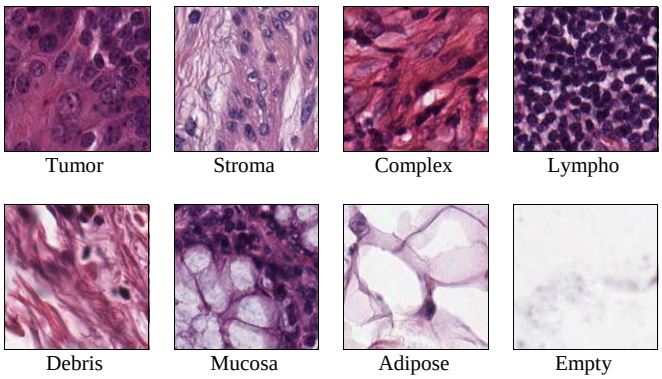

Fig. 5. Example of colorectal cancer histology images showing 8 different classes.

V. CONCLUSION

In this research, we propose an image preparation and ensemble learning techniques to classify colon cancer from a set of histology images. Our aim is to be an alternative assistance for a doctor in screening patients. We consider that

the colorectal cancer histology images contain multiple details that are difficult to separate cancerous cells from others. We design and demonstrate a technique of separating the nucleus from other surrounding objects in the images that had been prepared with the immunohistochemical staining technique, and then extract significant features from a color cancer histology images using image extraction techniques. Our model creation technique is based on the stacked ensemble technique with base-level contains three algorithms: ANN, SVM, and logistic regression. The meta-classifier level uses logistic regression algorithm. The experimental results show that our proposed method yields the best performance on classifying the cancerous cells.

REFERENCES

[1] V. Gupta and A. Bhavsar, "Breast cancer histopathological image classification: is magnification important?" in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[2] A. F. Spanhol, S. L. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455-1462, 2016.

[3] K. Chaiyakhan, N. Kerdprasop, and K. Kerdprasop, "Feature selection techniques for breast cancer image classification with support vector machine," in *Proc. Int. Multi. Conf. Eng. Comp. Sci. Hong Kong*, 2016.

[4] R. S. Shi, E. M. Key, and L. K. Kalra, "Antigen retrieval in formalin-fixed, paraffin-embedded tissues: An enhancement method for immunohistochemical staining based on microwave oven heating of tissue sections," *Journal of Histochemistry & Cytochemistry*, vol. 39, no. 6, pp. 741-748, 1991.

[5] C. A. Ruifrok and A. D. Johnston, "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology*, vol. 23, no. 4, pp. 291-299, 2001.

[6] Z. Wang and D. Zhang, "Progressive switching median filter for the removal of impulse noise from highly corrupted images," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 46, no. 1, pp. 78-80, 1999.

[7] S. Rahman, M. M. Rahman, M. Abdullah-Al-Wadud, D. G. Al-Quaderi, and M. Shoyaib, "An adaptive gamma correction for image enhancement," *EURASIP Journal on Image and Video Processing*, vol. 2016, no. 1, 2016.

[8] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[9] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971-987, 2002.

[10] I. Beheshti, N. Maikusa, H. Matsuda, H. Demirel, and G. Anbarjafari, "Histogram-based feature extraction from individual gray matter similarity-matrix for Alzheimer's disease classification," *Journal of Alzheimer's Disease*, vol. 55, no. 4, pp. 1571-1582, 2017.

[11] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. International Workshop on Multiple Classifier Systems*, June 2000, pp.1-15.

[12] H. D. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, 1992.

[13] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49-64, 1996.

[14] C. C. Aggarwal, *Data Classification: Algorithms and Applications*, USA: CRC Press, 2014.

Nuntawut Kaoungku is currently a lecturer at School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. He received his bachelor, master, and doctoral degrees in Computer Engineering from SUT in 2012, 2013, and 2015, respectively. His current research work includes Data Mining, Knowledge Engineering, and Semantic Web.

Kittisak Kerdprasop is an associate professor and chair of Computer Engineering School, SUT. He received bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, MS in Computer Science from the Prince of Songkla University, in 1991, and PhD in Computer Science from Nova Southeastern University, U.S.A., in 1999.

Nittaya Kerdprasop is an associate professor at the School of Computer Engineering, and head of Data and Knowledge Engineering Research Unit, SUT. She received her bachelor degree in Radiation Techniques from Mahidol University, Thailand, in 1985, MS in Computer Science from the Prince of Songkla University in 1991, and PhD in Computer Science from Nova Southeastern University, U.S.A, in 1999.

Ratiporn Chanklan is currently a researcher with the Data and Knowledge Engineering Research Unit, School of Computer Engineering, SUT. She received her bachelor, master, and doctoral degrees in Computer Engineering from SUT in 2013, 2014, and 2018, respectively. Her current research of interest is Data Mining and Artificial Intelligence.

Keerachart Suksut is currently a lecturer at Computer Engineering Department, Rajamangala University of Technology Isan, Thailand. He received his bachelor, master, and doctoral degrees in Computer Engineering from Suranaree University of Technology, Thailand, in 2012, 2014, and 2016, respectively. His current research of interest includes data mining, genetic algorithm, and imbalanced data classification.