Empirical Investigation of Multi-tier Ensembles for the Detection of Cardiac Autonomic Neuropathy Using Subsets of the Ewing Features

J. Abawajy[1], A.V. Kelarev[1], A. Stranieri[2], H.F. Jelinek[3]

[1]School of Information Technology, Deakin University
221 Burwood Highway, Burwood, Victoria 3125, Australia
{*jemal.abawajy,kelarev*}*@deakin.edu.au*

[2]School of Science, Information Technology and Engineering
University of Ballarat, P.O. Box 663, Ballarat, Victoria 3353, Australia
a.stranieri@ballarat.edu.au

[3]Centre for Research in Complex Systems and School of Community Health
Charles Sturt University, P.O. Box 789, Albury, NSW 2640, Australia
hjelinek@csu.edu.au

Abstract. This article is devoted to an empirical investigation of performance of several new large multi-tier ensembles for the detection of cardiac autonomic neuropathy (CAN) in diabetes patients using subsets of the Ewing features. We used new data collected by the diabetes screening research initiative (DiScRi) project, which is more than ten times larger than the data set originally used by Ewing in the investigation of CAN. The results show that new multi-tier ensembles achieved better performance compared with the outcomes published in the literature previously. The best accuracy 97.74% of the detection of CAN has been achieved by the novel multi-tier combination of AdaBoost and Bagging, where AdaBoost is used at the top tier and Bagging is used at the middle tier, for the set consisting of the following four Ewing features: the deep breathing heart rate change, the Valsalva manoeuvre heart rate change, the hand grip blood pressure change and the lying to standing blood pressure change.

1 Introduction

Cardiac autonomic neuropathy (CAN) is a condition associated with damage to the autonomic nervous system innervating the heart and highly prevalent in people with diabetes, [6, 7, 24]. The detection of CAN is important for timely treatment, which can lead to an improved well-being of the patients and a reduction in morbidity and mortality associated with cardiac disease in diabetes.

This article is devoted to empirical investigation of the performance of novel large binary multi-tier ensembles in a new application for the detection of cardiac autonomic neuropathy (CAN) in diabetes patients using subsets of the Ewing

features. This new construction belongs to the well known general and productive multi-tier approach, considered by the first author in [14, 15].

Standard ensemble classifiers can generate large collections of base classifiers, train them and combine into a common classification system. Here we deal with new large multi-tier ensembles, combining diverse ensemble techniques on two tiers into one scheme, as illustrated in Figure 1. Arrows in the diagram correspond to the generation and training stage of the system, and show that tier 2 ensemble generates and trains tier 1 ensembles and executes them in the same way as it is designed to handle simple base classifiers. In turn, each tier 1 ensemble applies its method to the base classifier in the bottom tier.

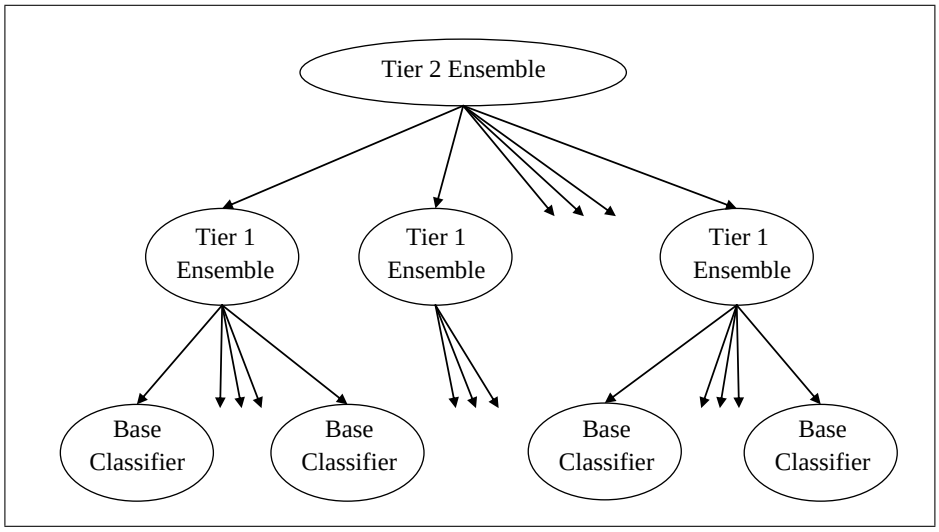

Fig. 1. The generation and training stage of multi-tier ensembles

Large multi-tier ensembles illustrated in Figure 1 have not been considered in the literature before in this form. They can be also regarded as a contribution to the very large and general direction of research devoted to the investigation of various multi-stage and multi-step approaches considered previously by other authors. Let us refer to [1, 14, 15] for examples, discussion and further references.

Our experiments used the Diabetes Screening Complications Research Initiative (DiScRi) data set collected at Charles Sturt University, Albury, Australia. DiScRi is a very large and unique data set containing a comprehensive collection of tests related to CAN. It has previously been considered in [5, 13, 21–23],

For the large DiScRi data set our new multi-tier ensembles produced better outcomes compared with those published in the literature previously. Our new results using multi-tier ensembles achieved substantially higher accuracies.

The paper is organised as follows. Section 2 describes the Diabetes Complications Screening Research Initiative, cardiac autonomic neuropathy and the

Ewing features. Section 3 deals with the base classifiers and standard ensemble classifiers. Section 4 describes our experiments and presents the experimental results comparing the effectiveness of base classifiers, ensemble classifiers and multi-tier ensembles for several subsets of the Ewing features. These outcomes are discussed in Section 5. Main conclusions are presented in Section 6.

2 Diabetes Complications Screening Research Initiative and the Ewing Features

This paper analysed the data set of test results and health-related parameters collected at the Diabetes Complications Screening Research Initiative, DiScRi, organised at Charles Sturt University, [5]. The collection and analysis of data has been approved by the Ethics in Human Research Committee of the university before investigations started. People participating in the project were attracted via advertisements in the media. The participants were instructed not to smoke and refrain from consuming caffeine containing drinks and alcohol for 24 hours preceding the tests as well as to fast from midnight of the previous day until tests were complete. The measurements were recorded in the DiScRi data base along with various other health background data including age, sex and diabetes status, blood pressure (BP), body-mass index (BMI), blood glucose level (BGL), and cholesterol profile. Reported incidents of a heart attack, atrial fibrillation and palpitations were also recorded.

The most essential tests required for the detection of CAN rely on assessing responses in heart rate and blood pressure to various activities, usually consisting of five tests described in [6] and [7]. Blood pressure and heart rate are very important features [2], [29]. The most important set of features recorded for detection of CAN is the *Ewing battery* [6], [7]. There are five Ewing tests in the battery: changes in heart rate associated with lying to standing, deep breathing and valsalva manoeuvre and changes in blood pressure associated with hand grip and lying to standing. In addition features from ten second samples of 12-lead ECG recordings for all participants were extracted from the data base. These included the QRS, PQ, QTc and QTd intervals, heart rate and QRS axis. (QRS width has also been shown to be indicative of CAN [9] and is included here.)

It is often difficult for clinicians to collect all test data. Patients are likely to suffer from other illnesses such as respiratory or cardiovascular dysfunction, obesity or arthritis, making it hard to follow correct procedures for all tests. This is one of the reasons why it particularly important to investigate various subsets of the Ewing battery.

The QRS complex and duration reflects the depolarization of the ventricles of the heart. The time from the beginning of the P wave until the start of the next QRS complex is the PQ interval. The period from the beginning of the QRS complex to the end of the T wave is denoted by QT interval, which if corrected for heart rate becomes the QTc. It represents the so-called refractory period of the heart. The difference of the maximum QT interval and the minimum QT interval over all 12 leads represents the QT dispersion (QTd). It is used as an

indicator of the repolarisation of the ventricles. The deflection of the electrical axis of the heart measured in degrees to the right or left is called the QRS axis.

The whole DiScRi database contains over 200 features. We used the following notation for the Ewing features and the QRS width:

$LSHR$	stands for the lying to standing heart rate change;
$DBHR$	is the deep breathing heart rate change;
$VAHR$	is the Valsalva manoeuvre heart rate change;
$HGBP$	is the hand grip blood pressure change;
$LSBP$	is the lying to standing blood pressure change;
QRS	is the width of the QRS segment, which is also known as a highly significant indicator of CAN [9].

The detection of CAN deals with a binary classification where all patients are divided into one of two classes: a 'normal' class consisting of patients without CAN, and a 'definite' class of patients with CAN. Detection of CAN allows clinicians to collect fewer tests and can be performed with higher accuracy compared with multi-class classifications of CAN progression following more detailed definitions of CAN progression classes originally introduced by Ewing. More details on various tests for CAN are given in the next section. This paper is devoted to the detection of CAN using subsets of the Ewing features.

A preprocessing system was implemented in Python to automate several expert editing rules that can be used to reduce the number of missing values in the database. These rules were collected during discussions with the experts maintaining the database. Most of them fill in missing entries of slowly changing conditions, like diabetes, on the basis of previous values of these attributes. Preprocessing of data using these rules produced 1299 complete rows with complete values of all fields, which were used for the experimental evaluation of the performance of data mining algorithms.

3 Binary Base Classifiers and Standard Ensemble Methods

Initially, we ran preliminary tests for many binary base classifiers available in Weka [12] and included the following classifiers for a series of complete tests with outcomes presented in Section 4. These robust classifiers were chosen since they represent most essential types of classifiers available in Weka [12] and performed well for our data set in our initial preliminary testing:

- *ADTree* classifier trains an Alternating Decision Tree, as described in [10]. Weka implementation of ADTree could process only binary classes.
- *J48* generates a pruned or unpruned C4.5 decision tree [31].
- *LibSVM* is a library for Support Vector Machines [8]. It can handle only attributes without missing values and only binary classes.
- *NBTree* uses a decision tree with naive Bayes classifiers at the leaves, [28].
- *RandomForest* constructs a forest of random trees following [4].

- *SMO* uses Sequential Minimal Optimization for training a support vector classifier, [19, 30]. Initially, we tested all kernels of SMO available in Weka and used it with polynomial kernel that performed best for our data set.

We used SimpleCLI command line in Weka [12] to investigate the performance of the following ensemble techniques:

- *AdaBoost* training every successive classifier on the instances that turned out more difficult for the preceding classifier [11];
- *Bagging* generating bootstrap samples to train classifiers and amalgamating them via a majority vote, [3];
- *Dagging* dividing the training set into a disjoint stratified samples [33];
- *Grading* labelling base classifiers as correct or wrong [32];
- *MultiBoosting* extending AdaBoost with the wagging [34];
- *Stacking* can be regarded as a generalization of voting, where meta-learner aggregates the outputs of several base classifiers, [35].

We used SimpleCLI command line in Weka [12] to train and test multi-tier ensembles of binary classifiers too.

4 Experimental Results

We used 10-trial 10-fold cross validation to evaluate the effectiveness of classifiers in all experiments. It is often difficult to obtain results for all five tests and we therefore included the largest subsets of four features from the Ewing battery. These subsets can help clinicians to determine whether CAN is present in those situations when one of the tests is missing. The following notation is used to indicate these subsets in the tables with outcomes of our experiments:

S_{Ewing}	is the set of all five Ewing features, i.e., $LSHR$, $DBHR$, $VAHR$, $HGBP$ and $LSBP$;
S_{LSHR}	is the set of four Ewing features with $LSHR$ excluded, i.e., $DBHR$, $VAHR$, $HGBP$ and $LSBP$;
S_{DBHR}	is the set of four Ewing features with $DBHR$ excluded, i.e., $LSHR$, $VAHR$, $HGBP$ and $LSBP$;
S_{VAHR}	is the set of four Ewing features with $VAHR$ excluded, i.e., $LSHR$, $DBHR$, $HGBP$ and $LSBP$;
S_{HGBP}	is the set of four Ewing features with $HGBP$ excluded, i.e., $LSHR$, $DBHR$, $VAHR$ and $LSBP$;
S_{LSBP}	is the set of four Ewing features with $LSBP$ excluded, i.e., $LSHR$, $DBHR$, $VAHR$ and $HGBP$;
S_4	is the set of two heart rate features $LSHR$, $DBHR$, one blood pressure feature $HGBP$, with QRS added.

Feature selection methods are very important, see [25], [26], [27]. In particular, the set S_4 was identified by the authors in [13] using feature selection.

First, we compared the effectiveness of base classifiers for these sets of features. We used accuracy to compare the classifiers, since it is a standard measure of performance. The accuracy of a classifier is the percentage of all patients classified correctly. It can be expressed as the probability that a prediction of the classifier for an individual patient is correct. The experimental results comparing all base classifiers are included in Table 1. These outcomes show that for the DiScRi database RandomForest is the most effective classifier. It is interesting that many classifiers worked more accurately when the LSHR feature had been excluded.

	Subsets of features						
	S_{Ewing}	S_{LSHR}	S_{DBHR}	S_{VAHR}	S_{HGBP}	S_{LSBP}	S_4
ADTree	84.14	84.68	75.31	80.08	81.02	71.73	80.77
J48	91.61	92.15	85.14	90.92	91.28	89.99	91.38
LibSVM	92.39	92.94	80.97	92.71	85.82	84.78	91.13
NBTree	90.15	91.07	81.83	87.45	87.22	86.99	87.76
RandomForest	94.46	94.84	91.76	93.61	94.23	93.76	94.35
SMO	74.13	73.75	64.36	71.98	73.83	71.36	74.44

Table 1. Accuracy of base classifiers for the detection of CAN using subsets of Ewing features

Second, we compared several ensemble classifiers in their ability to improve the results. Preliminary tests demonstrated that ensemble classifiers based on RandomForest were also more effective than the ensembles based on other classifiers. We compared AdaBoost, Bagging, Dagging, Grading, MultiBoost and Stacking based on RandomForest. The accuracies of the resulting ensemble classifiers are presented in Table 2, which shows improvement. We used one and the same base classifier, RandomForest, in all tests included in this table. We tested several other ensembles with different base classifiers, and they turned out worse.

Finally, we compared the results obtained by all multi-tier ensembles combining AdaBoost, Bagging and MultiBoost, since these ensembles produced better accuracies in Table 2. Tier 2 ensemble treats the tier 1 ensemble and executes it in exactly the same way as it handles a base classifier. In turn the tier 1 ensemble applies its method to the base classifier as usual. We do not include repetitions of the same ensemble technique in both tiers, since such repetitions were less effective. The outcomes of the multi-tier ensembles of binary classifiers are collected in Tables 3.

	Subsets of features						
	S_{Ewing}	S_{LSHR}	S_{DBHR}	S_{VAHR}	S_{HGBP}	S_{LSBP}	S_4
AdaBoost	96.84	97.23	94.07	95.99	96.59	96.11	96.51
Bagging	96.37	96.75	93.63	95.52	96.13	95.67	96.05
Dagging	89.75	90.13	87.18	88.94	89.54	89.10	89.46
Grading	94.49	94.87	91.79	93.61	94.26	93.78	94.18
MultiBoost	96.37	96.77	93.62	95.50	96.13	95.65	96.04
Stacking	95.44	95.81	92.70	94.56	95.20	94.73	95.09

Table 2. Accuracy of ensemble classifiers for the detection of CAN using subsets of Ewing features

		Subsets of features						
Tier 2	Tier 1	S_{Ewing}	S_{LSHR}	S_{DBHR}	S_{VAHR}	S_{HGBP}	S_{LSBP}	S_4
AdaBoost	Bagging	97.35	97.74	94.58	96.50	97.12	96.65	97.04
AdaBoost	MultiBoost	96.37	96.78	93.65	95.52	96.14	95.68	96.07
Bagging	AdaBoost	97.33	97.73	94.57	96.49	97.09	96.61	97.00
Bagging	MultiBoost	96.66	97.08	93.91	95.80	96.42	95.96	96.34
MultiBoost	AdaBoost	96.85	97.25	94.08	96.00	96.62	96.13	96.52
MultiBoost	Bagging	97.05	97.43	94.30	96.19	96.83	96.35	96.74

Table 3. Accuracy of multi-tier ensembles of binary classifiers for the detection of CAN using subsets of the Ewing features

5 Discussion

DiScRi is a very large and unique data set containing a comprehensive collection of tests related to CAN. It has been previously considered in [13, 21–23], New results obtained in this paper achieved substantially higher accuracies than the previous outcomes published in [13]. Overall, the results of the present paper are also appropriate for other data mining applications in general when compared to recent outcomes obtained for other data sets using different methods, for example, in [16] and [17].

AdaBoost has produced better outcomes than other ensemble methods for subsets of the Ewing features of the DiScRi data set; and the best outcomes were obtained by a novel combined ensemble classifier where AdaBoost is used after Bagging.

There are several reasons, why other techniques turned out less effective. First, Dagging uses disjoint stratified training sets to create an ensemble, which benefits mainly classifiers of high complexity. Our outcomes demonstrate that the base classifiers considered in this paper are fast enough and this benefit was not essential. Second, stacking and grading use an ensemble classifier to combine the outcomes of base classifiers. These methods are best applied to combine diverse collections of base classifiers. In this setting stacking performed worse than bagging and boosting.

Our experiments show that such large multi-tier ensembles of binary classifiers are in fact fairly easy to use and can also be applied to improve classifications, if diverse ensembles are combined at different tiers. It is an interesting question for future research to investigate multi-tier ensembles for other large datasets.

6 Conclusion

We have investigated the performance of novel multi-tier ensembles for the detection of cardiac autonomic neuropathy (CAN) using subsets of the Ewing features. Our experimental results show that large multi-tier ensembles can be used to increase the accuracy of classifications. They have produced better outcomes compared with previous results published in the literature. The best accuracy 97.74% of the detection of CAN has been achieved by the novel multi-tier combination of AdaBoost and Bagging, where AdaBoost is used at the top tier and Bagging is used at the middle tier, for the set consisting of the following four Ewing features: the deep breathing heart rate change, the Valsalva manoeuvre heart rate change, the hand grip blood pressure change and the lying to standing blood pressure change. This level of accuracy is also quite good in comparison with the outcomes obtained recently for other data sets in closely related areas using different methods, for example, in [18, 20, 16, 17, 36].

Acknowledgements

This work was supported by a Deakin-Ballarat collaboration grant. The authors are grateful to four referees for comments that have helped to improve the presentation, and for suggesting several possible directions for future research.

References

1. Al-Ani, A., Deriche, M.: A new technique for combining classifiers using the Dempster-Shafer theory of evidence. Journal of Artificial Intelligence Research 17, 333–361 (2012)
2. Al-Jaafreh, M., Al-Jumaily, A.A.: New model to estimate mean blood pressure by heart rate with stroke volume changing influence. In: IEEE EMBC 2006: 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 1803–1805 (2006)
3. Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996)
4. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
5. Cornforth, D., Jelinek, H.F.: Automated classification reveals morphological factors associated with dementia. Applied Soft Computing 8, 182–190 (2007)
6. Ewing, D.J., Campbell, J.W., Clarke, B.F.: The natural history of diabetic autonomic neuropathy. Q. J. Med. 49, 95–100 (1980)
7. Ewing, D.J., Martyn, C.N., Young, R.J., Clarke, B.F.: The value of cardiovascular autonomic function tests: 10 years experience in diabetes. Diabetes Care 8, 491–498 (1985)
8. Fan, R.E., Chen, P.H., Lin, C.J.: Working set selection using second order information for training SVM. J. Machine Learning Research 6, 1889–1918 (2005)
9. Fang, Z.Y., Prins, J.B., Marwick, T.H.: Diabetic cardiomyopathy: evidence, mechanisms, and therapeutic implications. Endocr. Rev. 25, 543–567 (2004)
10. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: Proc. 16th Internat. Conf. Machine Learning. pp. 124–133 (1999)
11. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proc. 13th Internat. Conf. Machine Learning. pp. 148–156 (1996)
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations 11, 10–18 (2009)
13. Huda, S., Jelinek, H.F., Ray, B., Stranieri, A., Yearwood, J.: Exploring novel features and decision rules to identify cardiovascular autonomic neuropathy using a hybrid of wrapper-filter based feature selection. In: Sixth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP 2010. pp. 297–302. IEEE Press, Sydney (2010)
14. Islam, R., Abawajy, J.: A multi-tier phishing detection and filtering approach. Journal of Network and Computer Applications p. to apper soon (2012)
15. Islam, R., Abawajy, J., Warren, M.: Multi-tier phishing email classification with an impact of classifier rescheduling. In: 10th International Symposium on Pervasive Systems, Algorithms, and Networks, ISPAN 2009. pp. 789–793 (2009)
16. Jelinek, H.F., Khandoker, A., Palaniswami, M., McDonald, S.: Heart rate variability and QT dispersion in a cohort of diabetes patients. Computing in Cardiology 37, 613–616 (2010)

17. Jelinek, H.F., Rocha, A., Carvalho, T., Goldenstein, S., Wainer, J.: Machine learning and pattern classification in identification of indigenous retinal pathology. In: 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5951–5954. IEEE Press (2011)
18. Kang, B., Kelarev, A., Sale, A., Williams, R.: A new model for classifying DNA code inspired by neural networks and FSA. In: Advances in Knowledge Acquisition and Management. Lecture Notes in Computer Science, vol. 4303, pp. 187–198 (2006)
19. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation 13(3), 637–649 (2001)
20. Kelarev, A., Kang, B., Steane, D.: Clustering algorithms for ITS sequence data with alignment metrics. In: AI 2006: Advances in Artificial Intelligence, 19th Australian Joint Conference on Artificial Intelligence. Lecture Notes in Artificial Intelligence, vol. 4304, pp. 1027–1031 (2006)
21. Kelarev, A.V., Dazeley, R., Stranieri, A., Yearwood, J.L., Jelinek, H.F.: Detection of CAN by ensemble classifiers based on Ripple Down Rules. In: Pacific Rim Knowledge Acquisition Workshop, PKAW2012. Lecture Notes in Artificial Intelligence, vol. 7457, pp. 147–159 (2012)
22. Kelarev, A.V., Stranieri, A., Yearwood, J.L., Jelinek, H.F.: Empirical investigation of consensus clustering for large ECG data sets. In: 25th International Symposium on Computer Based Medical Systems, CBMS2012. pp. 1–4 (2012)
23. Kelarev, A.V., Stranieri, A., Yearwood, J.L., Jelinek, H.F.: Empirical study of decision trees and ensemble classifiers for monitoring of diabetes patients in pervasive healthcare. In: Network-Based Information Systems, NBiS-2012. pp. 441–446 (2012)
24. Khandoker, A.H., Jelinek, H.F., Palaniswami, M.: Identifying diabetic patients with cardiac autonomic neuropathy by heart rate complexity analysis. BioMedical Engineering OnLine 8, http://www.biomedical–engineering–online.com/content/8/1/3 (2009)
25. Khushaba, R., Al-Jumaily, A., Al-Ani, A.: Evolutionary fuzzy discriminant analysis feature projection technique in myoelectric control. Pattern Recognition Letters 30, 699–707 (2009)
26. Khushaba, R., AlSukker, A., Al-Ani, A., Al-Jumaily, A.: A novel swarm-based feature selection algorithm in multifunction myoelectric control. Journal of Intelligent & Fuzzy Systems 20, 175–185 (2009)
27. Khushaba, R.N., Al-Ani, A., Al-Sukker, A., Al-Jumaily, A.: A combined ant colony and differential evolution feature selection algorithm. In: ANTS2008: Ant Colony Optimization and Swarm Intelligence. Lecture Notes in Computer Science, vol. 5217, pp. 1–12 (2008)
28. Kohavi, R.: Scaling up the accuracy of Naive-Bayes classifiers: a Decision-Tree hybrid. In: Second International Conference on Knowledge Discovery and Data Mining. pp. 202–207 (1996)
29. Mahmood, U., Al-Jumaily, A., Al-Jaafreh, M.: Type-2 fuzzy classification of blood pressure parameters. In: ISSNIP2007: The third International Conference on Intelligent Sensors, Sensor Networks and Information Processing. pp. 595–600 (2007)
30. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods – Support Vector Learning (1998)
31. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA (1993)

32. Seewald, A.K., Fuernkranz, J.: An evaluation of grading classifiers. In: Advances in Intelligent Data Analysis. Lecture Notes in Computer Science, vol. 2189/2001, pp. 115–124 (2001)
33. Ting, K.M., Witten, I.H.: Stacking bagged and dagged models. In: Fourteenth International Conference on Machine Learning. pp. 367–375 (1997)
34. Webb, G.I.: MultiBoosting: a technique for combining boosting and wagging. Machine Learning 40, 159 – 196 (2000)
35. Wolpert, D.H.: Stacked generalization. Neural Networks 5, 241–259 (1992)
36. Yearwood, J.L., Kang, B.H., Kelarev, A.V.: Experimental investigation of classification algorithms for ITS dataset. In: Pacific Rim Knowledge Acquisition Workshop, PKAW 2008. pp. 262–272. Hanoi, Vietnam, 15–16 December 2008 (2008)