The pertinent single-attribute-based classifier for small datasets classification

Mona Jamjoom
Department of Computer Sciences, Princess Nourah Bint Abdulrahman University, Kingdom of Saudi Arabia

Article Info	ABSTRACT
	Classifying a dataset using machine learning algorithms can be a big challenge when the target is a small dataset. The OneR classifier can be used for such cases due to its simplicity and efficiency. In this paper, we revealed the power of a single attribute by introducing the pertinent single-attribute-based-heterogeneity-ratio classifier (SAB-HR) that used a pertinent attribute to classify small datasets. The SAB-HR's used feature selection method, which used the Heterogeneity-Ratio (H-Ratio) measure to identify the most homogeneous attribute among the other attributes in the set. Our empirical results on 12 benchmark datasets from a UCI machine learning repository showed that the SAB-HR classifier significantly outperformed the classical OneR classifier for small datasets. In addition, using the H-Ratio as a feature selection criterion for selecting the single attribute was more effectual than other traditional criteria, such as Information Gain (IG) and Gain Ratio (GR). *Copyright © 2020 Institute of Advanced Engineering and Science.* *All rights reserved.*

Corresponding Author:

Mona Jamjoom,
Department of Computer Sciences,
Princess Nourah Bint Abdulrahman University,
Airport Road, Riyadh 11671, Kingdom of Saudi Arabia.
Email: mmjamjoom@pnu.edu.sa

1. INTRODUCTION

Classification is one of the main tasks of data mining and machine learning [1] that is widely used to predict different real-life situations. High accuracy is a key indicator for a successful prediction model. Building an accurate classifier is one of the important goals, and rich datasets make this task easier and more effective [2]. Classifying small datasets efficiently is essential as some real situations cannot provide a sufficient number of cases. A limited training set is challenging to learn and, as a result, base a decision on it. In many multivariable classification or regression problems, such as estimation or forecasting, we have a training set Tp = (x_i, t_i) of p pairs of input/output vector $x \in \Re_n$ and scalar target t. Thus, according to Vapnik's definition, a small dataset for Tp is determined as follows: "For estimating functions with VC dimension h, we consider the size p of data to be small if the ratio p/h is small (say $p/h < 20$)" [3].

The problem with the small dataset is that, if not elaborately collected, it is not a representative sample. Non-representative instances hinder the process of providing enough information for the learner model because of the gaps existing between instances; thus, the model does not generalize well. Many works have been proposed in the literature to solve the problem of small data size by using different methods. One of the common methods used is to increase the size of data by adding artificial instances [4], but this approach lacks data credibility and reflection on real-life use. Some researchers have used feature-selection methods [5-8], whereas a novel technique using multiple runs for model development was proposed by [9] and others.

A simple solution is one of the requirements when the problem is becoming increasingly complex. This philosophy has been stated by Occam's razor [1]. Literature in the field of classification has shown some successful attempts of very simple rules to achieve high accuracy with many datasets [10]. OneR is one of

the simple and widely used algorithms in machine learning to build a simple classifier. A trade-off between simplicity and high performance [10] makes OneR's performance slightly less accurate than state-of-the-art classification algorithms [11, 12], although sometimes it outperforms them [13, 14]. Its main advantage is that it balances the best accuracy possible with a model that is still simple enough for humans to understand [12].

OneR is a single-attribute-based classifier that involves only one attribute at the classification time. A single attribute concept is powerful if it can directly influence the classification accuracy of the dataset in a positive manner. Yet not all attributes have to positively contribute to the classification process which may increase the single attribute power. The single attribute rule can be more effective than complex methods when it is difficult to learn from the dataset due to it being simple, small, noisy, or complex. A study by [15] used the single attribute concept by creating multiple one-dimensional classifiers from the original dataset in the training phase and combining the results in the prediction phase. The new method is unlike OneR because it considers all attributes' contributions at the prediction time. Feature selection is a data-mining pre-processing step widely used to improve the classification and reduce the performance time. It is effective in reducing the dataset's dimensionality by eliminating non-contributable attributes. It uses different techniques to come up with a single attribute or a subset of attributes [16, 17]. Moreover, it has proven its effectiveness in improving various applications' predictive accuracy [18-20].

In this paper, we tackle the problem of classifying small datasets by expanding the power of a pertinent single attribute using SAB-HR classifier, which is similar to OneR classifier in using single attribute at classification phase, but different in which instead of generating a rule for each attribute, a feature selection method is employed to select the attribute that is less heterogenic among the other attributes. We calculated the H-Ratio [21] for each attribute (att) then identified the attribute with the lowest H-Ratio value ($att_{H-Ratio}$). We used the pair ($att_{H-Ratio}$, c), where c is the class value, to learn and classify the small dataset. The results were encouraging and showed a significant improvement compared to the classical OneR classifier. In addition, we created multiple classifiers in the same manner of SAB-HR, using different criteria to select the pertinent single attribute. We used IG and GR in the feature-selection process and created SAB-IG and SAB-GR classifiers, correspondingly. We individually compared the new classifier SAB-HR with others (i.e., SAB-IG and SAB-GR). The remainder of this paper is organized as follows: Section 2 reviews the background of our work. In Section 3, we propose the research method SAB-HR classifier. The experiments and a brief discussion of the findings is in subsections 3.1 and 3.2, consequently. Finally, Section 4 concludes the paper.

2.   BACKGROUND
In this section we will review some of the techniques that will be used in this study.

2.1. OneR classifier
OneR, is short for "One Rule", and has been introduced by Rob Holte [22, 10]. It is one of the most primitive techniques, based on a 1-level decision tree that creates one rule for each attribute in the dataset, then selects the rule with minimum classification errors as its "one rule". To create a rule for an attribute, it constructs a frequency table for each attribute against the class [22], Figure 1 shows the pseudocode of OneR algorithm. It has shown that OneR work distinctively well in practice with real-world data and can compete the state-of-the-art classification algorithms in some situations [13, 14, 23]. OneR is using one attribute for classification and many consider it as one of feature selection methods with feature subset containing a single attribute [24]. Comparing the OneR classifier with the baseline classifier ZeroR [14], OneR is a one step beyond. Both OneR and ZeroR are useful for determining a minimum standard classifier for other classification algorithms. OneR's accuracy is always higher or at least equal the baseline classifier when evaluated on the training data. The authors in [25] proposed attempts to enhance the performance of OneR by addressing two issues: the quantization of continuous-valued attributes, and the treatment of missing values.

```
For each attribute (att),
    For each value of that att, make a rule as follows;
        Count how often each value of class appears
        Find the most frequent class
        Make the rule assign that class to this value of the att
    Calculate the total error of the rules of each att
Choose the att with the smallest total error.
```

Figure 1. The pseudocode of OneR algorithm [15]

2.2. Feature selection

Feature selection methods attempt to find the minimal subset of features that do not significantly decrease the classification accuracy. Feature selection methods can be categorized as wrapper methods or filter methods [17]. Surveys done by [17] and [16] showed plenty of such methods. A wrapper method is a model-based approach where the quality of the features selected is measured by the classification accuracy of the classification algorithm being used. Some use a greedy search to select the subset [16]. Meanwhile, in a filter method, called a model-free approach, the selection of features is done independently from the classification algorithm. It selects the subset's features dependent on general measurable characteristics of the feature, such as information Gain, Gain Ratio, Pearson Correlation, Mutual Information (MI) [16], and Heterogeneity Ratio [21]. In this paper, we used feature selection that utilizes filter methods (i.e., attribute evaluation) and focused on some of the mentioned measures (i.e., IG, GR, and H-Ratio). A brief description of each follows.

- **Information gain** [21] measures the amount of information given by an attribute about the class. It is defined by formula (1):

$$IG(att) = H(Y) - H_{att}(Y) \qquad (1)$$

where $H_{att}(Y)$ measures the entropy of the attribute *att* by contributing to class Y while H(Y) calculates the entropy of class Y. In fact, entropy is the quantity of information contained or delivered by a source of information. It is also used in measuring the relevancy and defined by formula (2):

$$H(Y) = \sum_i - P(v_i) \, log_2 \, P(v_i) \qquad (2)$$

- **Gain ratio** [26] is a ratio of information gain to intrinsic information. It determines the relevancy of an attribute. GR is calculated using the formula (3):

$$GR(att) = \frac{IG(att)}{H(att)} \qquad (3)$$

where $H(att) = \sum_j -P(v_j) \, log_2 \, P(v_j)$ and $P(v_j)$ represents the probability to have the value v_j by contributing to overall values for attribute *j*.

- **Heterogeneity ratio** is a new measure defined by [21] that measures the ratio of heterogeneity of a nominal attribute among the dataset attributes. In other words, it quantifies the homogeneity of a set of instances sharing the same value of attributes. The H-Ratio is defined by formula (4):

$$H - Ratio(att) = \frac{H(att) + H_{att}(Y)}{H(Y)} \qquad (4)$$

The ratio $\frac{H(att)}{H(Y)}$ adds value to the homogeneity instances based on attributes and class simultaneously whereas the ratio $\frac{H_{att}(Y)}{H(Y)}$ appreciates the homogeneity instances of the same class and shares the same value of attributes.

3. RESULTS AND ANALYSIS

In this section, we introduce a new single-attribute-based classifier SAB-HR to classify the small datasets. The new algorithm uses a new criterion to select the powerful pertinent single attribute, which will contribute in the classification. SAB-HR is unlike OneR in generating a rule for each attribute. It calculates the H-Ratio for each attribute (*att_{H-Ratio}*) in the dataset to determine the attribute that is less heterogenic among the other attributes. The attribute with the lowest heterogeneity ratio value is used in pairs with the class *c* (*att_{H-Ratio}, c*) in the classification process while the remaining attributes are eliminated. The power of the single attribute selected for SAB-HR lies in its homogeneity with other attributes in which it provides enough information for the classifier to predict correctly. *att_{H-Ratio}* is a representative attribute that is sufficient for small datasets. The algorithmic description of SAB-HR is presented in Figure 2.

```
For each attribute (att),
    Calculate the att_H-Ratio;
Choose the att with the smallest att_H-Ratio value;
Remove all att in the dataset except the pairs (att_H-Rato, c);
```

Figure 2. The pseudocode of SAB-HR algorithm

3.1. Experiments

In the following experiments, we aim to evaluate the performance of the new SAB-HR classifier when dealing with small datasets. In addition, we want to compare the performance of SAB-HR with other single attribute classifiers that use different criteria, such as IG and GR, when selecting the single attribute during the feature-selection process. We used the well-known open source software WEKA [27]. The datasets were obtained from the UCI Repository for Machine Learning [28]. We selected 12 small datasets corresponding to Vapnik's definition [3]. Table 1 lists the main characteristics of the datasets collected and used in terms of number of instances, number of attributes, and Vapnik's ratio for determining the dataset's size. The number beside the dataset name will be its reference in the figures.

The OneR was used as a base classifier; a 10-fold cross-validation and a paired *t-test* with a confidence level of 95% were used to determine if the differences in classification accuracy were statistically significant, and underlined in the tables. We compared the different methods with respect to the average classification accuracy and the number of datasets for which each method achieved better results. Better results are shown in the tables in bold font.

In the tables, we named each technique using the abbreviation SAB for single-attribute-based name, suffixed with an abbreviation for the measure used for selecting the single attribute in the feature-selection process. The new classifiers, with respect to the different measures, are named as follows: SAB-HR, SAB-IG and SAB-GR. In our experiments, we applied the feature-selection process using different measures (H-Ratio, IG, and GR) to select the pertinent single attribute, then we eliminated the remaining (i.e., unselected) attributes and classified with a pair of attributes (pertinent single attribute, class).

Table 1. Characteristics of datasets used in the experiments

#	Dataset	# instances	# attributes	# instances/# attributes
1	Postoperative-patient-data	90	9	10
2	contact-lenses	24	4	6
3	weather-nominal	14	4	3.5
4	colic.ORIG	368	27	13.63
5	cylinder-bands	540	39	13.85
6	Dermatology	366	34	10.76
7	Flags	194	29	6.69
8	lung-cancer	32	56	0.57
9	spect_train	80	22	3.64
10	Sponge	72	45	1.6
11	Zoo	101	17	5.94
12	primary-tumor	339	17	19.94

3.2. Results and discussion

The experiment's results are combined in Table 2, which compares the performance of classical OneR with the new created classifiers. Noticeably, the performance of the classical OneR is insignificant when compared to the new applied classifiers. The overall average accuracy for the new classifiers (i.e., SAB-HR, SAB-IG and SAB-GR) is 64.6%, 49.72% and 61.31%, respectively, corresponding to 48.53% for the classical OneR classifier. Furthermore, the difference in average accuracy between SAB-HR compared to the classical OneR is statistically significant. The average difference between the classical OneR and the applied classifiers (i.e., SAB-HR, SAB-IG and SAB-GR) is 16.07%, 1.19% and 12.78%, respectively, favoring new classifiers.

Table 2. The performance's summary of applied classifiers compared to the classical OneR classifier

Dataset	OneR	SAB-HR	OneR	SAB-IG	OneR	SAB-GR
Postoperative-patient-data	67.78	**71.11**	67.78	**71.11**	67.78	**68.89**
contact-lenses	70.83	70.83	70.83	70.83	70.83	70.83
weather-nominal	42.86	**57.14**	42.86	**50**	42.86	**50**
colic.ORIG	**67.66**	65.76	67.66	67.66	**67.66**	63.86
cylinder-bands	49.63	**67.59**	49.63	49.63	49.63	**65**
dermatology	**49.73**	36.07	49.73	**50.27**	49.73	36.07
flags	4.64	**33.51**	4.64	4.64	4.64	**42.78**
lung-cancer	87.5	**96.88**	87.5	87.5	**87.5**	78.13
spect_train	67.5	**92.5**	67.5	**75**	67.5	**75**
sponge	4.17	**98.61**	4.17	4.17	4.17	**95.83**
zoo	42.57	**60.4**	42.57	42.57	42.57	**60.4**
primary-tumor	**27.43**	24.78	**27.43**	23.3	27.43	**28.9**
Average Accuracy	48.53	<u>**64.6**</u>	48.53	**49.72**	48.53	**61.31**
# of better dataset	3	**8**	1	**4**	3	**8**

Figure 3 (a-c) compare the applied classifiers to the classical OneR classifier in terms of average accuracy, with the less heterogenous attribute classifier (SAB-HR) ranking first, followed by SAB-GR with a slight difference (3.29%) from first, and SAB-IG classifier with a big difference from other classifiers but looking typical to the classical OneR, the two lines approximately identical as shown in Figure 3 (b). The (att_{IG}) attribute used in SAB-IG contains the largest amount of information about the class. In a small dataset case, it may be more important to be concerned about the consistency of the attribute with other attributes due to the limited number of instances in the dataset. This would minimize the gaps existing between the instances in the dataset. The homogeneity of the dataset helps make it more representative and, thus, more accurate to be learned. In addition, the new classifiers achieved better average accuracy in more datasets than OneR as shown in Table 2. Figure 4 (a-c) shows each new classifier in comparison to OneR. The number of better datasets achieved is 8, 4 and 8 for SAB-HR, SAB-IG and SAB-GR, respectively, corresponding to 3, 1, and 3 for OneR classifier.
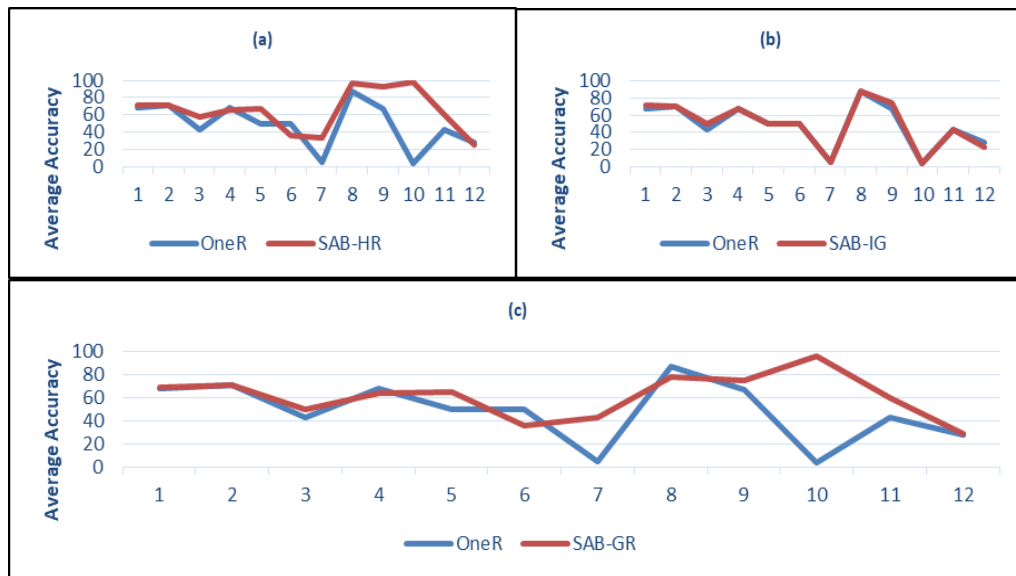
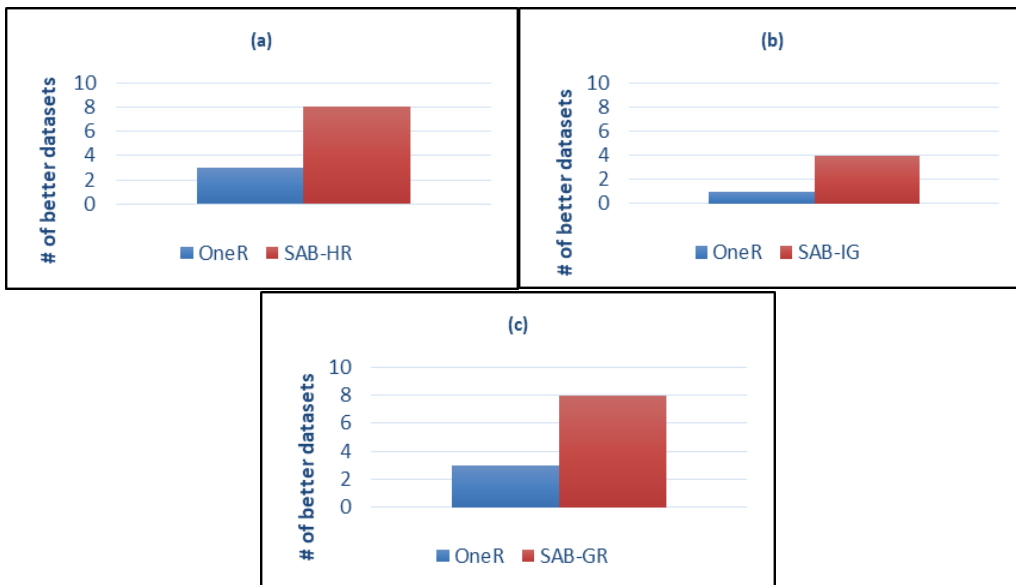Figure 3. Comparison of applied classifiers versus OneR classifier in term of average accuracy

Figure 4. Comparison of applied classifiers versus OneR classifier in term of number
of better datasets achieved

From Table 2, it is obvious that selecting the single attribute that has a lower classification error rate for the OneR classifier is not always optimal, especially in small datasets. Using a more deliberate technique to select the single attribute has a positive impact on classification accuracy and number of better datasets achieved. Meanwhile, we developed Table 3 to highlight the new classifier SAB-HR, which used homogeneity for the pertinent single attribute selection. Table 3 shows a comparison between the new classifier SAB-HR and the other created classifiers for the same purpose (i.e., SAB-IG and SAB-GR). The results showed that SAB-HR's average accuracy outperforms SAB-IG's average accuracy by nearly 14.88%, while with SAB-GR the difference is only 1.37%. In general, the performance of the SAB-HR classifier is remarkable when compared to the classical OneR or the applied classifiers (i.e., SAB-IG and SAB-GR). Figure 5 (a) and (b) show the difference of performance of each dataset between SAB-HR and the other applied classifiers in terms of average accuracy.

Table 3. A comparison between the new classifier SAB-HR and the other classifiers

| Dataset	SAB-HR	SAB-IG	SAB-HR	SAB-GR
Postoperative-patient-data	71.11	71.11	**71.11**	68.89
Contact-lenses	70.83	70.83	70.83	70.83
Weather-nominal	**57.14**	50	**57.14**	50
Colic.ORIG	65.76	**67.66**	65.76	63.86
Cylinder-bands	**67.59**	49.63	**67.59**	65
Dermatology	36.07	**50.27**	36.07	36.07
Flags	**33.51**	4.64	33.51	**42.78**
Lung-cancer	**96.88**	87.5	**96.88**	78.13
Spect_train	**92.5**	75	75	75
Sponge	**98.61**	4.17	93.06	**95.83**
Zoo	**60.4**	42.57	60.4	60.4
Primary-tumor	**24.78**	23.3	24.78	**28.9**
Average Accuracy	<u>**64.6**</u>	49.72	**62.68**	61.31
# of better dataset	**8**	2	**5**	3
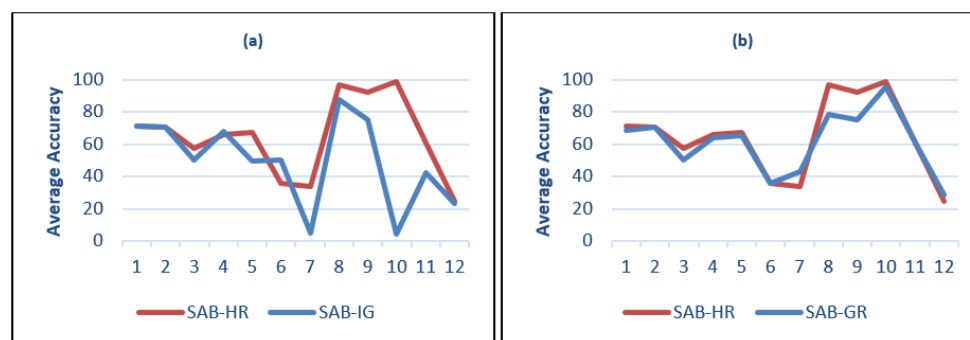
Figure 5. Comparison of applied classifiers versus SAB-HR classifier in term of Average Accuracy

In summary, we can conclude that, for small datasets, using a simple classifier, such as OneR, is one of the main options for enhancing its classification accuracy. In addition, employing the feature-selection method for selecting a single attribute using a common measure like H-Ratio, IG or GR will do so, with better results. On the other hand, considering the homogeneity of the attribute for pertinent single attribute selection can positively impact the classification process. It helped to reduce the gap between instances, and accordingly had a representative dataset. Consequently, it provided enough information for the classifier to learn and achieve a decent average accuracy. From the previous results, single-attribute-based classifier can be powerful for classifying small datasets when the pertinent attribute is selected. That is the case with the new SAB-HR, which is recommended among the tested classifiers in this work.

4.   CONCLUSION

In this work we have explored the power of the single attribute when selected using an effectual feature-selection criterion. We have addressed the small dataset mining problem as it is not always easy to gather a large amount of real data. The new algorithm SAB-HR is a pertinent single-attribute-based classifier

consisting of a pair of (simplicity, effectiveness) to contribute positively in classifying small datasets. The single attribute selected to be the most homogenous with the other attributes in the dataset gives more consistency between instances. Our empirical results used 12 benchmark datasets of a small size corresponding to Vapnik's definition. The results show that SAB-HR's performance significantly outperforms the classical OneR's performance. In addition, we compared the performance of SAB-HR with other single attribute classifiers that use different attribute selection criteria (e.g., IG and GR), and all the results confirmed the effectiveness of the SAB-HR classifier. In future work, we intend to investigate algorithms to improve the classification accuracy of small datasets using more progressive classifiers. In addition, we aim to propose more simple methods for classification.

REFERENCES

[1]   T. Mitchell, "Machine Learning," McGraw Hill, 1997.
[2]   T. Van Gemert, "On the influence of dataset characteristics on classifier performance," Bachelor Thesis, Faculty of Humanities, Utrecht University, pp. 1–13, 2017.
[3]   V. Vapnik, "Statistical Learning Theory," Wiley, New York, 2000.
[4]   N. H. Ruparel, N. M. Shahane, and D. P. Bhamare, "Learning from small data set to build classification model : A survey," *IJCA Proceedings on International Conference on Recent Trends in Engineering and Technology ICRTET*, vol. 4, pp. 23–26, 2013.
[5]   X. Chen and J. C. Jeong, "Minimum reference set based feature selection for small sample classifications," *Proceedings of the 24th International Conference on Machine Learning - ICML '07*, pp. 153–160, 2007.
[6]   S. L. Happy, R. Mohanty, and A. Routray, "An effective feature selection method based on pair-wise feature proximity for high dimensional low sample size data," *25th European Signal Processing Conference, EUSIPCO*, pp. 1574–1578, 2017.
[7]   A. Golugula, G. Lee, and A. Madabhushi, "Evaluating feature selection strategies for high dimensional, small sample size datasets," *Conference Proceedings Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society*, pp. 949–952, 2011.
[8]   I. Soares, J. Dias, H. Rocha, M. do Carmo Lopes, and B. Ferreira, "Feature selection in small databases: A medical-case study," *IFMBE Proceedings: XIV Mediterranean Conference on Medical and Biological Engineering and Computing*, vol. 57, pp. 808–813, 2016.
[9]   T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Machine learning for predictive modelling based on small data in biomedical engineering," *IFAC-PapersOnLine*, vol. 28, pp. 469–474, 2015.
[10] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–91, 1993.
[11] A. K. Dogra and T. Wala, "A comparative study of selected classification algorithms of data mining," *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 6, pp. 220–229, 2015.
[12] F. Alam and S. Pachauri, "Comparative study of J48, Naive Bayes and One-R classification technique for credit card fraud detection using WEKA," *Advances in Computational Sciences and Technology*, vol. 10, no. 6, pp. 1731–1743, 2017.
[13] V. S. Parsania, N. N. Jani, and N. H. Bhalodiya, "Applying Naïve Bayes, BayesNet, PART, JRip and OneR algorithms on hypothyroid database for comparative analysis," *IJDI-ERET*, vol. 3, pp. 1–6, 2015.
[14] C. Nasa and Suman, "Evaluation of different classification techniques for WEB data," *International Journal of Computer Applications*, vol. 52, pp. 34–40, 2012.
[15] L. Du and Q. Song, "A simple classifier based on a single attribute," *Proceedings of the 14th IEEE International Conference on High Performance Computing and Communications, HPCC-2012 & 9th IEEE International Conference on Embedded Software and Systems, ICESS-2012*, pp. 660–665, 2012.
[16] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997.
[17] L. Huan and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 491–502, 2005.
[18] M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," *Journal of Computing*, vol. 1, pp. 7–11, 2009.
[19] Y. Pan, "A proposed frequency-based feature selection method for cancer classification," *Master Theses & Specialist Projects, Top SHCOLAR, Faculty of the Department of Computer Science, Westers Kentucky University*, 2017.
[20] I. Sangaiah, A. V. A. Kumar, and A. Balamurugan, "An empirical study on different ranking methods for effective data classification," *Journal of Modern Applied Statistical Methods*, vol. 14, pp. 35–52, 2015.
[21] M. Trabelsi, N. Meddouri, and M. Maddouri, "A new feature selection method for nominal classifier based on formal concept analysis," *Procedia Computer Science*, vol. 112, pp. 186–194, 2017.

[22] R. Holte, "Machine learning," *Proceeding of the Tenth International Conference, University of Massachusetts, Amherst, June* 1993.

[23] D .I. Morariu, R. G. C. Ulescu, and M. Breazu, "Feature selection in document classification," *The Fourth International Conference in Romania of Information Science and Information Literacy, Romania*, 2013.

[24] J. Novakovic, "Using information gain attribute evaluation to classify sonar targets," *17 ThT Elecommunucation Forum*, pp. 1351–1354, 2009.

[25] C. G. Nevill-Manning, G. Holmes, and I. H. Witten, "The development of Holte's 1R classifier," *Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, 1995.

[26] J. Novaković, P. Strbac, and D. Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugoslav Journal of Operations Research*, vol. 21, pp. 119–135, 2011.

[27] U of Waikato, "WEKA: The Waikato environment for knowledge analysis," 2018. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[28] UCI, "UCI machine learning repository," 2018. [Online], Available: http://archive.ics.uci.edu/ml/machine-learningdatabases/