Better Classification of Pathological Tissue Classes from EIS data of Breast Tissue

Hyontai Sug

***Abstract*—**Electrical impedance spectroscopy (EIS) is an important technique to collect data from pathological tissue of breast tissue, and data mining algorithms are used to analyze and diagnose pathological tissue classes from the measured data. In order to build accurate data mining models the diversity and size of data are very important because such properties can make the trained data mining models cover more unseen cases. Neural networks and decision tree based machine learning algorithms are widely accepted for the related data mining task because of their robustness in errors and comprehensibility respectively. In order to confirm the diversity and size in data are important factors in the performance of machine learning algorithms experimentally, multilayer perceptron (MLP) and random forest algorithms are selected and used. A real world data set called breast tissue which consists of six classes and continuous attributes only was used. Over-sampling algorithm called SMOTE was applied for all classes in the data set to generate more diverse data. But, some of the over-sampled data may have wrong class values due to the property of the algorithm, so that those data instances may hinder our data mining task, because these instances may cause lower accuracy of our final data mining models. In order to avoid such things happen MLP and random forest are used to check the class of each training instances, and the class values were changed if the two algorithms generate the same result for the artificial instances whose original class values are different from the classification of the two algorithms. Extensive experiments using breast tissue data set in various over-sampling rates showed very good results.

***Keywords*—**Breast tissue**,** Random forest**,** MLP**,** electrical impedance spectroscopy, over-sampling.

I. Introduction

ELECTRICAL impedance spectroscopy(EIS) is a technique to collect pathological data from breast tissue [1]. A lot of researchers reported the utility of the method [2, 3], and many data mining algorithms have been suggested to find more accurate classification models [4, 5, 6]. But, no matter how we apply different data mining algorithms, the performance of built data mining models is limited by the quality and size of data set itself [7]. So, we may have the question on how we may improve the performance of the built data mining models with limited data sets. Being related to this matter recent world event in go games between human and machine give us some insight about the relationship between the performance of machine learning algorithms and the composition of data.

H. Sug is with the Division of Computer Engineering, Dongseo University, Busan, 617-716 Korea (phone: +82-51-320-1733; fax: +82-51-327-8955; e-mail: sht@ gdsu.dongseo.ac.kr). .

AlphaGo and AlphaGo Zero are two successful applications of machine learning algorithms called reinforcement learning and deep learning [8] that are based on very two different sources of data. While AlphaGo used the records of sixteen hundred thousand go games for its training, AlphaGo Zero used only the rules of go game and massive random generation of data [9]. It has been reported that AlphaGo Zero defeated AlphaGo in all games. Even though the slight difference in used hardware, the main reason for the winning of AlphaGo Zero against AlphaGo is that the characteristics of data themselves that were used for the training. While the supplied data to AlphaGo is real go game data records generated by human beings, AlphaGo Zero was trained using massive random data, so that it could be exposed to more novel data abundantly. In other words, the spectrum of training data is better for AlphaGo Zero. When a machine learning algorithm is trained with more diverse training data, the trained model could cover more future unseen cases, so that the performance of the model could become better like we can see from AlphaGo Zero's case. The second success point of AlphaGo Zero is that we know the rules of movement of go game well so that illegal movements can be checked easily, when we generate the random movement.

There are many domains that need to be data mined, where the available data is not good enough to build more accurate data mining models [10, 11], and there are almost no domains that have such rules of go game to check whether the generated random data can belong to the domain. Therefore, this paper suggests how to build more accurate data mining models in such situation, especially for the EIS of breast tissue data set. This paper is extension of the paper presented at CSCC2018 [12]. In section 2 related work is provided, in section 3 we discuss our experiment method, and in section 4 conclusions are provided.

II. RELATED WORK

Deep learning algorithms of artificial neural networks are known for their good performance, so that neural networks can be most preferred machine learning algorithms in data mining [13]. The performance of neural network algorithms are relatively stable compared to other machine learning algorithms, because the effect of training data is distributed evenly as weight in the structure of the network. As a result, their performances are affected less by the perturbation of training data than other less stable algorithms like decision trees. On the other hand, good point of decision tree is comprehensibility, because we can understand the tree structure easily. But, because decision tree

algorithms try to divide training data decisively in the structure, they are more sensitive to the quality of the data [14]. In order to surmount such problem in single decision tree, random forest has been developed [15]. It's been known that the performance of random forest is similar to that of neural networks. In this paper we want to generate more artificial data that will be used to build more accurate classifiers, so SMOTE algorithm [16] can be a good candidate, because it generates artificial data for a minority class based on k-nearest neighbours and random linear interpolation. But, SMOTE cannot guarantee the correctness of the artificial data.

Many researchers have been interested in building more accurate classifiers for the breast tissue data set. In [2] the principle on how data can be measured from breast tissue by electrical impedance tomography is described. As a sequel in [3] in order to analyse the data set linear discriminant analysis based technique was used. They reported accuracy of 66.37%. In [4] feedforward network using the backpropagation learning algorithm and radial basis function network were used to get better classification models, and reported that they achieved accuracy between 83.33% and 91.66%. Other machine learning algorithms like SVM and rough set based method was used also resulting in slightly less or similar accuracies than the above numbers [5, 6].

III. EMPIRICAL PROCEDURE

A. Experiment Method

Because we are interested in generating a data mining model of accuracy, we use multilayer perceptron (MLP) as a deep learning algorithm. We also use random forest because the performance of the forest is comparable to that of MLP. We also generate decision tree of C4.5 because it is one of mostly used data mining algorithm in medicine domain [17, 18, 19] because of its understandability. For our experiment we need randomness and diversity in the training data, so an artificial data generation algorithm like SMOTE could be good for our purpose. SMOTE algorithm was invented as a method of over-sampling for a minority class, because the minority class usually does not have enough training instances for accurate classifiers. SMOTE generates new data instances based on randomization in the linear interpolation of nearest neighbours in existing instances, so the algorithm satisfies our two purposes; novelty by the randomization in linear interpolation of neighbours and increase of training data set size by the generation of new artificial data. The correctness of new artificial data instances are checked by the two machine learning algorithms, MLP and random forest that are trained by the original data set. If the class of the artificial instance is different from the classification result of the two algorithms and the classification result is the same, then the class value is changed to the class value by the two algorithms.

B. Breast Tissue Data Set

For our experiment, a data set called 'Breast Tissue' from UCI machine learning depository [20] is used. The data set was chosen because all of its attributes are continuous attributes so that it's advantageous for the linear interpolation. Breast tissue data has 9 conditional attributes and one decision attribute. The decision attribute has 6 different class values which classify breast tissue depending on the values in conditional attributes. All the attributes in the data set have real values and no missing values in the attributes. Table 1 describes the attributes.

TABLE I
ATTRIBUTE INFORMATION

No.	attribute	meaning
1	I0	Impedivity at zero frequency
2	PA500	Phase angle at 500KHz
3	HFS	High frequency slope of phase angle
4	DA	Impedance distance between spectral ends
5	AREA	Area under spectrum
6	A/DA	Area normalized by DA
7	MAX IP	Maximum of the spectrum
8	DR	Distance between I0 and real part of the maximum frequency point
9	P	Length of the spectral curve
10	Class	Six classes like car(carcinoma), fad(fibro-adenoma), mas(mastopathy), gla(glandular), con(connective), adi(adipose)

Table 2 shows the property of each attributes with respect to minimum, maximum, mean, and standard deviation, where the total number of instances is 106.

TABLE II
THE PROPERTY OF EACH ATTRIBUTE

attribute	min	max	mean	StdDev
I0	103	2800	784.252	753.95
PA500	0.812	0.358	0.12	0.069
HFS	-0.066	0.468	0.115	0.101
DA	19.648	1063.441	190.509	190.801
AREA	70.426	174480.476	7335.155	18580.314
A/DA	1.596	164.072	23.474	23.355
MAX IP	7.969	436.1	75.381	81.346
DR	-9.258	977.552	166.711	181.31
P	124.979	2896.582	810.638	763.019

Table 3 shows the number of instances for each class.

TABLE III
THE NUMBER OF INSTANCES FOR EACH CLASS

Class#	1	2	3	4	5	6
Class name	car	fad	mas	gla	con	adi
# of instances	21	15	18	16	14	22

As deep learning neural network, ensemble-based algorithm, and decision tree algorithm, MLP, random forest, and C4.5 were used respectively. Table 4 shows the accuracy of the algorithms for the data set. In the experiment 10 fold cross-validation was used. For MLP epochs of 4000 and the learning rate of 0.08 and the number of hidden layers of 5 were used. For the parameters of random forest the number of candidate attributes to choose in each subtree of each tree is one and the number of generated tree is 500. Default parameters were applied for C4.5.

TABLE IV
THE RESULT FOR THE ORIGINAL DATA SET

algorithm	Accuracy (%)	Kappa statistic
Random Forest	74.5283	0.6923
MLP	72.6415	0.6696
C4.5	66.0377	0.5893

The accuracy by C4.5 is comparable to the result by Silva et el's result which is 66.37% based on linear discriminant analysis [3]. The accuracy achieved by MLP is comparable to the result by Norte's SVM which is 70.598% in 3 fold cross-validation [21]. SVM is known for its ability to get high accuracy in classification task [22]. The true positive rate of random forest for each class of the 6 classes is 0.857, 0.667, 0.444, 0.688, 0.857, and 0.909, and true positive rate of MLP for each class is 0.81, 0.667, 0.556, 0.5, 0.857, and 0.909.

SMOTE based over-sampling was applied to add more new data for each class. For each class over-sampling rate of 100% ~ 800% and 1600% are applied with the parameter of 5 nearest neighbours. Table 5 shows the change of the number of data instances for each class after the over-sampling.

TABLE V
OVER-SAMPLING BY SMOTE

Over-sampling rate	car	fad	mas	gla	con	adi	Total # of insta-nces
100%	42	30	36	32	28	44	212
200%	63	45	54	48	42	66	318
300%	84	60	72	64	56	88	424
400%	105	75	90	80	70	110	530
500%	126	90	108	96	84	132	636
600%	147	105	126	112	98	154	742
700%	168	120	144	128	112	176	848
800%	189	135	162	144	126	198	954
1600%	357	255	306	272	238	374	1802

After generating over-sampled data set by SMOTE, the class of each artificial data instance has been checked by the random forest and MLP from the original data set. Because SMOTE may generate artificial instances that have wrong class value, we have changed the class of artificial instances of SMOTE, when the classification result of both of random forest and MLP are the same. Table 6 shows the number of class value-changed instances for each over-sampling rate.\

TABLE VI
THE NUMBER OF CLASS VALUE-CHANGED INSTANCES FOR EACH OVER-SAMPLING RATE

Over-sampling rate	Total # of instances	Total # of class valued-changed instances
100%	212	3
200%	318	14
300%	424	22
400%	530	32
500%	636	32
600%	742	35
700%	848	47
800%	954	54
1600%	1802	110

Table VII-1 to tableVII-9 compares the two methods for the three machine learning algorithms for each over-sampling rate. Table VII-1 compares the suggested method that modifies the class values generated by SMOTE with the original method by SMOTE when over-sampling rate for each class is 100%.

TABLE VII-1
THE RESULT FOR OVER-SAMPLING RATE 100%

algorithm		Accuracy (%)	Kappa statistic
Random Forest	Suggested	89.1509	0.8689
	Original	88.6792	0.8633
MLP	Suggested	81.6038	0.7775

	Original	78.7736	0.7434
C4.5	Suggested	82.5472	0.7893
	Original	82.5472	0.7896

Table VII-2 compares the suggested method that modifies the class values generated by SMOTE with the original method by SMOTE when over-sampling rate for each class is 200%.

TABLE VII-2
THE RESULT FOR OVER-SAMPLING RATE 200%

algorithm		Accuracy (%)	Kappa statistic
Random Forest	Suggested	91.195	0.8934
	Original	88.9937	0.8672
MLP	Suggested	83.9623	0.806
	Original	81.4465	0.7757
C4.5	Suggested	87.1069	0.8443
	Original	85.5346	0.8255

Table VII-3 compares the suggested method that modifies the class values generated by SMOTE with the original method by SMOTE when over-sampling rate for each class is 300%.

TABLE VII-3
THE RESULT FOR OVER-SAMPLING RATE 300%

algorithm		Accuracy (%)	Kappa statistic
Random Forest	Suggested	92.9545	0.9114
	Original	92.6887	0.9117
MLP	Suggested	83.2547	0.7975
	Original	80.8962	0.7695
C4.5	Suggested	88.6792	0.8632
	Original	88.6792	0.8633

Table VII-4 compares the suggested method that modifies the class values generated by SMOTE with the original method by SMOTE when over-sampling rate for each class is 400%.

TABLE VII-4
THE RESULT FOR OVER-SAMPLING RATE 400%

algorithm		Accuracy (%)	Kappa statistic
Random Forest	Suggested	94.9057	0.9382
	Original	93.0189	0.9157
MLP	Suggested	88.6792	0.8628

	Original	84.1509	0.8085
C4.5	Suggested	90.3774	0.8835
	Original	88.6792	0.8634

Table VII-5 compares the suggested method that modifies the class values generated by SMOTE with the original method by SMOTE when over-sampling rate for each class is 500%.

TABLE VII-5
THE RESULT FOR OVER-SAMPLING RATE 500%

algorithm		Accuracy (%)	Kappa statistic
Random Forest	Suggested	95.1258	0.941
	Original	93.7107	0.9241
MLP	Suggested	88.8365	0.8648
	Original	84.7484	0.8158
C4.5	Suggested	91.0377	0.8916
	Original	90.0943	0.8806

Table VII-6 compares the suggested method that modifies the class values generated by SMOTE with the original method by SMOTE when over-sampling rate for each class is 600%.

TABLE VII-6
THE RESULT FOR OVER-SAMPLING RATE 600%

algorithm		Accuracy (%)	Kappa statistic
Random Forest	Suggested	95.0135	0.9397
	Original	94.8787	0.9382
MLP	Suggested	89.3531	0.8712
	Original	86.7925	0.8407
C4.5	Suggested	91.3747	0.8958
	Original	89.8922	0.878

Table VII-7 compares the suggested method that modifies the class values generated by SMOTE with the original method by SMOTE when over-sampling rate for each class is 700%.

TABLE VII-7
THE RESULT FOR OVER-SAMPLING RATE 700%

algorithm		Accuracy (%)	Kappa statistic
Random Forest	Suggested	95.4009	0.9443
	Original	93.8679	0.926
MLP	Suggested	88.7972	0.8643

	Original	82.783	0.7922
C4.5	Suggested	91.2736	0.8945
	Original	88.5613	0.862

Table VII-8 compares the suggested method that modifies the class values generated by SMOTE with the original method by SMOTE when over-sampling rate for each class is 800%.

TABLE VII-8
THE RESULT FOR OVER-SAMPLING RATE 800%

algorithm		Accuracy (%)	Kappa statistic
Random Forest	Suggested	94.9686	0.939
	Original	95.5975	0.9469
MLP	Suggested	88.8889	0.8654
	Original	85.6394	0.8266
C4.5	Suggested	92.2432	0.9062
	Original	90.2516	0.8824

Table VII-9 compares the suggested method that modifies the class values generated by SMOTE with the original method by SMOTE when over-sampling rate for each class is 1600%.

TABLE VIII-9
THE RESULT FOR OVER-SAMPLING RATE 1600%

algorithm		Accuracy (%)	Kappa statistic
Random Forest	Suggested	95.3385	0.9436
	Original	97.0033	0.9638
MLP	Suggested	90.788	0.8886
	Original	87.3374	0.8473
C4.5	Suggested	92.3973	0.908
	Original	92.7303	0.9123

Figure 1 to figure 3 shows the accuracy and kappa statistic for three machine learning algorithms for each class as over-sampling rate increases. Figure 1 is the graphs for random forest.

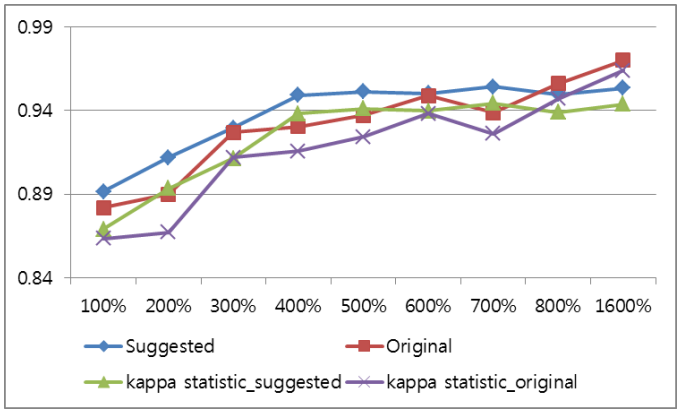

Fig. 1 accuracy and kappa statistic of random forest for each over-sampling rate

Figure 2 is the graphs for MLP.

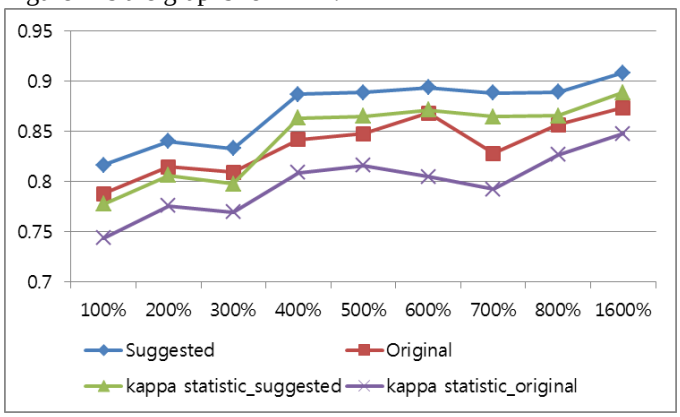

Fig. 2 accuracy and kappa statistic of MLP for each over-sampling rate

Figure 2 is the graphs for decision tree C4.5.

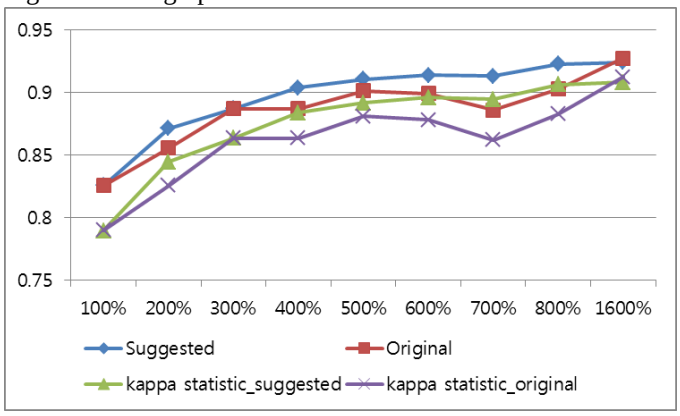

Fig. 3 accuracy and kappa statistic of decision tree C4.5 for each over-sampling rate

Because our suggested method uses random forest and MLP to check and change the class of the artificial data, we compare true positive rate, false positive, ROC area of our suggested method and the original method by SMOTE also as over-sampling rate changes. Figure 4-1 to 4-6 compares the true positive (TP) rate and ROC area of the original method and our suggested method in random forest as over-sampling rate changes. Figure 4-1 is TP rate and ROC area of the original method and suggested method in random forest for class 1 as

over-sampling rate changes.

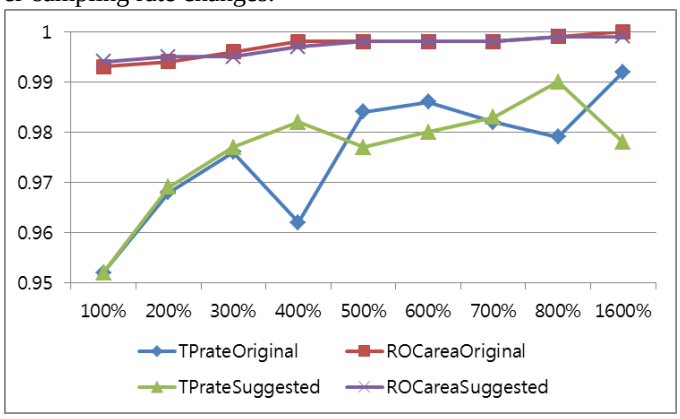

Fig. 4-1 TP rate and ROC area of the original method and suggested method in random forest for class 1 as over-sampling rate changes

Figure 4-2 is TP rate and ROC area of the original method and suggested method in random forest for class 2 as over-sampling rate changes.

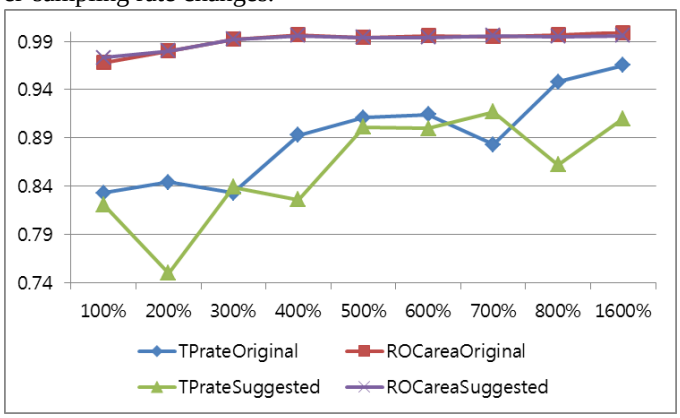

Fig.4-2 TP rate and ROC area of the original method and suggested method in random forest for class 2 as over-sampling rate changes

Figure 4-3 is TP rate and ROC area of the original method and suggested method in random forest for class 3 as over-sampling rate changes.

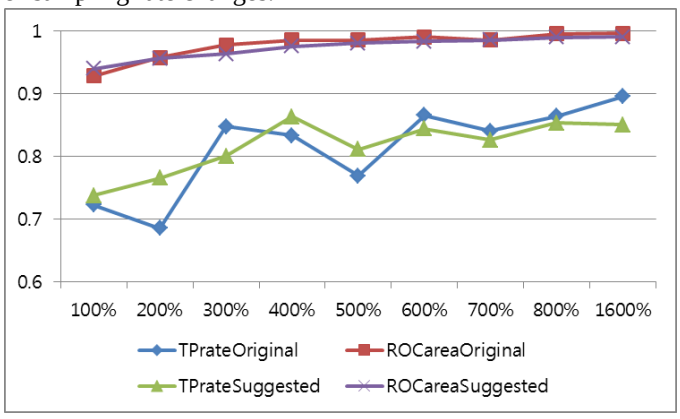

Fig. 4-3 TP rate and ROC area of the original method and suggested method in random forest for class 3 as over-sampling rate changes

Figure 4-4 is TP rate and ROC area of the original method and suggested method in random forest for class 4 as over-sampling rate changes.

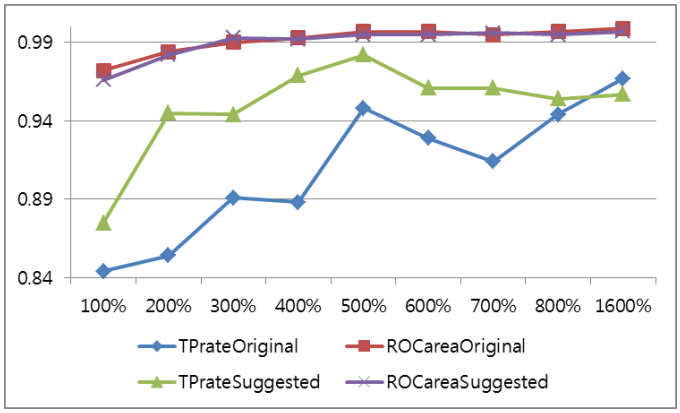

Fig. 4-4 TP rate and ROC area of the original method and suggested method in random forest for class 4 as over-sampling rate changes

Figure 4-5 is TP rate and ROC area of the original method and suggested method in random forest for class 5 as over-sampling rate changes.

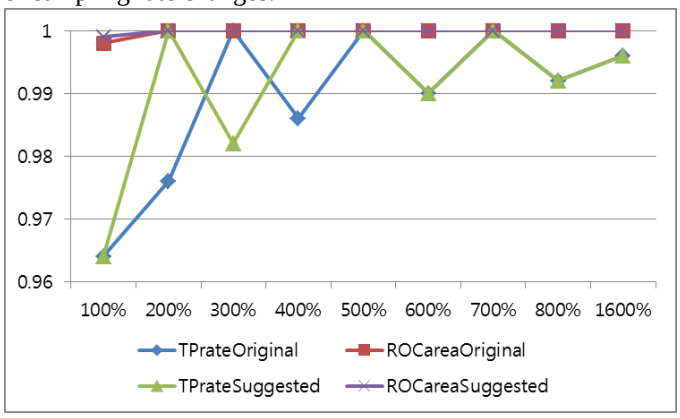

Fig. 4-5 TP rate and ROC area of the original method and suggested method in random forest for class 5 as over-sampling rate changes

Figure 4-6 is TP rate and ROC area of the original method and suggested method in random forest for class 6 as over-sampling rate changes.

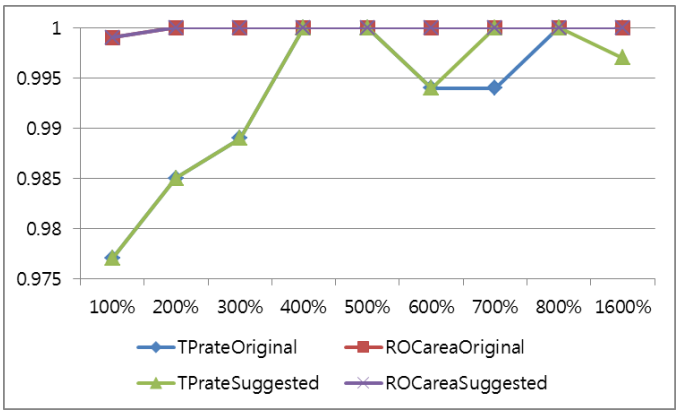

Fig. 4-6 TP rate and ROC area of the original method and suggested method in random forest for class 6 as over-sampling rate changes

Figure 5-1 to 5-6 compares the true positive (TP) rate and ROC area of the original method and suggested method in MLP as over-sampling rate changes. Figure 5-1 is TP rate and ROC

area of the original method and suggested method in MLP for class 1 as over-sampling rate changes.

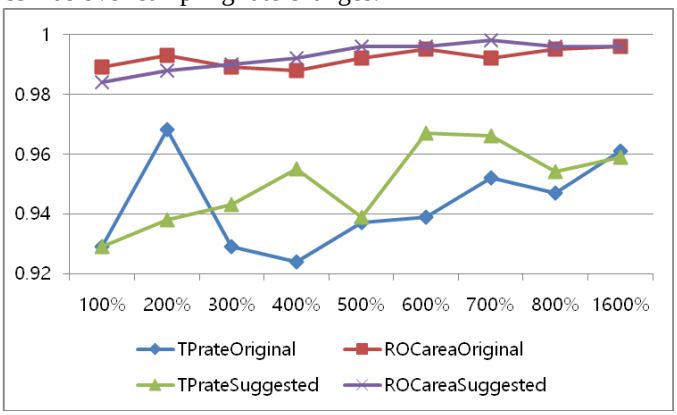

Fig. 5-1 TP rate and ROC area of the original method and suggested method in MLP for class 1 as over-sampling rate changes

Figure 5-2 is TP rate and ROC area of the original method and suggested method in MLP for class 2 as over-sampling rate changes.

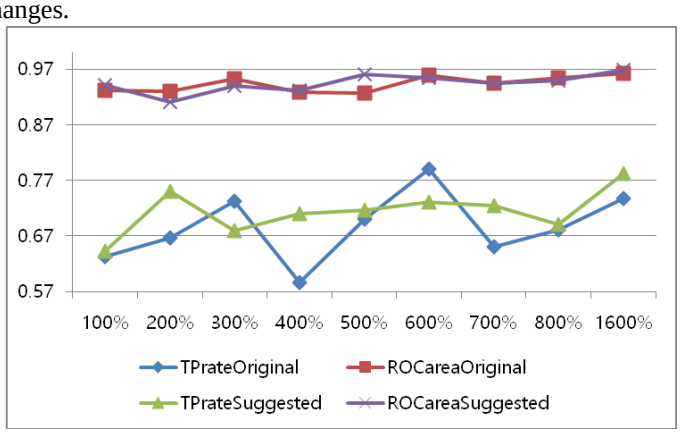

Fig. 5-2 TP rate and ROC area of the original method and suggested method in MLP for class 2 as over-sampling rate changes

Figure 5-3 is TP rate and ROC area of the original method and suggested method in MLP for class 3 as over-sampling rate changes.

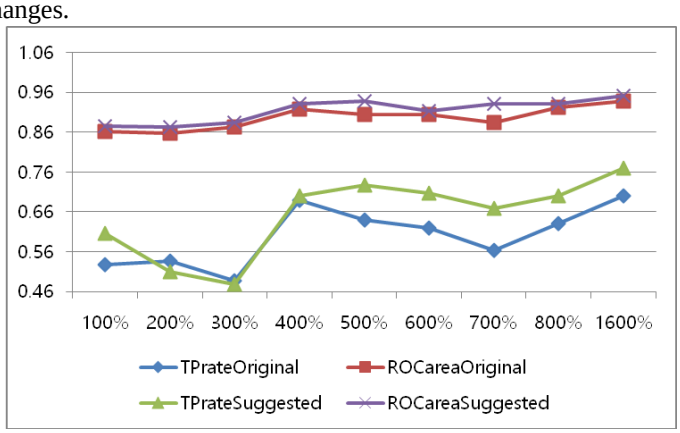

Fig. 5-3 TP rate and ROC area of the original method and suggested method in MLP for class 3 as over-sampling rate changes

Figure 5-4 is TP rate and ROC area of the original method and suggested method in MLP for class 4 as over-sampling rate changes.

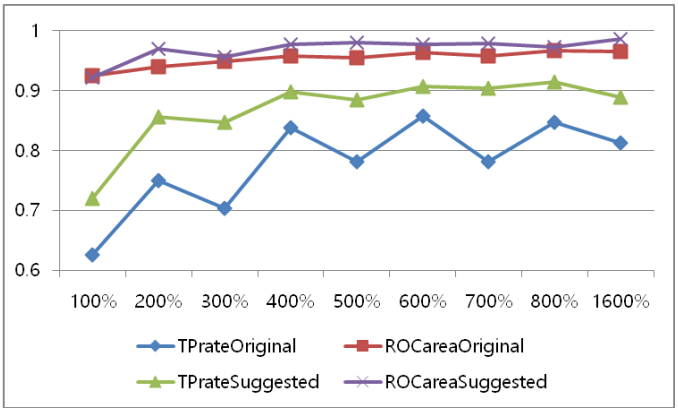

Fig. 5-4 TP rate and ROC area of the original method and suggested method in MLP for class 4 as over-sampling rate changes

Figure 5-5 is TP rate and ROC area of the original method and suggested method in MLP for class 5 as over-sampling rate changes.

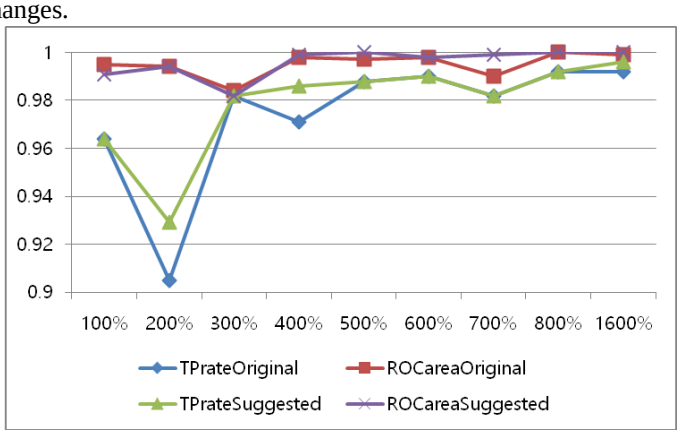

Fig. 5-5 TP rate and ROC area of the original method and suggested method in MLP for class 5 as over-sampling rate changes

Figure 5-6 is TP rate and ROC area of the original method and suggested method in MLP for class 6 as over-sampling rate changes.

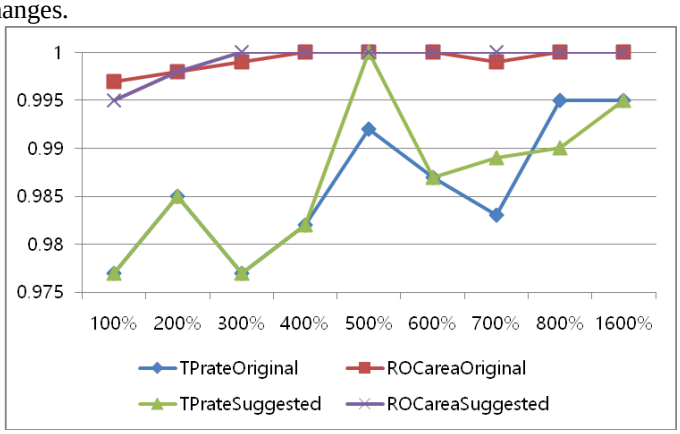

Fig. 5-6 TP rate and ROC area of the original method and suggested method in MLP for class 6 as over-sampling rate changes

We can also think of the performance with respect to false positive rate. Figure 6-1 to 6-6 compares the false positive (FP) rate of the original method and suggested method in random forest as over-sampling rate changes. Figure 6-1 is FP rate of the original method and suggested method in random forest for

class 1 as over-sampling rate changes.

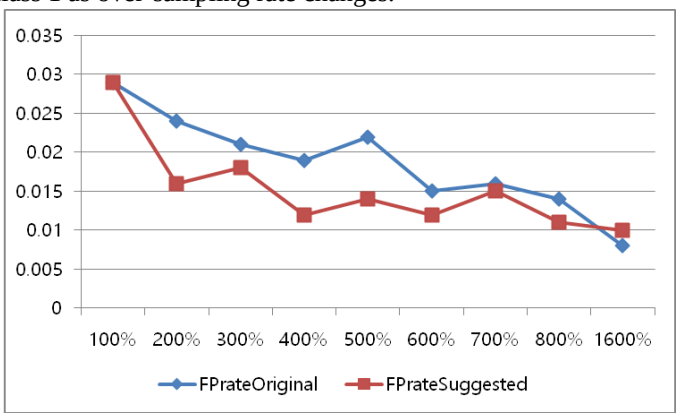

Fig. 6-1 FP rate of the original method and suggested method in random forest for class 1 as over-sampling rate changes

Figure 6-2 is FP rate of the original method and suggested method in random forest for class 2 as over-sampling rate changes.

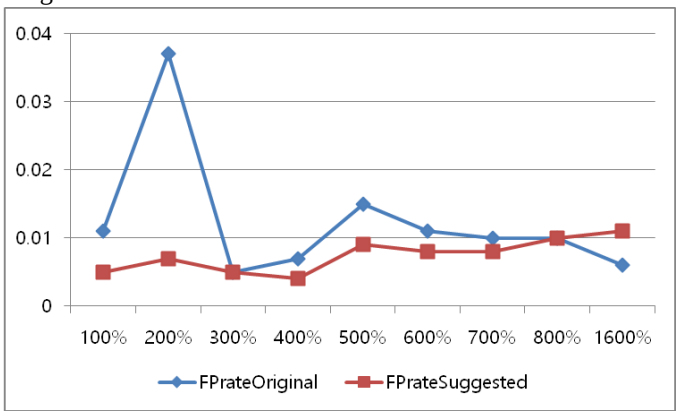

Fig. 6-2 FP rate of the original method and suggested method in random forest for class 2 as over-sampling rate changes

Figure 6-3 is FP rate of the original method and suggested method in random forest for class 3 as over-sampling rate changes.

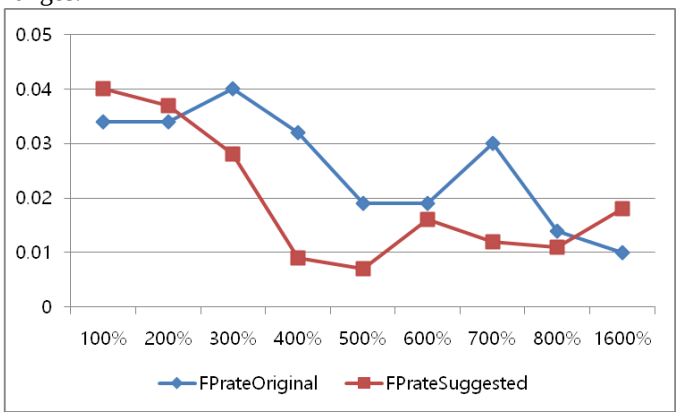

Fig. 6-3 FP rate of the original method and suggested method in random forest for class 3 as over-sampling rate changes

Figure 6-4 is FP rate of the original method and suggested method in random forest for class 4 as over-sampling rate changes.

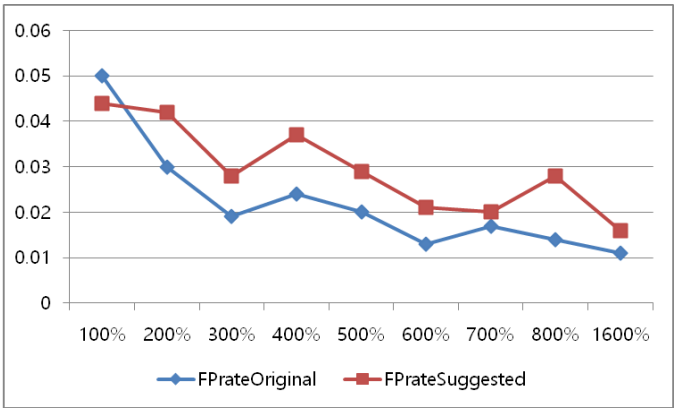

Fig. 6-4 FP rate of the original method and suggested method in random forest for class 4 as over-sampling rate changes

Figure 6-5 is FP rate of the original method and suggested method in random forest for class 5 as over-sampling rate changes.

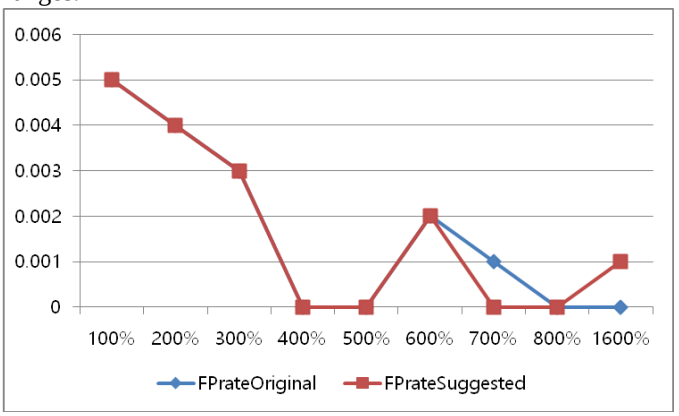

Fig. 6-5 FP rate of the original method and suggested method in random forest for class 5 as over-sampling rate changes

Figure 6-6 is FP rate of the original method and suggested method in random forest for class 6 as over-sampling rate changes.

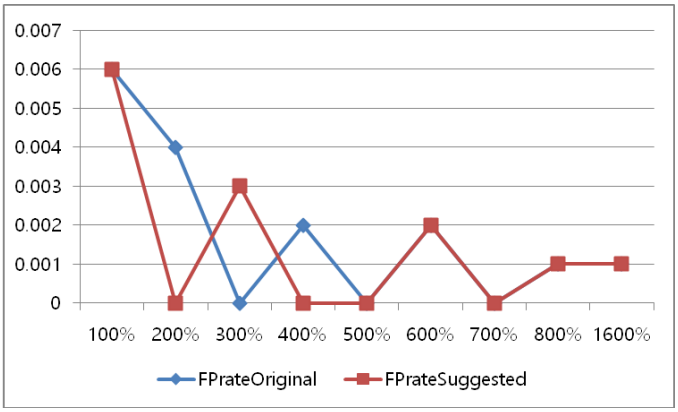

Fig. 6-6 FP rate of the original method and suggested method in random forest for class 6 as over-sampling rate changes

Figure 7-1 to 7-6 compares the false positive (FP) rate of the original method and suggested method in MLP as over-sampling rate changes. Figure 7-1 is FP rate of the original method and suggested method in MLP for class 1 as

over-sampling rate changes.

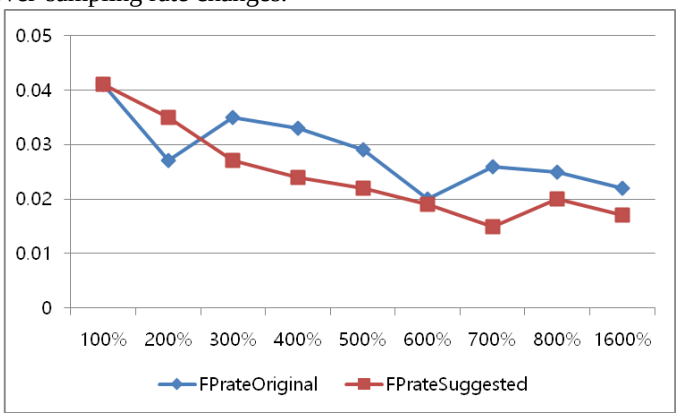

Fig. 7-1 FP rate of the original method and suggested method in MLP for class 1 as over-sampling rate changes

Figure 7-2 is FP rate of the original method and suggested method in MLP for class 2 as over-sampling rate changes.

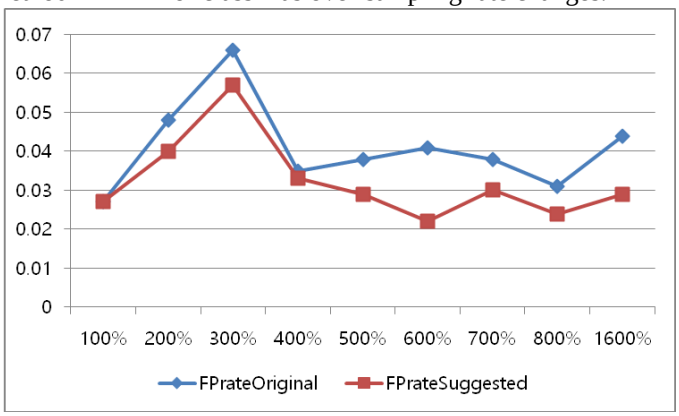

Fig. 7-2 FP rate of the original method and suggested method in MLP for class 2 as over-sampling rate changes

Figure 7-3 is FP rate of the original method and suggested method in MLP for class 3 as over-sampling rate changes.

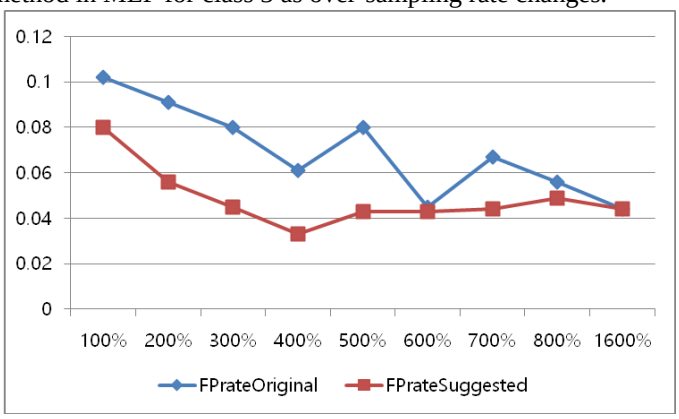

Fig. 7-3 FP rate of the original method and suggested method in MLP for class 3 as over-sampling rate changes

Figure 7-4 is FP rate of the original method and suggested method in MLP for class 4 as over-sampling rate changes.

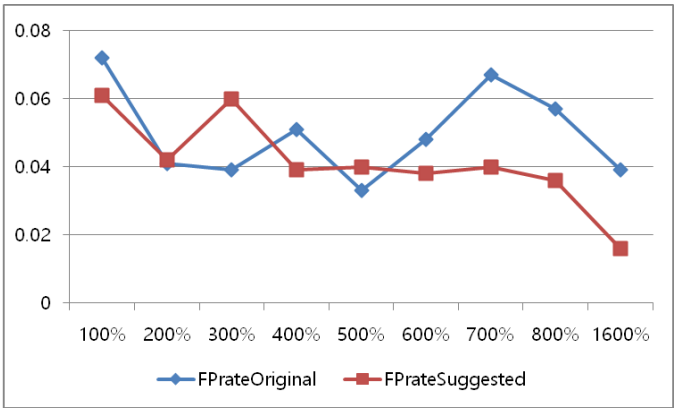

Fig. 7-4 FP rate of the original method and suggested method in MLP for class 4 as over-sampling rate changes

Figure 7-5 is FP rate of the original method and suggested method in MLP for class 5 as over-sampling rate changes.

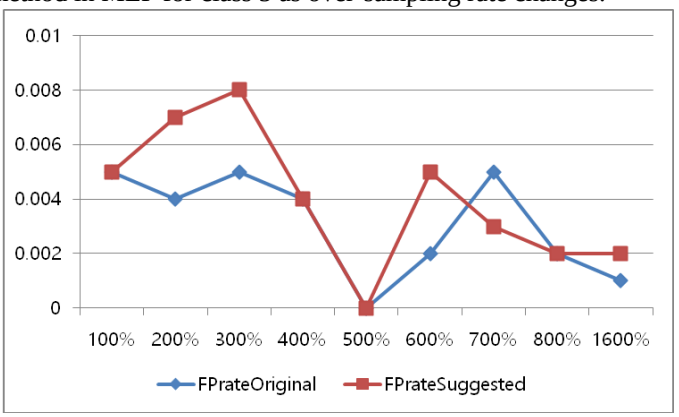

Fig. 7-5 FP rate of the original method and suggested method in MLP for class 5 as over-sampling rate changes

Figure 7-6 is FP rate of the original method and suggested method in MLP for class 6 as over-sampling rate changes.

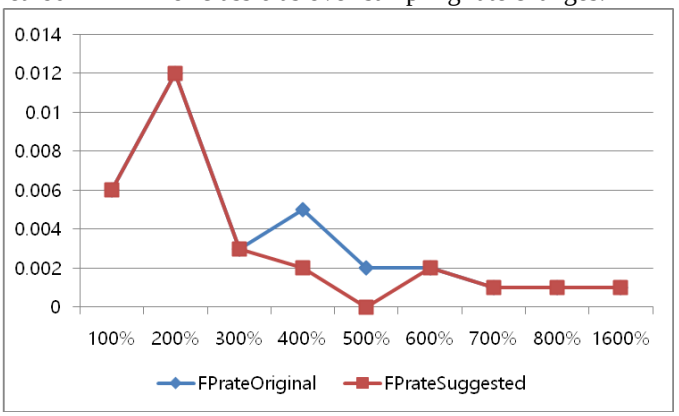

Fig. 7-6 FP rate of the original method and suggested method in MLP for class 6 as over-sampling rate changes

All in all we can confirm that we can generate better data mining models with our suggested method than the original SMOTE only method, especially when the over-sampling rate is not too high.

IV. Conclusion

A machine go player called AlphaGo Zero reminds us the fact of amazing progress of machine learning technologies. When AlhpaGo Zero is playing go game, it uses massive random moves that were checked by the rules of go game, resulting in total victory at every go games. The random moves supplied more novel training data set to the program, which implies the fact that the diversity and size in data are very important for the success of the machine learning algorithms. Neural networks and decision tree based machine learning algorithms are widely accepted in data mining because of their robustness in errors and comprehensibility respectively. In order to confirm the diversity and size in data are important factors in the performance of machine learning algorithms experimentally, MLP and random forest algorithms are selected and used. A real world data set called breast tissue which consists of continuous attributes only was used. As a way to prepare more diverse data over-sampling algorithm called SMOTE that is based on random linear interpolation of neighbours was applied for all classes in the data set, and the correctness of the class of new instances of over-sampling was checked, and modified if necessary. In other words, if the two machine learning algorithms classify an artificial instance differently from its own class value and the classification results of the two algorithms are the same, then the class of the instance is changed for the new instances to have the classified value of the two algorithms. Experiment showed that by supplying the class corrected data instances to target machine learning algorithms for training, the accuracy of the machine learning algorithms becomes generally better as the size and diversity of data increase.

References

[1] M. Grossi, B. Riccò, "Electrical impedance spectroscopy (EIS) for biological analysis and food characterization: a review", *Journal of Sensors and Sensor Systems*, vol.6, 2017, pp.303-325.

[2] J. Jossinet, "Variability of impedivity in normal and pathological breast tissue", *Medical & Biological Engineering & Computing*, vol. 34, 1996, pp. 346-350.

[3] J.E. Silva, J.P.M. Sá, J. Jossinet, "Classification of breast tissue by electrical impedance spectroscopy", , *Medical & Biological Engineering & Computing*, vol. 38, 2000, pp. 26-30.

[4] A. Helwan, J.B. Idoko, R.H. Abiyev, "Machine learning techniques for classification of breast tissue", *Procedia Computer Science*, vol. 120, 2017, pp. 402-410.

[5] L. Chang, C. Tiantian, L. Changxing, "Breast tissue classification based on Electrical Impedance Spectroscopy", *Proceeding of International Conference on Industrial technology and management*, 2015, pp. 237-240.

[6] V.P. Sumathi, K. Kousalya, V. Vanitha, "Rough set based approach for multiclass breast tissue classification", *Asian Journal of Information Technology*, vol. 15, no. 22, 2016, pp. 4438-4444.

[7] J.V. Hulse, *Data quality in data mining and machine learning*, doctoral dissertation, Florida Atlantic University, 2007.

[8] J. Schmidhuber, "Deep learning in neural networks: An overview", *Neural Networks*, vol. 61, 2015, pp.85-117.

[9] D. Silver, J. Schrittwieser,K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, D. Hassabis, "Mastering the game of go without human knowledge", *Nature*, vol. 550, no. 7676, 2017, pp. 354-359.

[10] N.H. Ruparel, N.M. Shahane, D.P. Bhamare, "Learning from small data set to build classification model: a survey", *International Journal of Computer Applications*, 2013, pp.23-26.

[11] R. Andonie, "Extreme data mining: Inference from small datasets", *International Journal of Computers, Communications & Control*, vol. 5, no. 3, 2010, pp. 280-291.

[12] H. Sug, "Performance of machine learning algorithms and diversity in data", MATEC *Web of Conferences*, vol. 210, 04019, 2018.

[13] P. Tan, M. Steinbach, A. Karpatne, V. Kumar, *Introduction to Data Mining*, 2nd ed., Pearson, 2018.

[14] W. Sun, *Stability of machine learning algorithms*, PhD thesis, Purdue University, 2015.

[15] L. Breiman, "Random forests", *Machine Learning*, vol. 45, issue 1, pp. 5-32, 2001.

[16] N.V. Chawla, K.W. Dowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 321-357

[17] L. Rokach, O. Maimon, *Data Mining with Decision Trees: Theories and Applications*, 2nd ed., World Scientific Publishing Company, 2014.

[18] J.R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., 1993

[19] K. Madadipouya, "A new decision tree method for data mining in medicine", *Advanced Computational Intelligence: An International Journal*, vol. 2, no. 3, 2015, pp. 31-37.

[20] A. Frank and A. Suncion, *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Sciences, 2010.

[21] M.W. Nonte, *Classification of Breast Tissue Based on Electrical Impedance Spectroscopy Data*, ECE/CS /ME 539 Introduction to Artificial Neural Networks and Fuzzy System-Fall 2013 Semester Class Projects, University of Wisconsin, 2013.

[22] E.A. Zanaty, "Support vector machines (SVMs) versus Multilayer perceptron (MLP) in data classification", *Egyptian Informatics Journal*, vol. 13, issue 3, 2012, pp. 177-183.

Hyontai Sug received the B.S. degree in Computer Science and Statistics from Busan National University, Busan, Korea, in 1983, the M.S. degree in Computer Science from Hankuk University of Foreign Studies, Seoul, Korea, in 1986, and the Ph.D. degree in Computer and Information Science & Engineering from University of Florida, Gainesville, FL, USA in 1998. He is a professor of the Division of Computer Engineering of Dongseo University, Busan, Korea since 2001. From 1999 to 2001, he was a full time lecturer of Pusan University of Foreign Studies, Busan, Korea. He was also a researcher of Agency for Defense Development, Korea from 1986 to 1992. His areas of research include data mining and database applications.