Improving Classifications for Cardiac Autonomic Neuropathy Using Multi-level Ensemble Classifiers and Feature Selection Based on Random Forest

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

[1]School of Science, Information Technology and Engineering
University of Ballarat, P.O. Box 663, Ballarat, Victoria 3353, Australia
Email: {a.kelarev,a.stranieri,j.yearwood}@ballarat.edu.au

[2]School of Information Technology,
Deakin University, Victoria 3125, Australia
Email: {kelarev,jemal.abawajy}@deakin.edu.au

[3]School of Community Health, Charles Sturt University
P.O. Box 789, Albury, NSW 2640, Australia
Email: hjelinek@csu.edu.au

Abstract

This paper is devoted to empirical investigation of novel multi-level ensemble meta classifiers for the detection and monitoring of progression of cardiac autonomic neuropathy, CAN, in diabetes patients. Our experiments relied on an extensive database and concentrated on ensembles of ensembles, or multi-level meta classifiers, for the classification of cardiac autonomic neuropathy progression. First, we carried out a thorough investigation comparing the performance of various base classifiers for several known sets of the most essential features in this database and determined that Random Forest significantly and consistently outperforms all other base classifiers in this new application. Second, we used feature selection and ranking implemented in Random Forest. It was able to identify a new set of features, which has turned out better than all other sets considered for this large and well-known database previously. Random Forest remained the very best classifier for the new set of features too. Third, we investigated meta classifiers and new multi-level meta classifiers based on Random Forest, which have improved its performance. The results obtained show that novel multi-level meta classifiers achieved further improvement and obtained new outcomes that are significantly better compared with the outcomes published in the literature previously for cardiac autonomic neuropathy.

Keywords: Random Forest, ensembles of ensembles, multi-level ensembles, meta classifiers, feature selection, cardiac autonomic neuropathy.

1 Introduction

The investigation of medical applications of data mining is very important and has been considered, for example, in recent articles by Al-Oqaily et al. (2008), Han et al. (2006), Kennedy et al. (2008), Li et al. (2009), Liang & Zhang (2011), Sinha et al. (2011), Shouman et al. (2011), Sun et al. (2011), Tayebi et al. (2011), Van et al. (2011), where more background information and further references can be found. In particular, valuable information concerning cardiac patients has been obtained using data mining methods, for example, by Cornforth & Jelinek (2007), Han et al. (2006), Jelinek et al. (2010) and Van et al. (2011)

This article is devoted to experimental investigation of several data mining methods for a new application to the study of cardiac autonomic neuropathy (CAN), which is a cardiac condition quite common in diabetes patients. We used an extensive database created by the Diabetes Complications Screening Research Initiative (DiScRi) at Charles Sturt University and concentrated on the particular task of monitoring the progression of cardiac autonomic neuropathy.

First, we compared the performance of many base classifiers for various sets of the most essential features in this database and determined that Random Forest significantly and consistently outperforms all base classifiers in this new application. Second, we used Random Forest feature selection and found a new set of features, which has turned out much better than all sets of features considered previously for CAN in the literature. We verified that Random Forest remained the very best classifier for the new set of features too. Third, we carried out a systematic investigation of various ensemble meta classifiers and found that ensembles based on Random Forest also outperform ensemble meta classifiers based on other classifiers, and that ensemble techniques can be used for further improvement of the performance of Random Forest for this dataset.

Many effective applications of ensemble techniques in data mining have been developed recently. Let us refer, for example, to Ting et al. (2009), Ting et al. (2011), Webb (2008), Webb & Zheng (2004), Yang et al. (2005) In particular, it is well known that various constructions of meta classifiers creating ensembles of base classifiers are capable of improving the stability and effectiveness of classifications.

This article concentrates, in particular, on a systematic empirical investigation of the performance of novel large multi-level meta classifiers for monitoring of CAN progression in diabetes patients. To the best of our knowledge such ensembles of ensembles or multi-level ensemble meta classifiers have not been considered in the literature before, probably because personal computers routinely used in research have only recently become powerful enough to train them for data sets large enough to justify the use of such large classification systems. On the other hand, the motivation and inspiration for our study originally came from many different multi-stage procedures that had been treated previously, for example, by Christen (2007), Islam & Abawajy (2012), Jiangning et al. (2012) and Madjarov et al. (2011).

Diabetes is a condition requiring continuous everyday monitoring of medical tests to adjust the diet, administer medication, update or modify treatment plans and provide further assistance (Wickramasinghe et al. 2011). These tasks make the development of data mining algorithms for the analysis of test results for diabetes patients particularly valuable. To monitor the progression of a specific clinical condition one has to find a small set of features to be collected and efficient algorithms for the processing of these features.

Experimental research comparing various algorithms applied to particular areas is important, since previous experience of such investigations can be used to guide further implementations and achieve better performance in future practical applications. Indeed, there does not exist a single algorithm that is best for all application domains. The effectiveness of any given category of algorithms depends on the size of a data set, number and types of attributes, and the nature of functional relations and dependencies among the attributes. This is also confirmed by the so-called "no-free-lunch" theorems, which imply that there does not exist one algorithm, which is best for all problems (Wolpert 1996). The present paper concentrates on testing multi-level meta classifiers for the classification of cardiac autonomic neuropathy progression, see Section 3 for details. Our experiments included multi-level meta classifiers combining diverse meta classifiers on two levels. These new results show, in particular, that Random Forest performed best in this setting, and that novel multi-level meta classifiers can be used to achieve further improvement of the classification outcomes for cardiac autonomic neuropathy progression. The multi-level meta classifiers based on Random Forest achieved better performance compared with the results published in the literature (Huda et al. 2010, Kelarev, Dazeley, Stranieri, Yearwood & Jelinek 2012, Kelarev, Stranieri, Yearwood & Jelinek 2012).

The paper is organised as follows. Section 2 describes the Diabetes Complications Screening Research Initiative (DiScRi) organised at Charles Sturt University,2 and the corresponding data set. Section 3 contains background information on cardiac autonomic neuropathy. Section 4 deals with the methods used in our experiments. Section 6 presents the experimental results and discussion comparing the efficiencies of several base classifiers and multi-level meta classifiers based on Random Forest for this application. The conclusions are presented in Section 7.

2 Diabetes Complications Screening Research Initiative

In order to investigate the data mining algorithms for diabetes patients, we used a large database of test results and health-related parameters collected at the Diabetes Complications Screening Research Initiative (DiScRi) organised at Charles Sturt University (Cornforth & Jelinek 2007). Many patients suffering from diabetes develop complications that require 24/7 cardiac monitoring.

The collection and analysis of data in the project has been approved by the Ethics in Human Research Committee of the university. 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 conducted from 9:00am until 12midday and were recorded in the DiScRi database along with various other clinical 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 important set of features recorded for CAN determination is the *Ewing battery* (Ewing et al. 1980, 1985). 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 the ten second samples of 12-lead ECG for all participants were extracted from the database. These included the QRS, PQ, QTc and QTd intervals, heart rate and QRS axis explained below. The QRS complex reflects the depolarization of the ventricles of the heart. The duration of the QRS complex is called the QRS duration. The time from the beginning of the P wave until the start of the next QRS complex is called the PQ interval. The longest distance from the Q wave to the next T wave is called the QT 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 is known as the QT dispersion denoted by QTd. It is used as an indicator of repolarisation of ventricular. The deflection of the electrical axis of the heart measured in degrees to the right or left is called the QRS axis.

Several expert editing rules were used to reduce the number of missing values in the database. These rules were collected during discussions with the experts maintaining the database. Preprocessing of data using these rules produced 1029 complete rows with complete values of all fields, which were used for the experimental evaluation of the performance of data mining algorithms. The whole database contained over 200 features.

3 Cardiac Autonomic Neuropathy

Cardiac autonomic neuropathy (CAN) is a condition associated with damage to the autonomic nervous system innervating the heart (Ewing et al. 1980, 1985, Khandoker et al. 2009). The classification of disease progression associated with CAN is important, because it has implications for planning of timely treatment, which can lead to an improved well-being of the patients and a reduction in morbidity and mortality associated with cardiac arrhythmias in diabetes. The most important tests required for identification of CAN rely on assessing responses in heart rate and blood pressure to various activities, usually consisting of tests described by Ewing et al. (1980, 1985): lying to standing heart rate change (LSHR), deep breathing heart rate change (DBHR), valsalva manoeuvre heart rate change (VAHR), hand grip blood pressure change (HGBP), lying to standing blood pressure change (LSBP). QRS width has also been shown to be indicative of CAN (Fang et al. 2004) and is also included. For discussion of the outcomes of our experiments we use the acronyms for the DiScRi features listed in Figure 1. The same acronyms are used in the original DiScRi database.

Acronym	Feature
LSHR	Lying to standing heart rate change
LSHRresu	Categorical variable based on LSHR defined by Ewing et al. (1980)
DBHR	Deep breathing heart rate change
DBHRresu	Categorical variable based on LSHR defined by Ewing et al. (1980)
VAHR	Valsalva manoeuvre heart rate change
VAHRresu	Categorical variable based on LSHR defined by Ewing et al. (1980)
HGBP	Hand grip blood pressure change
LSBP	Lying to standing blood pressure change
QRSaxis	QRS axis degree 10sec
SLHR	Standing to lying heart rate
QRS 10sec	QRS duration
PQ 10sec	PQ duration
QTc 10sec	Corrected QT interval duration
QTd 10sec	QT dispersion

Figure 1: Acronyms for several features in DiScRi database

We investigated three original classifications of cardiac autonomic neuropathy progression introduced by Ewing et al. (1980, 1985). They have 2, 3 and 4 classes, respectively. The first one divides all patients into two classes allocating each patient either to the 'normal' class, or to 'definite' class. The second one divides all patients into three classes allocating each patient to one of the following classes: 'normal', 'early', 'definite'. The fourth classification divides all patients into four classes, allocated each patient to one of the following classes: 'normal', 'early', 'definite', and 'severe'.

4 Methods

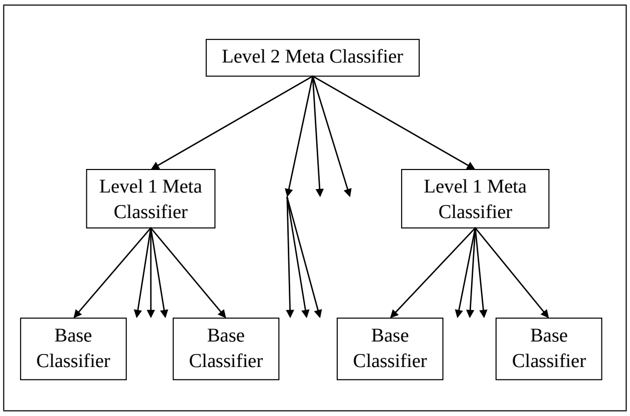

Figure 2: Multi-level meta classifiers

4.1 Random Forest

Random Forest plays a special role in this paper, and so we introduce it in a separate subsection. Random Forest is an ensemble meta classifier hardwired to a particular base classifier, Random Tree. It constructs a forest of random trees following Breiman (2001) building many decision tree predictors with randomly selected variable subsets and utilizing a different subset of training and validation data for each individual model. After generating many trees, the resulting class prediction is based on votes from the single trees. Consequently, lower ranked variables are eliminated based on empirical performance heuristics (Han et al. 2006). We used Random Forest feature selection in R (version 2.15.1) with Rattle (Williams 2009, 2011). Weka implementation of Random Forest was used to combine it with other meta classifiers available in Weka via SimpleCLI. (This implementation can handle missing values.) In applying the Random Forest classifier and its feature selection we followed the recommendations and conclusions based on previous experiments for a different database of cardiac patients presented by Van et al. (2011).

4.2 Base Classifiers

We tested many preliminary base classifiers available in Weka (Hall et al. 2009) and have chosen the following classifiers for a series of complete tests with outcomes presented in this paper. These robust classifiers performed well for DiScRi data set during our initial testing. They represent several essential categories of classifiers.

- *DecisionTable* builds and uses a decision table majority classifier (Kohavi 1995).
- *FURIA* is a fuzzy unordered rule induction algorithm due to Huehn & Huellermeier (2009).
- *J48* generates a pruned or unpruned C4.5 decision tree (Quinlan 1993).
- *NBTree* uses a decision tree with naive Bayes classifiers at the leaves (Kohavi 1996).
- *SMO* uses Sequential Minimal Optimization for training a support vector classifier (Hastie & Tibshirani 1998, Keerthi et al. 2001, Platt 1998). Initially, we tested all kernels of SMO available in Weka and used it with polynomial kernel that performed best for our data set.

4.3 Meta Classifiers

We investigate the performance of the following meta classifiers: Bagging, Boosting, Dagging, Decorate, Grading, HBGF, MultiBoost and Stacking.

- *Bagging* (bootstrap aggregating), generates a collection of new sets by resampling the given training set at random and with replacement. These sets are called *bootstrap samples*. New classifiers are then trained, one for each of these new training sets. They are amalgamated via a majority vote (Breiman 1996, Liang & Zhang 2011).
- *Boosting* trains several classifiers in succession. Every next classifier is trained on the instances that have turned out more difficult for the preceding classifier. To this end all instances are assigned weights, and if an instance turns out difficult to classify, then its weight is increased

at the next boosting step. We used highly successful AdaBoost classifier described by Freund & Schapire (1996).

- *Consensus functions* can be used as a replacement for voting to combine the outputs of classifiers in the ensemble. Here we used the HBGF consensus function, following the recommendations of Fern & Brodley (2004) and our previous experience with consensus functions for other data sets (Yearwood et al. 2009). It utilizes a bipartite graph with two sets of vertices: clusters and elements of the data set. A cluster C and an element d are connected by an edge in this bipartite graph if and only if d belongs to C. An appropriate graph partitioning algorithm is then applied to the whole bipartite graph, and the final clustering is determined by the way it partitions all elements of the data set.

- *Dagging* is useful in situations where the base classifiers are slow. It divides the training set into a collection of disjoint (and therefore smaller) stratified samples, trains copies of the same base classifier and averages their outputs using vote (Ting & Witten 1997).

- *Decorate* constructs special artificial training examples to build diverse base classifiers (Melville & Mooney 2005).

- *Grading* trains meta-classifiers, which grade the output of base classifiers as correct or wrong labels, and these graded outcomes are then combined (Seewald & Fuernkranz 2001).

- *MultiBoost* extends the approach of AdaBoost with the wagging technique (Webb 2000). Wagging is a variant of bagging where the weights of training instances generated during boosting are utilized in selection of the bootstrap samples (Bauer & Kohavi 1999).

- *Stacking* can be regarded as a generalization of voting, where meta-learner aggregates the outputs of several base classifiers (Wolpert 1992).

4.4 Multi-level Meta Classifiers

The main focus of this paper is on a systematic investigation of several novel multi-level meta classifiers for DiScRi data set. These classifiers have not been considered in the literature before, since personal computers regularly used in research have only recently become powerful enough to train them for large data sets. It turns out easy to set up and use these multi-level meta classifiers in Weka SimpleCLI command line. To demonstrate how such classifiers can be set up and executed, we include Figure 7 with complete commands used in SimpleCLI to run two very best options in our experiments and Figure 8, which shows how to enter these commands, see also Section 6. Our experiments compared several multi-level meta classifiers with two levels and various base classifiers. However, the best results were obtained by classifiers, which can also be viewed as multi-level meta classifiers with three levels of ensembles, since they are based on Random Forest, see Section 6 and Subsection 4.1.

5 Measures of Performance of Classifiers

We looked at several standard measures of performance of classifiers: Area Under Curve, accuracy, precision, recall, sensitivity and specificity. Following Van et al. (2011), we used the Area Under Curve, AUC, as the main measure of performance of classifiers. It is also known as the Receiver Operating Characteristic or ROC area. Let us refer to Van et al. (2011) for more detailed discussion of measures of performance and references to other relevant articles. In the rare cases where two classifiers produced equal AUC, or where one classifier performed with the same AUC for two sets of features, to finetune the ordering of such cases we used accuracy of the classification as the second important metric to guide our experiments. The tables presenting the results of our experiments in this paper contain only AUC as the main measure of performance.

Here we include only a brief overview of the measures we used conducting our experiments, since there is a variety of terms used to discuss clinical experiments. Notice that for multi-class classifiers, like those considered in the present article, weighted average values of the performance metrics are usually used. This means that they are calculated for each class separately, and a weighted average is found then. In particular, our tables in this paper include the weighted average values of AUC over all classes. In contrast, the *accuracy* is defined for the whole classifier as the percentage of all patients classified correctly, which means that this definition does not involve weighted averages in the calculation. The accuracy can be expressed as the probability that the prediction of the classifier for an individual patient is correct.

The *Area Under Curve*, AUC, for a given class, is an area under the ROC graph that plots true positive rates for this class against false positive rates for a series of cut-off values. Equivalently, the ROC graph can be defined as a curve graphically displaying the trade-off between sensitivity and specificity for each cut-off value. *Sensitivity* is the proportion of positives (patients with CAN) that are identified correctly. *Specificity* is the proportion of negatives (patients without CAN) that are identified correctly.

Sensitivity and specificity are measures evaluating binary classifications. For multi-class classifications they can be also used with respect to one class and its complement. Sensitivity is also called *True Positive Rate*. *False Positive Rate* is equal to 1 - specificity. These measures are related to recall and precision. *Precision* of a classifier, for a given class, is the ratio of true positives to combined true and false positives. *Recall* is the ratio of true positives to the number of all positive samples (i.e., to the combined true positives and false negatives). The recall calculated for the class of patients with CAN is equal to sensitivity of the whole classifier.

For example, in the case of the two-class classification of CAN For the class of patients with CAN, the precision is the ratio of the number of patients correctly identified as having CAN to the number of all patients identified as having CAN. For the cohort of patients without CAN, the precision is the ratio of the number of patients correctly identified as having no CAN to the number of all patients identified as free from CAN. The precision of the classifier as a whole is a weighted average of its precisions for these classes.

Likewise, for the class of patients with CAN, the recall is the ratio of the number of patients correctly identified as having CAN to the number of all patients with CAN. For the cohort of patients without CAN, the recall is the ratio of the number of patients correctly identified as being free from CAN to the number of all patients without CAN. The recall of the classifier is a weighted average of its recalls for both classes.

6 Experiments and Discussion

We used Rattle (Williams 2009) and R software (Williams 2011) for Random Forest feature selection, and Weka SimpleCLI command line to train and test classifiers and meta classifiers. One of the standard options for preventing overfitting is 10-fold cross validation. It is implemented in Weka and is invoked in SimpleCLI by default as stratified 10-fold cross validation. (Default stratified 10-fold cross validation can be switched off or modified by indicating command line arguments -no-cv, -split-percentage and -preserve-order in SimpleCLI.) It divides data into ten stratified folds and creates training sets and hold out testing sets ten times for ten consequtive tests with hold out sets automatically. Thus, we used 10-fold cross validation to assess the performance of various base classifiers, meta-classifiers and multi-level meta classifiers. All tables with outcomes included in this paper contain average performance against the validate sets found in stratified 10-fold cross validation.

First, we tested the performance of DecisionTable, FURIA, J48, NBTree, RandomForest and SMO for all subsets of the Ewing battery, which is the set of the most important features. All of these base classifiers are available in Weka Explorer, and we use Weka to test them. These experiments demonstrated that Random Forest consistently outperformed all other classifiers for all of these subsets of features. To illustrate these results, we include only Table 1.

	Number of classes		
	2	3	4
DecisionTable	0.905	0.900	0.897
FURIA	0.942	0.936	0.932
J48	0.933	0.930	0.922
NBTree	0.923	0.917	0.914
RandomForest	0.982	0.977	0.973
SMO	0.861	0.856	0.854

Table 1: AUC of base classifiers for the subset LSHR, DBHR, VAHR, LSBP of Ewing features

Then we used Random Forest feature selection in Rattle (Williams 2009). It produced feature ranking presented in Figure 3. We tested the performance of Random Forest for all sets beginning with the most significant feature and adding more features in the order of their significance. This demonstrated that the best set of features consists of the first 8 attributes: DBHR, VAHRresu, VAHR, DBHRresu, LSHRresu, LSHR, HGBP, QRS axis (degree) 10sec.

We tested all base classifiers for this set of 8 features too. The results of these experiments are given in Table 2 and Figure 4. The outcomes show that Random Forest remains the best classifier for this set of attributes too. Thus, in all our tests Random Forest has consistently performed as the very best base classifier for all sets of features of DiScRi database. We see that Random Forest feature selection has made it possible to improve the outcomes significantly.

Next, we used SimpleCLI command line in Weka to investigate the performance of meta classifiers in their ability to achieve further improvement to performance. We tested the following meta classifiers: AdaBoost, Bagging, Dagging, Decorate, Grading, HBGF, MultiBoost, and Stacking. Our tests have also shown that the outcomes remained better when Random Forest was used as a base classifier for these meta classifiers and that the results became worse when Random Forest was replaced by other base classifiers. We conduced complete set of evaluations of the meta classifiers based on Random Forest. These results are included in Table 3 and Figure 5. We see that AdaBoost, Bagging, Decorate and MultiBoost performed better than other meta classifiers.

Finally, for the four meta classifiers that performed well in the previous step, we investigated all their multi-level combinations. The experimental results comparing the performance of multi-level meta classifiers are presented in Table 4 and Figure 6. To provide more details on how these multi-level classifiers can be set up and executed, we include Figure 7 with

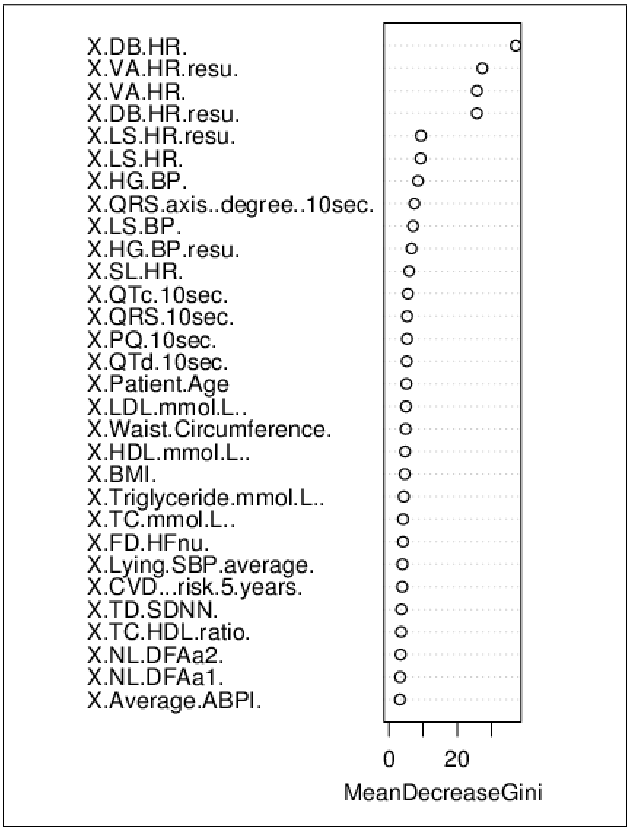

Figure 3: Random Forest feature ranking

	Number of classes		
	2	3	4
DecisionTable	0.963	0.960	0.956
FURIA	0.979	0.976	0.968
J48	0.967	0.964	0.959
NBTree	0.980	0.975	0.972
RandomForest	0.990	0.987	0.981
SMO	0.954	0.949	0.944

Table 2: AUC of base classifiers for the best subset DBHR, VAHRresu, VAHR, DBHRresu, LSHRresu, LSHR, HGBP, QRS axis (degree) 10sec.

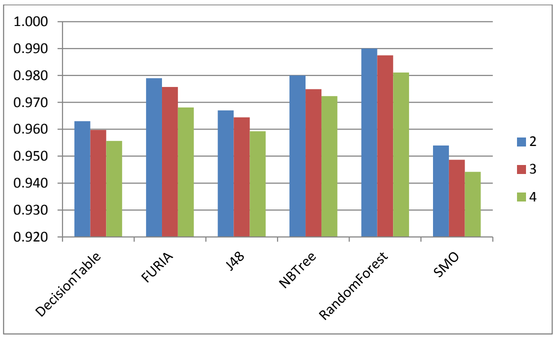

Figure 4: Base classifiers

	Number of classes		
	2	3	4
AdaBoost	0.987	0.984	0.979
Bagging	0.995	0.990	0.987
Dagging	0.971	0.967	0.964
Decorate	0.996	0.993	0.988
Grading	0.969	0.965	0.960
HBGF	0.974	0.969	0.967
MultiBoost	0.988	0.984	0.981
Stacking	0.978	0.976	0.969

Table 3: AUC of meta classifiers based on Random Forest

SimpleCLI command line arguments used to run two very best options given in our Table 4. These results show that several multi-level combinations of ensemble classifiers made additional improvements and produced very good outcomes in Table 4. The very best result was obtained by two options combining Bagging and Decorate into one multi-level ensemble classifier. In the first option Bagging was used in the 2nd level after applications of Decorate based on Random Forest in the first level. In the second option Decorate was used in the 2nd level to combine the results of Bagging applied to Random Forest as a base classifier.

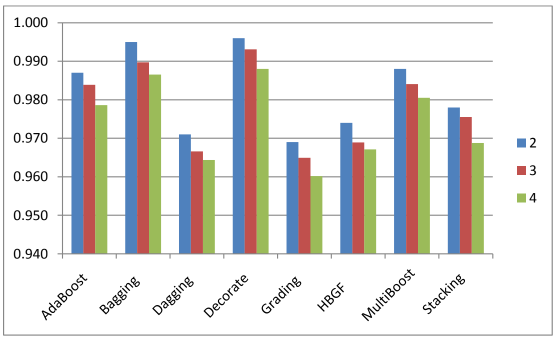

Figure 5: Meta classifiers based on Random Forest

		Number of classes		
Level 2	Level 1	2	3	4
AdaBoost	Bagging	0.990	0.987	0.984
AdaBoost	Decorate	0.992	0.988	0.986
AdaBoost	MultiBoost	0.989	0.987	0.982
Bagging	AdaBoost	0.994	0.990	0.987
Bagging	Decorate	0.997	0.993	0.990
Bagging	MultiBoost	0.994	0.992	0.988
Decorate	AdaBoost	0.996	0.992	0.989
Decorate	Bagging	0.997	0.994	0.990
Decorate	MultiBoost	0.996	0.992	0.990
MultiBoost	AdaBoost	0.989	0.986	0.983
MultiBoost	Bagging	0.985	0.982	0.979
MultiBoost	Decorate	0.990	0.987	0.983

Table 4: AUC of multi-level meta classifiers based on Random Forest

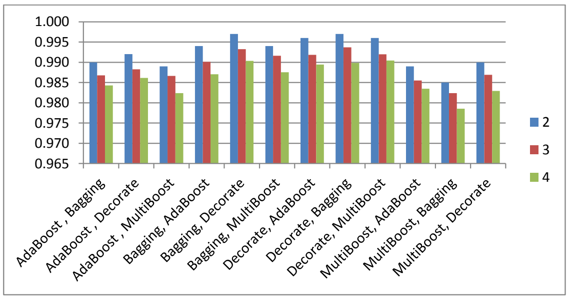

Figure 6: Multi-level meta classifiers

2 levels	SimpleCLI command line
Decorate, Bagging	java weka.classifiers.meta.Decorate -E 10 -R 1.0 -S 1 -I 10 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.RandomForest – -I 10 -K 0 -S 1
Bagging, Decorate	java weka.classifiers.meta.Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.meta.Decorate – -E 10 -R 1.0 -S 1 -I 10 -W weka.classifiers.trees.RandomForest – -I 10 -K 0 -S 1

Figure 7: SimpleCLI command lines with parameters of two best multi-level meta classifiers

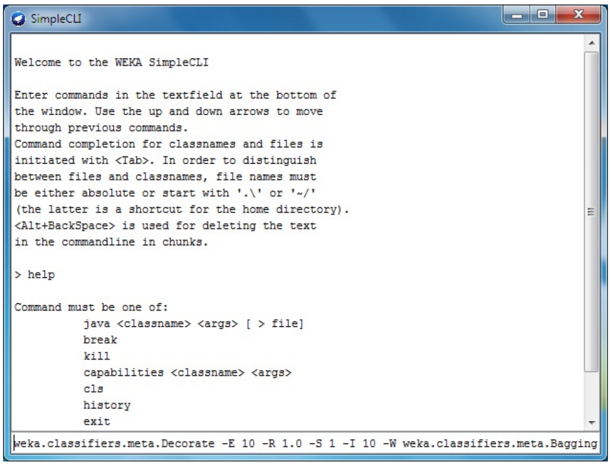

Figure 8: Multi-level meta classifier in SimpleCLI

7 Conclusion

Our experiments demonstrated that for DiScRi data set Random Forest consistently produced better outcomes than all other base classifiers. Feature selection based on the ranking obtained by the implementation of Random Forest in Rattle further improved the outcomes of all classifiers, and again Random Forest produced the best outcomes for the set of features obtained. Finally, the results show that meta classifiers and multi-level ensemble meta classifiers can be used to improve the classifications even more. The best outcomes have been obtained by the novel combined multi-level ensemble classifiers combining Bagging and Decorate based on Random Forest. These methods can be recommended for the monitoring of cardiac autonomic neuropathy progression in those situations where the energy and memory used are not an issue. In situations where it is very important to conserve the energy and use less memory, as it is the case for example, in mobile applications, then Random Forest can be recommended, since it has also produced excellent outcomes.

DiScRi is a very large and unique data set containing a comprehensive collection of tests related to CAN. Using Random Forest feature selection and multi-level meta classifiers has made it possible to achieve a serious improvement in performance compared with outcomes obtained in previous publications using only basic decision trees for classification.

The level of performance of multi-level classifiers for DiScRi data set is also quite good in comparison with the outcomes obtained recently for other data sets in closely related areas using different methods, for example, by Kang et al. (2006), Kelarev et al. (2006), Jelinek et al. (2010, 2011), Yearwood et al. (2008).

In conclusion, let us note that Random Forest is also an ensemble classifier hard wired to a particular base classifier, Random Tree. Therefore, in fact the multi-level meta classifiers included in Table 4 can be considered as ensemble classifiers with three levels where ensemble methods are used.

Acknowledgements

This work was supported by a Deakin-Ballarat collaboration grant.

The authors are grateful to three referees for comments and corrections that have helped to improve the paper.

References

Al-Oqaily, A., Kennedy, P., Catchpoole, D. & Simoff, S. (2008), Comparison of visualization methods of genome-wide SNP profiles in childhood acute lymphoblastic leukaemia, *in* J. Roddick, J. Li, P. Christen & P. Kennedy, eds, 'Data Mining and Analytics 2008, Proceedings of the Seventh Australasian Data Mining Conference, AusDM 2008', Vol. 87 of *CRPIT*, ACS, Glenelg, South Australia, pp. 111–121.

Bauer, E. & Kohavi, R. (1999), 'An empirical comparison of voting classification algorithms: Bagging, boosting, and variants', *Machine Learning* **36**, 105–139.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**, 123–140.

Breiman, L. (2001), 'Random Forests', *Machine Learning* **45**, 5–32.

Christen, P. (2007), A two-step classification approach to unsupervised record linkage, *in* P. Christen, P. Kennedy, J. Li, I. Kolyshkina & G. Williams, eds, 'Sixth Australasian Data Mining Conference (AusDM 2007)', Vol. 70 of *CRPIT*, ACS, Gold Coast, Australia.

Cornforth, D. & Jelinek, H. (2007), 'Automated classification reveals morphological factors associated with dementia', *Applied Soft Computing* **8**, 182–190.

Ewing, D., Campbell, J. & Clarke, B. (1980), 'The natural history of diabetic autonomic neuropathy', *Q. J. Med.* **49**, 95–100.

Ewing, D., Martyn, C., Young, R. & Clarke, B. (1985), 'The value of cardiovascular autonomic function tests: 10 years experience in diabetes', *Diabetes Care* **8**, 491–498.

Fang, Z., Prins, J. & Marwick, T. (2004), 'Diabetic cardiomyopathy: evidence, mechanisms, and therapeutic implications', *Endocr. Rev.* **25**, 543–567.

Fern, X. & Brodley, C. (2004), Solving cluster ensemble problems by bipartite graph partitioning, *in* '21st International Conference on Machine Learning, ICML'04', Vol. 69, ACM, New York, NY, USA, pp. 36–43.

Freund, Y. & Schapire, R. (1996), Experiments with a new boosting algorithm, *in* 'Proc. 13th Internat. Conf. Machine Learning', pp. 148–156.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. (2009), 'The WEKA data mining software: an update', *SIGKDD Explorations* **11**, 10–18.

Han, L., Embrechts, M., Szymanski, B., Sternickel, K. & Ross, A. (2006), Random forests feature selection with K-PLS: detecting ischemia from magnetocardiograms, *in* 'Proc. European Symposium on Artificial Neural Networks, ESANN', Vol. 14, pp. 221–226.

Hastie, T. & Tibshirani, R. (1998), Classification by pairwise coupling, *in* 'Advances in Neural Information Processing Systems'.

Huda, S., Jelinek, H., Ray, B., Stranieri, A. & Yearwood, J. (2010), 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.

Huehn, J. & Huellermeier, E. (2009), 'FURIA: An algorithm for unordered fuzzy rule induction', *Data Mining and Knowledge Discovery* **19**, 293–319.

Islam, R. & Abawajy, J. (2012), 'A multi-tier phishing detection and filtering approach', *Journal of Network and Computer Applications* p. to apper soon.

Jelinek, H., Khandoker, A., Palaniswami, M. & McDonald, S. (2010), 'Heart rate variability and QT dispersion in a cohort of diabetes patients', *Computing in Cardiology* **37**, 613–616.

Jelinek, H., Rocha, A., Carvalho, T., Goldenstein, S. & Wainer, J. (2011), Machine learning and pattern classification in identification of indigenous retinal pathology, *in* 'Proceedings IEEE Conference Eng. Med. Biol. Soc.', pp. 5951–5954.

Jiangning, S., Tan, H., Wang, M., Webb, G. & Akutsu, T. (2012), 'TANGLE: Two-level support vector regression approach for protein backbone torsion angle prediction from primary sequences', *PLoS ONE* **7**, e30361.

Kang, B., Kelarev, A., Sale, A. & Williams, R. (2006), A new model for classifying DNA code inspired by neural networks and FSA, *in* 'Advances in Knowledge Acquisition and Management', Vol. 4303 of *Lecture Notes in Computer Science*, pp. 187–198.

Keerthi, S., Shevade, S., Bhattacharyya, C. & Murthy, K. (2001), 'Improvements to Platt's SMO algorithm for SVM classifier design', *Neural Computation* **13**(3), 637–649.

Kelarev, A., Dazeley, R., Stranieri, A., Yearwood, J. & Jelinek, H. (2012), Detection of CAN by ensemble classifiers based on Ripple Down Rules, *in* 'Pacific Rim Knowledge Acquisition Workshop, PKAW2012', Vol. 7457 of *Lecture Notes in Artificial Intelligence*, pp. 147–159.

Kelarev, A., Kang, B. & Steane, D. (2006), Clustering algorithms for ITS sequence data with alignment metrics, *in* 'AI 2006: Advances in Artificial Intelligence, 19th Australian Joint Conference on Artificial Intelligence', Vol. 4304 of *Lecture Notes in Artificial Intelligence*, pp. 1027–1031.

Kelarev, A., Stranieri, A., Yearwood, J. & Jelinek, H. (2012), 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.

Kennedy, P., Simoff, S., Catchpoole, D., Skillicorn, D., Ubaudi, F. & Al-Oqaily, A. (2008), Integrative visual data mining of biomedical data: Investigating cases in chronic fatigue syndrome and acute lymphoblastic leukaemia, *in* 'Visual Data Mining', Vol. 4404/2008 of *Lecture Notes in Computer Science*, pp. 367–388.

Khandoker, A., Jelinek, H. & Palaniswami, M. (2009), '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.

Kohavi, R. (1995), The power of decision tables, *in* '8th European Conference on Machine Learning', pp. 174–189.

Kohavi, R. (1996), Scaling up the accuracy of Naive-Bayes classifiers: A Decision-Tree hybrid, *in* 'Second International Conference on Knoledge Discovery and Data Mining', pp. 202–207.

Li, J., Fu, A. & Fahey, P. (2009), 'Efficient discovery of risk patterns in medical data', *Artificial Intelligence in Medicine* **45**, 77–89.

Liang, G. & Zhang, C. (2011), Empirical study of bagging predictors on medical data, *in* P. Vamplew, A. Stranieri, K.-L. Ong, P. Christen & P. J. Kennedy, eds, 'Australasian Data Mining Conference, AusDM 2011', Vol. 121 of *CRPIT*, ACS, Ballarat, Australia, pp. 31–40.

Madjarov, G., Gjorgjevikj, D. & Delev, T. (2011), Efficient two stage voting architecture for pairwise multi-label classification, *in* J. Li, ed., 'AI 2010: Advances in Artificial Intelligence', Vol. 6464 of *Lecture Notes in Artificial Intelligence*, pp. 164–173.

Melville, P. & Mooney, R. (2005), 'Creating diversity in ensembles using artificial data', *Information Fusion* **6**, 99–111.

Platt, J. (1998), Fast training of support vector machines using sequential minimal optimization, *in* 'Advances in Kernel Methods – Support Vector Learning'.

Quinlan, R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Seewald, A. & Fuernkranz, J. (2001), An evaluation of grading classifiers advances in intelligent data analysis, *in* 'Advances in Intelligent Data Analysis', Vol. 2189/2001 of *Lecture Notes in Computer Science*, pp. 115–124.

Shouman, M., Turner, T. & Stocker, R. (2011), Using decision tree for diagnosing heart disease patients, *in* P. Vamplew, A. Stranieri, K.-L. Ong, P. Christen & P. Kennedy, eds, 'Australasian Data Mining Conference, AusDM 11', Vol. 121 of *CRPIT*, ACS, Ballarat, Australia, pp. 23–30.

Sinha, A., Tayebi, H., Krishnaswamy, S., Waluyo, A. & Gaber, M. (2011), Resource-aware ECG analysis on mobile devices, *in* 'Proceedings of the 2011 ACM Symposium on Applied Computing, SAC11', pp. 1012–1013.

Sun, B., Zhu, Z., Li, J. & Linghu, B. (2011), 'Combined feature selection and cancer prognosis using support vector machine regression', *IEEE/ACM Transactions on Computational Biology Bioinformatics* **8**, 1671–1677.

Tayebi, H., Krishnaswamy, S., Waluyo, A., Sinha, A. & Gaber, M. (2011), RA-SAX: Resource-aware symbolic aggregate approximation for mobile ECG analysis, *in* 'Proceedings of the 12th IEEE International Conference on Mobile Data Management, MDM11', Vol. 1, pp. 289–290.

Ting, K., Wells, J., Tan, S., Teng, S. & Webb, G. (2009), FaSS: Ensembles for stable learners, *in* 'MCS 2009: 8th International Workshop on Multiple Classifier Systems,', Vol. 5519 of *Lecture Notes in Compuer Science*, pp. 364–374.

Ting, K., Wells, J., Tan, S., Teng, S. & Webb, G. (2011), 'Feature-subspace aggregating: Ensembles for stable and unstable learners', *Machine Learning* **82(3)**, 375–397.

Ting, K. & Witten, I. (1997), Stacking bagged and dagged models, *in* 'Fourteenth international Conference on Machine Learning', pp. 367–375.

Van, A., Gay, V., Kennedy, P., Barin, E. & Leijdekkers, P. (2011), Understanding risk factors in cardiac rehabilitation patients with random forests and decision trees, *in* P. Vamplew, A. Stranieri, K.-L. Ong, P. Christen & P. Kennedy, eds, 'Australasian Data Mining Conference, AusDM 2011', Vol. 121 of *CRPIT*, ACS, Ballarat, Australia, pp. 11–22.

Webb, G. (2000), 'Multiboosting: A technique for combining boosting and wagging', *Machine Learning* **40**, 159 – 196.

Webb, G. (2008), Multi-strategy ensemble learning, ensembles of bayesian classifiers, and the problem of false discoveries, *in* J. Roddick, J. Li, P. Christen & P. Kennedy, eds, 'Data Mining and Analytics 2008, Proceedings of the Seventh Australasian Data Mining Conference, AusDM 2008', Vol. 87 of *CRPIT*, ACS, Glenelg, South Australia, p. 15.

Webb, G. & Zheng, Z. (2004), 'Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques', *IEEE Transactions on Knowledge and Data Engineering* **16**, 980–991.

Wickramasinghe, K., Alahakoon, D., Schattner, P. & Georgeff, M. (2011), Self-organizing maps for translating health care knowledge: A case study in diabetes management, *in* D. Wang & M. Reynolds, eds, 'AI 2011: Advances in Artificial Intelligence', Vol. 7106 of *Lecture Notes in Artificial Intelligence*, pp. 162–171.

Williams, G. (2009), 'Rattle: a data mining GUI for R', *The R Journal* **1**, 45–55.

Williams, G. (2011), *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!)*, Springer, New York, Dordrecht, Heidelberg, London.

Wolpert, D. (1992), 'Stacked generalization', *Neural Networks* **5**, 241–259.

Wolpert, D. (1996), 'The lack of a priori distinctions between learning algorithms', *Neural Computation* **8**, 1341–1390.

Yang, Y., Korb, K., Ting, K.-M. & Webb, G. (2005), Ensemble selection for superparent-one-dependence estimators, *in* 'AI 2005: Advances in Artificial Intelligence, 18th Australiasian Joint Conference on Artificial Intelligence', Vol. 3809 of *Lecture Notes in Artificial Intelligence*, pp. 102–111.

Yearwood, J., Kang, B. & Kelarev, A. (2008), Experimental investigation of classification algorithms for ITS dataset, *in* 'Pacific Rim Knowledge Acquisition Workshop, PKAW 2008', Hanoi, Vietnam, 15–16 December 2008, pp. 262–272.

Yearwood, J., Webb, D., Ma, L., Vamplew, P., Ofoghi, B. & Kelarev, A. (2009), Applying clustering and ensemble clustering approaches to phishing profiling, *in* P. Kennedy, K. Ong & P. Christen, eds, 'Data Mining and Analytics 2009, Proc. 8th Australasian Data Mining Conference, AusDM 2009', Vol. 101 of *CRPIT*, ACS, Melbourne, Australia, pp. 25–34.