Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques

Satyabrata Aich[*], Ahmed Abdulhakim Al-Absi[**], Kueh Lee Hui[***], and Mangal Sain[****]

[*]*Department of Computer Engineering, Inje University, South Korea*

[**]*Department of Computer Engineering, Kyungdong University- Global Campus, Gangwondo, South Korea*

[***]*Dept. of Electrical Engineering, Dong-A University, South Korea*

[****]*Department of Computer Engineering, Dongseo University, South Korea*

satyabrataaich@gmail.com, absiahmed@kduniv.ac.kr, leehkueh@dau.ac.kr, mangalsain1@gmail.com

***Abstract*—In recent years, most of the industries promoting their products based on the quality certification they received on the products. The traditional way of assessing the product quality is time consuming, however with the invent of machine learning techniques the processes has become more efficient and consumed less time than before. In this paper we have explored, some of the machine learning techniques to assess the quality of wine based on the attributes of wine that depends on quality. We have used white wine and red wine quality dataset for this research work. We have used different feature selection technique such as genetic algorithm (GA) based feature selection and simulated annealing (SA) based feature selection to check the prediction performance. We have used different performance measure such as accuracy, sensitivity, specificity, positive predictive value, negative predictive value for comparison using different feature sets and different supervised machine learning techniques. We have used nonlinear, linear and probabilistic classifiers. We have found that feature selection-based feature sets able to provide better prediction than considering all the features for performance prediction. We have found accuracy ranging from 95.23% to 98.81% with different feature sets. This analysis will help the industries to access the quality of the products at less time and more efficient way.**

Keywords—machine learning; feature selection; classifiers; performance metrics;wine qulaity

I. INTRODUCTION

In recent years there is a modest increase in the wine consumption as it has been found that wine consumption has a positive correlation to the heart rate variability [1]. With the increase in the consumption wine industries are looking for alternatives to produce good quality wine at less cost. Different wines have different purposes. Although most of the chemicals are same for different type of wine based on the chemical tests, the quantity of each chemicals have different level of concentration for different type of wine. These days it is really important to classify different wine for quality assurance [2]. In the past due to lack of technological resources it become difficult for most of the industries to classify the wines based on the chemical analysis as it takes lot of time and also need more money. These days with the advent of the machine learning techniques it is possible to classify the wines as well as it is possible to figure out the importance of each chemical analysis parameters in the wine and which one to ignore for reduction of cost. The performance comparison with different feature sets will also help to classify it in a more distinctive way. In this paper an intelligent approach is proposed by considering genetic algorithm (GA) based feature selection as well as simulated annealing-based feature selection considering the nonlinear classifiers, linear classifiers and probabilistic classifiers to predict the quality in red wine as well as the white wine.

The structure of the paper is organized as follows: Section II presents the past work related to this field. Section 3 describes about the methodologies used for this research work. Section 4 describes about the result of feature selection as well

Manuscript received January 02, 2018. This work was supported by Kyungdong University grant, 2018 and a follow-up the invited journal to the outstanding paper of the 20th International Conference on Advanced Communication Technology (ICACT 2018).

Satyabrata Aich is with the Department of Computer Engineering, Inje University, South Korea (e-mail: satyabrataaich@gmail.com)

Ahmed Abdulhakim Al-Absi is with the Department of Computer Engineering, Kyungdong University- Global Campus, Gangwondo, South Korea (e-mail: absiahmed@kduniv.ac.kr)

Kueh Lee Hui is with the Department of Electrical Engineering, Dong-A University, South Korea (e-mail: leehkueh@dau.ac.kr)

Mangal Sain is with the Department of Computer Engineering, Dongseo University, South Korea. He is the corresponding author of this paper. (Corresponding author phone: +8251-320-2009; e-mail: mangalsain1@gmail.com).

ISBN 979-11-88428-02-1

as the result of classification. Section 5 describes about the conclusion and future work.

II. RELATED WORKS

In the past few attempts have been made to use different machine learning approaches and feature selection techniques to the wine dataset. Er and Atasoy proposed a method to classify the quality of wines using three different classifier such as support vector machines, Random forest and k-nearest neighborhood. They have used principal component analysis for feature selection and they found good result using Random forest algorithm [3]. Chen et al proposed an approach that will predict the grade of wine using the human savory reviews. They have used hierarchical clustering approach and association rule algorithm to process the reviews and predict the wine grade and they found an accuracy of 85.25% while predicting the grade [4]. Appalasamy et al proposed a method to predict wine quality based on physiochemical test data. They have pointed out that classification approach helps to improve the quality of wine during the production [5]. Beltrán et al proposed an approach to classify the wine based on aroma chromatograms and they have used PCA for dimensionality reduction and wavelet transform for feature extraction and classifiers such as neural network, linear discriminant analysis and support vector machine and found that support vector machine with wavelet transforms perform better than other classifiers [6]. Thakkar et al., used analytical hierarchy process (ahp) to rank the attributes and then used different machine learning classifiers such as support vector machine and random forest and they found accuracy of 70.33% using random forest and 66.54% using SVM [7]. Reddy and Govindarajulu used a user centric clustering approach to recommend the product. They have used red wine data set for the survey purpose. They have allocated relative voting to the attributes based on the literature review. Then they assigned weight to the attributes using Gaussian Distribution Process. They judged the quality based on the user preference group [8]. The above past work motivated us to try different feature selection algorithm as well as different classifiers to compare the performance metrics. This paper proposed GA based feature selection and SA based feature selection and used different classifiers such as PART, RPART, Bagging, C5.0, random forest, svm, lda, naïve bayes etc.

III. METHODOLOGIES

The flow chart of the proposed methodology is shown in the Fig. 1.

A. Data Collection

The wine data set is publicly available in the database of UCI. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. This data set contains the physiochemical variables as well as sensory variables; altogether there are 12 attributes [9]. We have used genetic algorithm (GA)-based feature sets for feature selection. Pledsoe first presented an adaptive optimization search methodology is called genetic algorithm and Holland mathematically presented the genetic algorithm-based approach by getting inspiration from Darwin's theory of evolution. A variable is mentioned as a gene. A chromosome is nothing but a sequence of gene. An initialization is done randomly by using population of chromosome. The quality of the chromosomes is evaluated according to a predefined fitness function. High performance chromosomes are used to produce the offspring. The genetic operators such as mutation and crossover are used to form the offspring. In this process the chromosomes are competing with each other and the fittest one survives at the end. The optimal solution comes after a series of iterative computations. [10, 11].

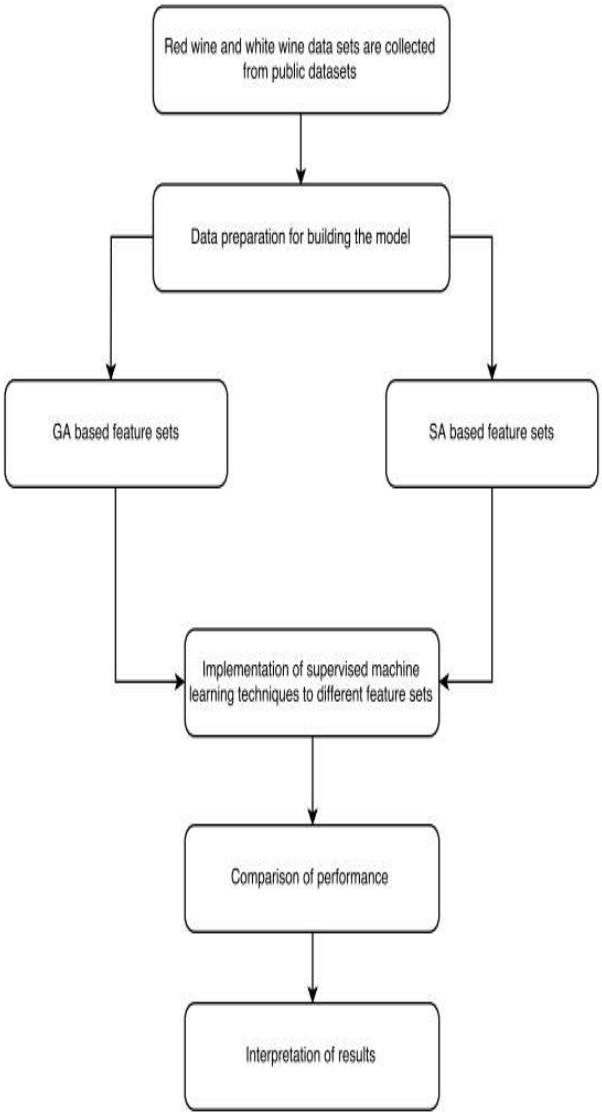

Fig. 1. Flow chart of proposed method

We have also used simulated annealing-based feature set for feature selection. The widely used combinatorial optimization method is called simulated annealing. It is one of the most popular search algorithms. This method used probabilistic technique to find the local optima that ultimately find a better solution [12]. This method is widely used for feature selection method. The simulated algorithm procedure is mentioned below. It runs based on the number of classes. If the number of classes is n then it runs n^{th} times. In each run j, the

ISBN 979-11-88428-02-1

subset of the feature for the j^{th} class is found. All the j^{th} class patterns are taken into one class and other pattern belong to the other class while evaluating the current string. This process helps to give the features which classify patterns as belonging to class j or not class j [13]. After selecting the features by using simulated annealing (SA) and genetic algorithm (GA), we have implemented the data sets into various classifiers and compare the performance parameters.

B. *Performance Measure Metrics*

The parameters used to compare the performance and validations of classifier are as follows: accuracy, sensitivity, specificity, positive predictive value (ppv), negative predictive value (npv). The sensitivity is defined as the ratio of true positives to the sum of true positives and false negatives. The specificity is defined as the ratio of true negatives to the sum of false positives and true negatives. In our research we have used the Positive predictive value and negative predictive value to check the present and absent of one type of wine. So, the ppv is the probability that the one type of wine is present given a positive test result and npv is the probability that the one type of wine is absent given a negative test result [14]. Accuracy is defined as the ratio of number of correct predictions made to the total prediction made and the ratio is multiplied by 100 to make it in terms of percentage.

IV. RESULTS AND DISCUSSION

We have divided the data into two groups such as train data and test data. We trained each classifier based on the trained data and predict the power of classifier on the test data. So, each classifier able to show all the performance metrics such as accuracy, sensitivity, specificity, PPV, and NPV based on the test data. We have applied all the classification techniques to the GA based reduced feature sets for two types of wine as well as SA based reduced feature sets for two types of wine to measures the performance parameter with respect to each classifier. We separated each performance measures with respect to GA and SA sets and plot the column plot for better visualization. The results of each performance measure with respect to two feature sets are shown in the Fig. 2, 3, 4, 5, and 6 respectively for red wine and 7, 8, 9, 10, 11 for white wine.

A. *Comparison of Accuracy for red wine*

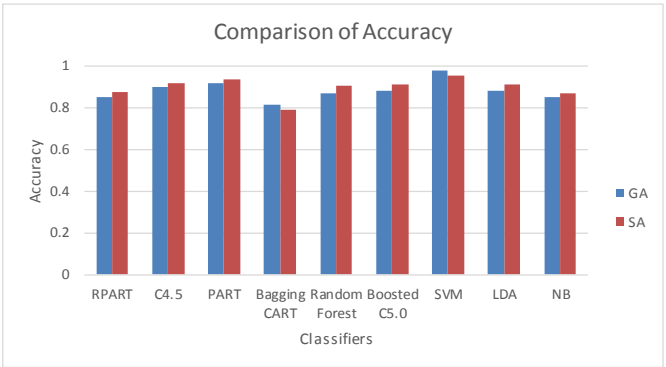

Fig. 2. Comparison of Accuracy on PCA and RFE sets

Fig. 2 show that SVM classifier shows maximum accuracy among all the classifiers. It is performed better with the SA based feature sets. The accuracy of SVM classifier with SA feature set found to be 95. 23%.

B. *Comparison of Sensitivity for red wine*

Fig. 3 shows the sensitivity plot of all the classifiers with two different feature sets. The plot shows SVM has the highest sensitivity compared to others and it was found to be 0.9717 with the SA based feature sets

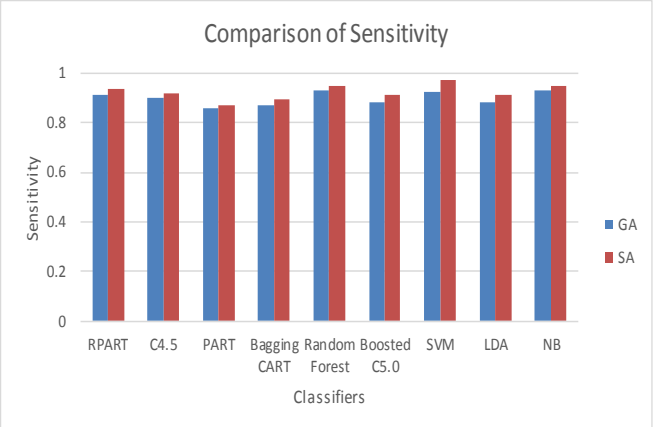

Fig. 3. Comparison of Sensitivity on PCA and RFE sets

C. *Comparison of Specificity for red wine*

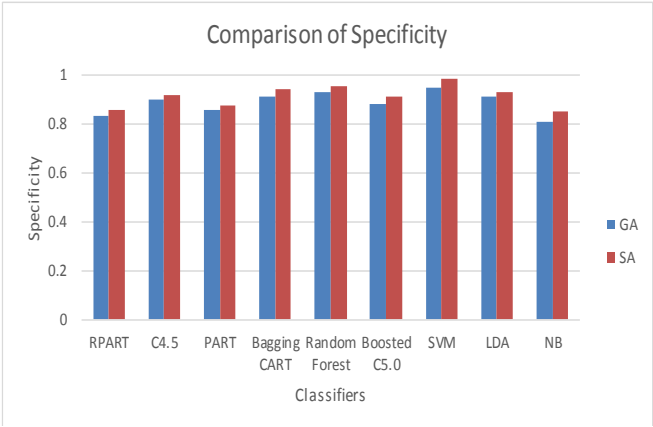

Fig. 4. Comparison of Specificity on PCA and RFE sets

Fig. 4 shows that SVM classifier shows maximum specificity among all the classifiers. It is performed better with the SA based feature sets. The specificity of SVM classifier with SA feature set found to be 0.9835

D. *Comparison of PPV for red wine*

ISBN 979-11-88428-02-1

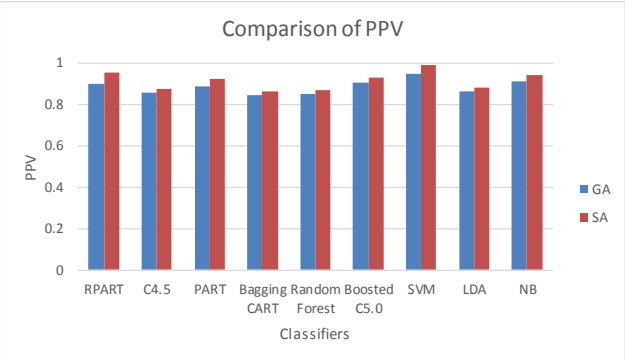

Fig. 5. Comparison of PPV on PCA and RFE sets

Fig. 5 shows the PPV plot of all the classifiers with two different feature sets. The plot shows SVM has the highest PPV compared to others and it was found to be 0.9912 with the SA based feature sets

E. Comparison of NPV for red wine

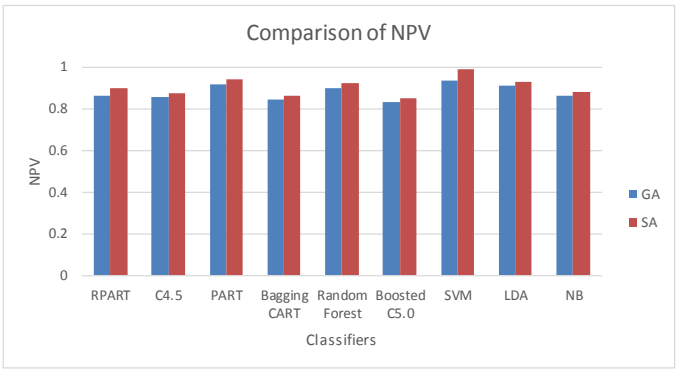

Fig. 6. Comparison of NPV on PCA and RFE sets

Fig. 6 shows that SVM classifier shows maximum NPV among all the classifiers. It is performed better with the SA based feature sets. The NPV of SVM classifier with SA feature set found to be 0.9907

F. Comparison of Accuracy for white wine

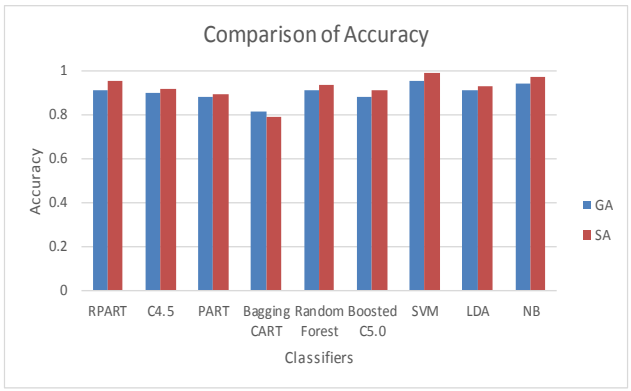

Fig. 7. Comparison of Accuracy on PCA and RFE sets

Fig. 7 shows that SVM classifier shows maximum accuracy among all the classifiers. It is performed better with the SA based feature sets. The accuracy of SVM classifier with SA feature set found to be 98. 81% for white wine data set.

G. Comparison of Sensitivity for white wine

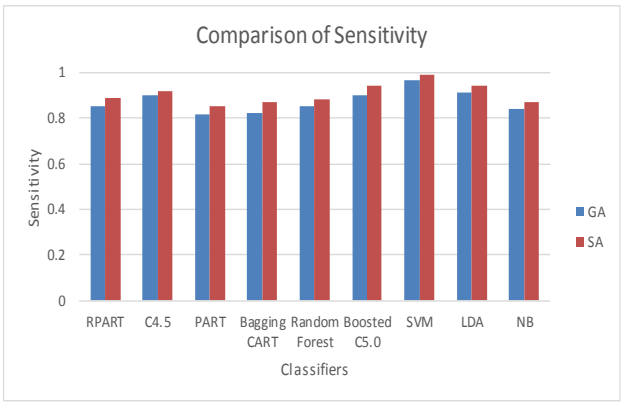

Fig. 8. Comparison of Sensitivity on PCA and RFE sets

Fig. 8 shows the sensitivity plot of all the classifiers with two different feature sets. The plot shows SVM has the highest sensitivity compared to others and it was found to be 0.9934 with the SA based feature sets for white wine data set.

H. Comparison of Specificity for white wine

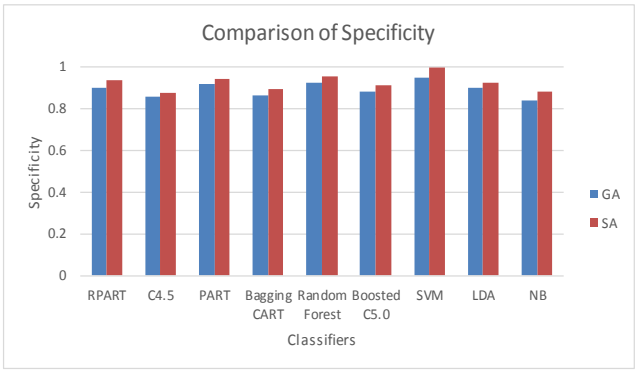

Fig. 9. Comparison of specificity on PCA and RFE sets

Fig. 9 shows that SVM classifier shows maximum specificity among all the classifiers. It is performed better with the SA based feature sets. The specificity of SVM classifier with SA feature set found to be 0.9956 for white wine data set.

I. Comparison of PPV for white wine

ISBN 979-11-88428-02-1

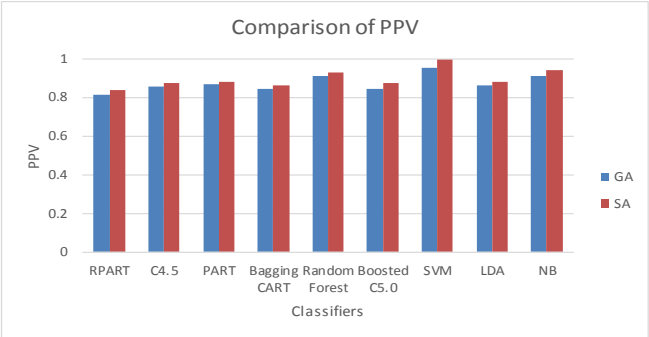

Fig. 10. Comparison of PPVon PCA and RFE sets

Fig. 10 shows the PPV plot of all the classifiers with two different feature sets. The plot shows SVM has the highest PPV compared to others and it was found to be 0.9987 with the SA based feature sets for white wine data set.

J. Comparison of NPV for white wine

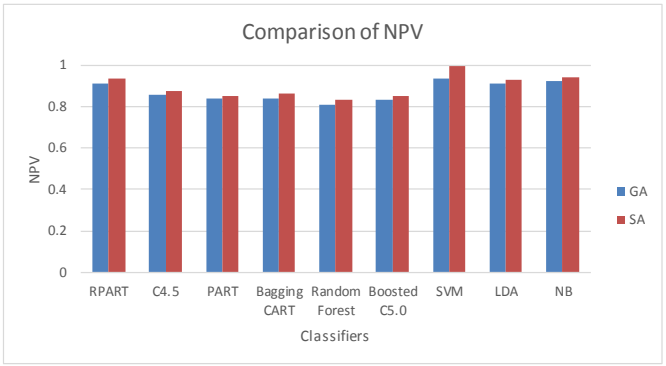

Fig. 11. Comparison of NPV onPCA and RFE sets

Fig. 11 shows that SVM classifier shows maximum NPV among all the classifiers. It is performed better with the SA based feature sets. The NPV of SVM classifier with SA feature set found to be 0.9992 for white wine data set.

The above plots show the performance metrics comparison of different type of wines based on the metrics parameters such as accuracy, sensitivity, specificity, ppv and npv on two different feature sets. The result shows that the SVM classifier performs better for both type of data sets. Specially it is performing better in SA based feature sets. Although it is easy to say based on our result that Simulated Annealing is better algorithm for feature selection compared to genetic algorithm-based feature selection method, however the result could be different for other datasets as well as it could be different for bigger datasets. Similarly based on our result we can say that SVM classifier is best, but in practical lot of other parameters also come into picture that could change the scenario completely. This analysis will give a clear idea about the important attributes for the prediction of quality as well as it saves lot of time and money for the industries.

V. CONCLUSION AND FUTURE WORK

This paper mentioned about potential of genetic algorithm as well as simulated annealing algorithm for feature selection as well as the potentials of the classifiers to predict accurately based on the new feature sets. The feature selection algorithm provided a clear idea about the importance of the attributes for prediction of quality, which was time consuming and expensive when done in the traditional way. We have also compared the performance metrics of linear, nonlinear, and probabilistic based classifiers and it was found that these classifiers performed well with the new feature sets. We have found that the SA based feature sets performed better than the GA based feature sets. We have also found that the SVM classifier performed better compared to all other classifiers for red wine and white wine data sets. We have found accuracy ranging from 95.23% to 98.81% with different feature sets. In future we can try other performance measures and other machine learning techniques for better comparison on results. This analysis will help the industries to predict the quality of the different type of wines based on certain attributes and also it will helpful for them to make good product in the future.

REFERENCES

[1] I.Janszky, M.Ericson, M.Blom, A. Georgiades, J.O.Magnusson, H.Alinagizadeh, and S.Ahnve, "Wine drinking is associated with increased heart rate variability in women with coronary heart disease," *Heart,* 91(3), pp.314-318,2005.

[2] V. Preedy, and M. L. R. Mendez, "Wine Applications with Electronic Noses," in Electronic Noses and Tongues in Food Science, Cambridge, MA, USA: *Academic Press*, 2016, pp. 137-151.

[3] Y.Er, and A.Atasoy, "The Classification of White Wine and Red Wine According to Their Physicochemical Qualities,"*International Journal of Intelligent Systems and Applications in Engineering*,4,pp.23-26,2016.

[4] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, "Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel," *IEEE International Conference on Data Mining Workshop*, pp. 142-149, Dec. 2014.

[5] P.Appalasamy, A.Mustapha, N.D.Rizal, F.Johari, and A.F.Mansor, "Classification-based Data Mining Approach for Quality Control in Wine Production*," Journal of Applied Sciences*, 12(6), pp.598-601,2012

[6] N. H. Beltran, M. A. Duarte- MErmound, V. A. S. Vicencio, S. A. Salah, and M. A. Bustos, "Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer," Instrum. Measurement, *IEEE Trans*., 57: 2421-2436, 2008.

[7] K.Thakkar,J.Shah,R.Prabhakar,A.Narayan,A.Joshi, "AHP and MACHINE LEARNING TECHNIQUES for Wine Recommendations," *International Journal of Computer Science and Information Technologies*,7(5),pp. 2349-2352 ,2016

[8] Reddy, Y. S., & Govindarajulu, P. (2017). An Efficient User Centric Clustering Approach for Product Recommendation Based on Majority Voting: A Case Study on Wine Data Set. *IJCSNS,* 17(10), 103.

[9] M.Forina, R. Leardi, C. Armanino, and S. Lanteri, "PARVUS An Extendible Package for Data Exploration," *Classification and Correla*,1988.

[10] Bledsoe, W. W. (1961). The use of biological concepts in the analytical study of systems. In the ORSA-*TIMS National Meeting*

[11] Holland, J. H. (1992). Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. *MIT press.*

[12] Jeong, I. S., Kim, H. K., Kim, T. H., Lee, D. H., Kim, K. J., & Kang, S. H. (2018). A Feature Selection Approach Based on Simulated Annealing

ISBN 979-11-88428-02-1

for Detecting Various Denial of Service Attacks. *Software Networking*, 2018(1), 173-190.

[13] Devi, V. S. (2015, December). Class Specific Feature Selection Using Simulated Annealing. In International Conference on Mining Intelligence and Knowledge Exploration(pp. 12-21). *Springer, Cham.*

[14] H.B.Wong, and G.H. Lim, "Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV," *Proceedings of Singapore healthcare*, 20(4), pp.316-318,2011.

Satyabrata Aich is working as a researcher in the field of computer engineering He has over four years of teaching, research and industry experience in India and abroad. He has published many research papers in journals and conferences in the realms of Supply Chain Management and data analytics. His research interests are natural language processing, Machine learning, supply chain management, data mining.

Kueh Lee Hui is working as an assistant professor at the department of Electrical Engineering, Dong-A University since 2012. She completed her PhD Degrees from Department of Electrical Engineering, Dong-A University, Korea. In 2009 she completed her BS degree in Electronic and Communication, Department of Electronic Engineering, University Malaysia of Sarawak, Malaysia. She also done MS in 2007 from Malaysia. Her research interests are image processing, face recognition, digital image forensic, intelligent control and control application, power system.

Ahmed Abdulhakim Al-Absi is an assistant professor in Department of Computer Engineering (Smart Computing) at Kyungdong University in South Korea. He earned a Ph.D. in computer science from Dongseo University in 2015. He received M.Sc. degree in information technology at University Utara Malaysia in 2011, and B.Sc. degree in computer applications at Bangalore University in 2008. His research interests include Big Data processing, Hadoop, Cloud computing, IoT, Distributed systems, Parallel computing, Bioinformatics, Security, and VANETs.

Mangal Sain received the M.Sc. degree in computer application from India in 2003 and the Ph.D. degree in computer science in 2011. Since 2012, he has been an Assistant Professor with the Department of Computer Engineering, Dongseo University, South Korea. His research interest includes wireless sensor network, cloud computing, Internet of Things, embedded systems, and middleware. He has authored over 50 international publications including journals and international conferences. He is a member of TIIS and a TPC member of more than ten international conferences.

ISBN 979-11-88428-02-1