Elsevier Editorial System(tm) for Computers in Biology and Medicine
Manuscript Draft

Manuscript Number: CBM-D-12-00474R2

Title: Predicting cardiac autonomic neuropathy category for diabetic data with missing values

Article Type: Full Length Article

Keywords: cardiac autonomic neuropathy; diabetes; missing value imputation; ensemble classifiers; regression learners; meta regression methods; Ewing formula;

Corresponding Author: Dr. Andrei Kelarev, PhD

Corresponding Author's Institution: Deakin University

First Author: Jemal Abawajy, PhD

Order of Authors: Jemal Abawajy, PhD; Andrei Kelarev, PhD; Morshed Chowdhury, PhD; Andrew Stranieri, PhD; Herbert F Jelinek, PhD

Abstract: Cardiovascular autonomic neuropathy (CAN) is a serious and well known complication of diabetes. Previous articles circumvented the problem of missing values in CAN data by deleting all records and fields with missing values and applying classifiers trained on different sets of features that were complete. Most of them also added alternative features to compensate for the deleted ones. Here we introduce and investigate a new method for classifying CAN data with missing values. In contrast to all previous papers, our new method does not delete attributes with missing values, does not use classifiers, and does not add features. Instead it is based on regression and meta regression combined with the Ewing formula for identifying the classes of CAN. This is the first article using the Ewing formula and regression to classify CAN. We carried out extensive experiments to determine the best combination of regression and meta regression techniques for classifying CAN data with missing values. The best outcomes have been obtained by the additive regression meta learner based on M5Rules and combined with the Ewing formula. It has achieved the best accuracy of 99.78\% for two classes of CAN, and 98.98\% for three classes of CAN. These outcomes are substantially better than previous results obtained in the literature by deleting all missing attributes and applying traditional classifiers to different sets of features without regression. Another advantage of our method is that it does not require practitioners to perform more tests collecting additional alternative features.

Author Biography

Andrew Stranieri is an Associate Professor and the Director of the Centre for Informatics and Applied Optimisation at the University of Ballarat. His research into cognitive models of argumentation and artificial intelligence was instrumental in modelling decision making in refugee law, copyright law, eligibility for legal aid and sentencing. His research in health informatics spans data mining in health, complementary and alternative medicine informatics, telemedicine and intelligent decision support systems. Andrew Stranieri is the author of over 120 peer reviewed journal and conference articles and has published two books.

Jemal H. Abawajy is a Professor and the Director of the Parallel and Distributing Computing Lab at Deakin University, Australia. Prof. Abawajy is a senior member of IEEE and was a member of the organizing committees for over 100 international conferences serving in various capacities including chair, general co-chair, vice-chair, best paper award chair, publication chair, session chair and program committee member. Prof. Abawajy has published more than 200 refereed articles, supervised numerous PhD students to completion and is on the editorial boards of many journals.

Andrei Kelarev is an author of two books and 194 journal articles, and an editor of 5 international journals. Andrei worked as an Associate Professor in the University of Wisconsin and University of Nebraska in USA and as a Senior Lecturer in the University of Tasmania in Australia. Andrei was a Chief Investigator of a large Discovery grant from Australian Research Council and was a member of the program committees of several conferences. Andrei Kelarev is working for a research grant at the Parallel and Distributing Computing Lab at Deakin University, Australia.

Morshed U. Chowdhury is a faculty of School of Information Technology, and the founder member of the Parallel and Distributing Computing Lab at Deakin University, Australia. He has organised IEEE sponsored ACIS2007 and SNPD2012 conferences as a Co-Chairs. Dr. Chowdhury was the member of the organizing committees for over 50 international conferences serving in various capacities. He received the best research paper award in the Annual International conference on "Information Technology Security (ITS2010)" in 2010. He has published more than 60 refereed articles and also member of the editorial board of International Journal of Software Innovation (IJSI), IGI Global, US. He has been supervising seven PhD. Students.

Herbert F. Jelinek is a Clinical Associate Professor with the Australian School of Advanced Medicine, Macquarie University, Sydney, Australia, and a member of the Centre for Research in Complex Systems, Charles Sturt University, Albury, Australia. Dr Jelinek is currently a visiting Associate Professor at Khalifa University of Science, Technology and Research, Abu Dhabi, UAE. Herbert Jelinek received the B.Sc. (Hons.) degree in human genetics from the University of New South Wales, Sydney, Australia, in 1984, followed by the Graduate Diploma in neuroscience from the Australian National University, Canberra, Australia, in 1986 and the Ph.D. degree in medicine from the University of Sydney, Sydney, Australia, in 1996. He is a member of the IEEE Biomedical Engineering Society and the Australian Diabetes Association.

Predicting cardiac autonomic neuropathy category for diabetic data with missing values

Jemal Abawajy[a], Andrei Kelarev[a], Morshed Chowdhury[a], Andrew Stranieri[b], Herbert F. Jelinek[c]

[a]*School of Information Technology, Deakin University, 221 Burwood Hwy, Victoria 3125, Australia*
[b]*School of Science, Information Technology and Engineering, University of Ballarat, P.O. Box 663, Ballarat, Victoria 3353, Australia*
[c]*School of Community Health, Charles Sturt University, P.O. Box 789, Albury, NSW 2640, Australia*

Abstract

Cardiovascular autonomic neuropathy (CAN) is a serious and well known complication of diabetes. Previous articles circumvented the problem of missing values in CAN data by deleting all records and fields with missing values and applying classifiers trained on different sets of features that were complete. Most of them also added alternative features to compensate for the deleted ones. Here we introduce and investigate a new method for classifying CAN data with missing values. In contrast to all previous papers, our new method does not delete attributes with missing values, does not use classifiers, and does not add features. Instead it is based on regression and meta regression combined with the Ewing formula for identifying the classes of CAN. This is the first article using the Ewing formula and regression to classify CAN. We carried out extensive experiments to determine the best combination of regression and meta regression techniques for classifying CAN data with missing values. The best outcomes have been obtained by the additive regression meta learner based on M5Rules and combined with the

Ewing formula. It has achieved the best accuracy of 99.78% for two classes of CAN, and 98.98% for three classes of CAN. These outcomes are substantially better than previous results obtained in the literature by deleting all missing attributes and applying traditional classifiers to different sets of features without regression. Another advantage of our method is that it does not require practitioners to perform more tests collecting additional alternative features.
Keywords: cardiac autonomic neuropathy, diabetes, missing value imputation, regression learners, meta regression techniques, Ewing formula.
2010 MSC: 68T05, 68T10.

---

1. Introduction

The applications of data mining methods for the development of computer systems analyzing data related to cardiac patients are very important and have been investigated, for example, in the recent articles [1–6]. Cardiac autonomic neuropathy (CAN) is a diabetes complication due to abnormal functioning of the autonomic nervous system, which may be associated with sudden cardiac death ([7, 8]). Based on the results of the cardiac autonomic function tests, CAN progression is categorized as normal pattern, early pattern, definite pattern or severe pattern. Categories of CAN are determined using rules introduced by Ewing and Clarke [9]. These rules can also be referred to as the Ewing formula. Features used in the Ewing formula to define CAN progression are called the Ewing features, or the Ewing fields, or the Ewing attributes, or the Ewing battery. More information on the Ewing battery and the Ewing formula is given in the next section.

The aim of this paper is to introduce and investigate a new method of handling CAN data with missing values of the Ewing features for the detection of CAN

and its progression from normal, early and definite to severe categories. Missing Ewing features are a common occurrence in cardiac autonomic function tests, because many patients are unable to perform some of the tests, as noted in [9]. The task of classifying CAN with missing Ewing attributes has implications for timely treatment. It can lead to an improved prognosis of the patients and a reduction in morbidity and mortality associated with cardiac arrhythmias in diabetes.

Algorithms for the classification of CAN data have been considered recently, for example, in [10–17]. All these articles investigated different classification schemes and applied one and the same general approach. Their approach is different from the method proposed in the present paper. It is described in more details below. All previous articles only circumvented the problem of missing values by deleting all attributes with missing values, deleting all records with missing values from the training set, designing alternative sets of features and applying various classifiers. Here we introduce and apply a new method of handling CAN data with missing values. It utilizes regression and meta regression techniques combined with the Ewing formula.

Let us first briefly explain our new method, and then compare it with the studies undertaken in the previous articles. The main problem is to determine the CAN category of a new instance I_N of CAN data using a database containing instances with known categories. In practice, the instance I_N may come from a new patient or an unclassified database record, where the CAN category has not been indicated yet because of missing values. If all features of the Ewing battery B_{Ewing} are complete in the record I_N, then the Ewing formula can determine the CAN category of I_N. The problem is in treating instances I_N with missing Ewing values.

Suppose that a Ewing feature is missing in I_N. Denote this Ewing feature with a missing value in I_N by F_M. Likewise, by $C_{Ewing}(I_N)$ we denote the set of the Ewing features with complete values in I_N.

In this article, we propose to impute the value of the feature F_M in I_N with a new predicted value and then apply the Ewing formula. To make a prediction, we create a training set by selecting all records with complete Ewing features and Ewing categories. Let us denote this set by S_{train}. A regression or meta regression learner L can be trained on S_{train} to predict the value of the feature F_M in any new instance. The regression learner L can be applied to the new instance I_N. It will produce a predicted value $V(F_M) = L(I_N)$ for the value of the missing feature F_M in I_N. Now, the union of the set $C_{Ewing}(I_N)$ of all complete Ewing features in I_N and the predicted value $V(F_M)$ covers the whole Ewing battery. We can substitute these values in the Ewing formula to derive the CAN category of I_N.

Although our new method is natural, it has never been considered in previous articles. The task of selecting appropriate regression and meta regression techniques as ingredients for our method is quite sophisticated. It is thoroughly investigated in the present article. We conducted comprehensive experiments and identified the best combination of the Ewing formula, regression and meta regression techniques for classifying CAN data with missing values. The best combination produced outcomes, which are substantially better than previous results obtained in the literature. A comparison of the outcomes of our new method with the results of previous papers is given in Section 5.

All previous papers investigated a different general approach to the main problem mentioned above. They deleted the attribute F_M from the instance I_N and from all records in the database, and searched for alternative sets S_A of features,

which have complete values in I_N. Most of the previous papers added additional features to the set S_A to compensate for the loss of accuracy caused by deleting the Ewing attribute F_M. A classifier was then trained on the set of records from the database with complete values of all features in S_A and with given CAN categories. After the training and deletion of the missing feature F_M from the new instance I_N, this classifier was applied to I_N to predict its category. In contrast to all previous papers, our new method does not use classifiers, does not delete attributes with missing values, and does not add alternative features, which would require practitioners to perform more tests.

The rest of the paper is organized as follows. Section 2 gives background information on CAN, the Ewing battery of tests and the Ewing rules/formula used to determine the CAN category. Section 3 contains preliminaries on the regression techniques used in our experiments. The base regression learners investigated in our experiments are presented in Subsection 3.1. Meta regression techniques employed to enhance their performance are covered in Subsection 3.2. Section 4 contains more information on the Diabetes Screening Research Initiative (DiScRi) database and preparing data for experiments. Experimental results and discussion are given in Section 5. A summary of conclusions is contained in Section 6.

2. Background on cardiac autonomic neuropathy

Cardiac autonomic neuropathy (CAN) is a complication of diabetes that involves damage to the autonomic nerve fibres that innervate the heart and blood vessels. The resulting abnormalities in heart rate control and vascular dynamics are thought to account in part for the incidence of sudden cardiac death often observed in people with diabetes [7, 8, 18].

The most important tests required for a risk assessment of CAN rely on responses in heart rate and blood pressure to various activities, usually consisting of the following five Ewing tests described in [9, 19, 20].

(1) Heart rate response to the Valsalva manoeuvre (VAHR); where the patient exhales against 40mmHg pressure while the heart rate is observed.

(2) Heart rate variation during deep breathing (DBHR); where the patient sits quietly and breathes deeply while an electrocardiogram records the heart rate variation over 6 breathing cycles.

(3) Blood pressure response to sustained hand-grip (HGBP); where the systolic blood pressure variation is recorded before and after a sustained hand grip.

(4) Heart rate response to moving from a lying to a standing position (LSHR); where the beat to beat (R-R) interval change in response to standing from a lying position is measured.

(5) Blood pressure response moving from lying to standing (LSBP); where the blood pressure change in response to standing from a lying position is measured.

Table 1 contains the boundary points for each test derived in [9, 19, 20] from physiological evidence in association with in-field trials. These boundary values are also explained by Ewing et al. [20] in great detail. The categorical variables *abnormal*, *borderline* and *normal* are introduced in the Ewing and Clark formulation for each test.

The rules, or the Ewing formula for determining the five categories for a CAN risk assessment, are given in Table 2. These rules were originally defined by Ew-

Test	Value		
	Normal	Borderline	Abnormal
VAHR (ratio)	$\geq$1.21	1.11-1.20	$\leq$1.10
DBHR (beats/min)	$\geq$15	11-14	$\leq$10
HGBP (mmHg)	$\geq$16	11-15	$\leq$10
LSHR (ratio)	$\geq$1.04	1.01-1.03	$\leq$1.00
LSBP (mmHg)	$\leq$10	11-29	$\geq$30

Table 1: Ranges and boundary values determining categorical variables for the Ewing battery.

ing et al. [19, 20]. Ewing et al. [20] also compared the categorization given in Table 2 with two scoring systems used by other researchers: (1) giving 0 for a normal test, 0.5 for a borderline result, and 1 for an abnormal result, thus giving a score of 0 to 5 for each patient; and (2) counting the number of tests that are definitely abnormal, again giving a score of 0 to 5 for each patient. Ewing et al. [20] demonstrated that these scoring systems give roughly equivalent categorizations and seem to carry no real advantages.

It is not always possible for all patients to perform all of the Ewing tests. For instance, the hand grip test may be difficult to do due to arthritis. The lying to standing tests often cannot be done due to mobility challenges and some patients have conditions where forceful breathing required for the Valsalva manoeuvre is contra-indicated. These issues result in CAN risk assessments being made in practice on the basis of only a subset of the Ewing tests ([9, 19, 20]).

Category	Test values
Normal	All tests normal or one borderline.
Early	One of the three heart rate tests abnormal or two borderline
Definite	Two or more of the heart rate tests abnormal.
Severe	Two or more of the heart rate tests abnormal plus one or both of the blood pressure tests abnormal, or both borderline.
Atypical	Any other combination of abnormal tests.

Table 2: CAN categories defined by Ewing et al. [20].

3. Regression and meta regression learners

This sections contains brief preliminaries on the regression and meta regression techniques applied in this paper. Every regression learner is always trained on a training set with complete values to predict the value of one attribute given values of all other features in the set. The way regression and meta regression techniques are used in our experiments is explained in Section 5.

3.1. Base regression learners

This subsection deals with prerequisites on the base regression learners, which are compared as ingredients of our method during experiments. The following robust base regression learners were selected for a complete experimental evaluation of their performance, because they represent the most essential types of regression

techniques.

- *ConjunctiveRule* is a learner of conjunctive rules that consist of a set of antecedents and a consequent. The antecedents are grouped together by conjunction, i.e., logical AND, and the consequent is the class value for classification/regression. The antecedents are chosen in the order of their information gain defined, for classification, in terms of the weighted average of the entropies of the data covered and not covered by the rule and, for regression, in terms of the weighted average of the mean-squared errors of the data covered and not covered by the rule. The distribution of the class labels or means of the numeric class value in the dataset are used as the consequent. Reduced Error Pruning (REP) or simple pre-pruning based on the number of antecedents is applied to the generated rule. If a test instance is not covered by the conjunctive rule, then it is predicted on the basis of the default class distributions/value of the data not covered by the rule in the training set. ConjunctiveRule can predict numeric and nominal class labels.

- *EMImputation* uses the Expectation Maximization and a multivariate normal model for replacing missing numeric values.

- *IBk* is a K-nearest neighbours regression learner. It can select an appropriate value of K based on cross-validation.

- *Kstar* uses an entropy-based distance function for instance-based regression, where the predicted class value of a test instance comes from the values of training instances similar to it.

- *LinearRegression* applies the Akaike Information Criterion for model selec-

tion in linear regression for prediction.

- *M5Rules* generates a decision list for regression problems using separate-and-conquer. In each iteration it builds a model tree using M5 algorithm originally proposed by R.J. Quinlan [21], and makes the "best" leaf into a rule, as described in [22].

- *REPTree* is a fast decision tree learner building a decision tree based on information gain and pruning it via reduced-error pruning with backfitting.

We use WEKA implementations of these base regression learners and refer to [23–27] for more information.

3.2. Meta regression techniques

It is well known that meta regression can improve the performance of the base regression learners. Each meta regression learner is built by applying one of the known meta regression techniques to a base regression learner ([27]). Our experiments investigated and compared the following five meta regression techniques in their ability to improve the performance of the base regression learners as ingredients of our method of treating the missing Ewing values.

- *AdditiveRegression* successively enhances the performance of a base regression learner. Each iteration fits a model to the residuals left by the previous regression learner. Final prediction is made by adding the predictions of all regression learners.

- *Bagging* is a regression scheme for bagging base regression model to reduce variance.

- *MultiScheme* is a regression scheme using cross validation on the training data or the performance on the training data to select a base regression model from several models according to the mean-squared error.

- *RandomSubSpace* constructs a decision tree based classifier that maintains highest accuracy on training data and improves on generalization accuracy as it grows in complexity. The classifier consists of multiple trees constructed systematically by pseudorandomly selecting subsets of components of the trees constructed in randomly chosen subspaces.

- *RegressionByDiscretization* employs any regression learner on a copy of the data that has the class attribute discretized. The predicted value is the expected value of the mean class value for each discretized interval based on the predicted probabilities for each interval and on conditional density estimation accomplished by building a univariate density estimator from the target values in the training data weighted by the class probabilities.

More information on these meta regression techniques and their WEKA implementations is given in [24, 27].

4. Diabetes screening research initiative database

We use a new and unique database from a diabetes screening research initiative (DiScRi) project [18–20, 28, 29]. DiScRi is a diabetes complications screening program in Australia where members of the general public participate in a comprehensive health review consisting of tests including electrocardiogram (ECG), the Ewing battery, retinal scans, peripheral nerve function and assessment of diverse biomarkers associated with risk and early detection of diabetes and cardio-

vascular disease. ECG data is crucial for medical applications, as illustrated, for example, by [30–33]. The DiScRi database is more than ten times larger than the data set used by Ewing in terms of the number of participants involved. Data on over 200 variables from over two thousand attendances have been collected in DiScRi, see [18, 29].

Since there are few atypical and severe patients in the DiScRi database, we deleted all instances with severe and atypical Ewing category and investigated two classifications for cardiac autonomic neuropathy progression originally defined by Ewing et al. [19, 20]. The first classification divides all patients into two categories allocating each patient either to the *normal* category, or to *definite* category. The second one divides all patients into three categories allocating each patient to one of the following categories: *normal*, *early*, *definite*. An alternative option was to merge all atypical and severe instances into other categories. This option was not used here, since such mergers are arbitrary and so they may lead to a certain reduction of the accuracy of the method. Note that since there are only small numbers of atypical and severe instances in the DiScRi database, this may result only in relatively minor changes.

5. Experiments and discussion

The aim of our experiments was to find a combination of the regression and meta regression techniques with the Ewing formula that achieves the best prediction of the Ewing category for a new instance I_N with a missing Ewing value.

There are five Ewing attributes, and so the missing feature F_M can take on one of the five values: VAHR, DBHR, HGBP, LSHR, LSBP. We considered all combinations of the regression and meta regression techniques with the Ewing formula

and used the DiScRi database to determine the accuracy each combination can achieve in predicting each of the missing variables VAHR, DBHR, HGBP, LSHR, LSBP. This means that for each combination we conducted five tests to determine the accuracy it achieves in predicting these five variables. Accordingly, for each particular combination, the charts with the outcomes of our experiments contain five bars labelled by the variables VAHR, DBHR, HGBP, LSHR, LSBP. These bars represent results corresponding to these Ewing variables.

We use the prediction accuracy of the CAN category as a measure to compare the outcomes. Our objective is not to learn how to impute the missing values with high precision with respect to appropriate metrics, as for example in [34]. Here we investigate the problem of predicting the CAN category for data with missing values. Precision of the imputation as one step of the whole procedure plays only an intermediate role, but the quality of the scheme has to be assessed looking at the final outcome. Therefore, the accuracy of the final classification of CAN is the most appropriate measure for evaluating the effectiveness of each method of handling missing values in this paper. In all previous articles using DiScRi our experiments have shown that for this database different measures of performance of classifiers correlate well, as it is often the case for well balanced data sets. This means that the algorithms with higher accuracy tend to produce better specificity and other metrics. This is why we included only the accuracies in the diagrams with outcomes in this paper.

To prevent overfitting of the regression models during tests, we applied tenfold cross validation. This is a standard and very well known procedure explained, for example, in [27]. Here we include a brief overview of how it works in our case. First, we selected all records with complete Ewing values and categories from

DiScRi database. Let us denote the set of all these records by D_{all}. The numbers of instances in each of the two or three categories of CAN in the set D_{all} are given in Table 3.

	2 classes	3 classes
Normal	461	461
Early	-	442
Definite	717	275

Table 3: Breakdown of the CAN categories in the dataset selected for experiments.

Applying tenfold cross validation means that we divided D_{all} into ten stratified folds, D_1, ..., D_{10}. We used these folds to conduct ten consecutive tests for each combination of a Ewing feature F_M and a regression learner L.

In the first test, we used the set $Train_1 = D_{all} \setminus D_1$ as the training set. The set $Train_1$ consists of all records from D_{all} that do not belong to D_1. The regression learner L was trained on $Train_1$ to impute the missing value F_M from all other Ewing values. The testing set $Test_1$ was obtained from D_1 by creating a copy of the set D_1 and turning all values of the feature F_M into missing values there. (Notice that this is different from deleting the whole attribute F_M from the set.) The regression learner L was used to impute the missing Ewing values in all records of $Test_1$. After that the Ewing formula was applied to derive the category of all instances in $Test_1$. A Python script was written by the second author to apply the Ewing rules automatically. The predicted categories were compared with the correct ones contained in D_1. The accuracy of the predicted categories was the

outcome of the first of the ten consecutive tests.

The other nine consecutive tests for the same L and F_M were organized in the same way for the sets D_2, ..., D_9 instead of D_1. The average accuracy obtained in all ten consecutive tests is the outcome of tenfold cross validation for L and F_M. It is included as a bar in the diagram with the outcomes, where it is labelled by the Ewing feature F_M.

The accuracies of the detection of CAN and the classification of CAN progression with three categories using base regression learners ConjunctiveRule, EMImputation, Kstar, LinearRegression, M5Rules and REPTree are presented in Figures 1 and 2. The best outcomes were obtained by M5Rules.

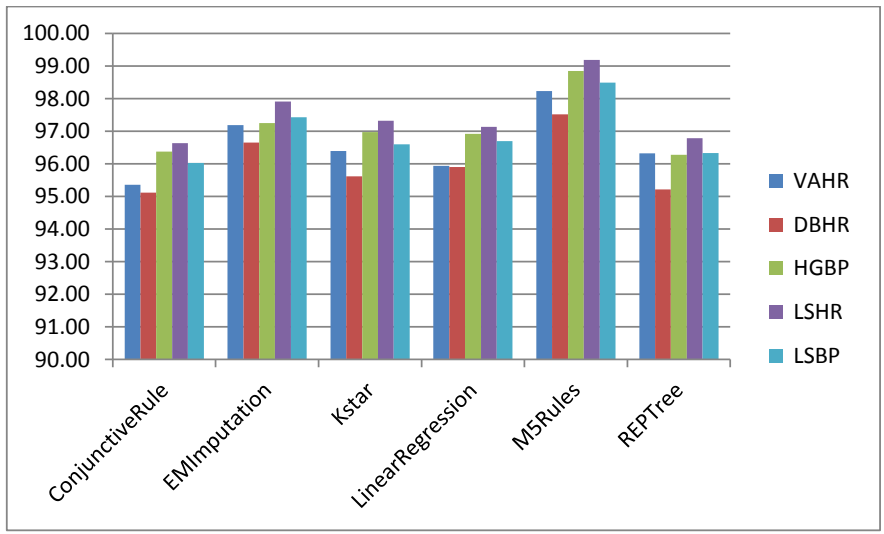

Figure 1: Accuracy of the detection of CAN with two categories for instances with a missing Ewing feature by using base regression learners to impute missing values and then applying the Ewing formula to determine the CAN category.

Next, we investigate the ability of the meta regression learners AdditiveRegression, Bagging, MultiScheme, RandomSubSpace, and RegressionByDiscretiza-

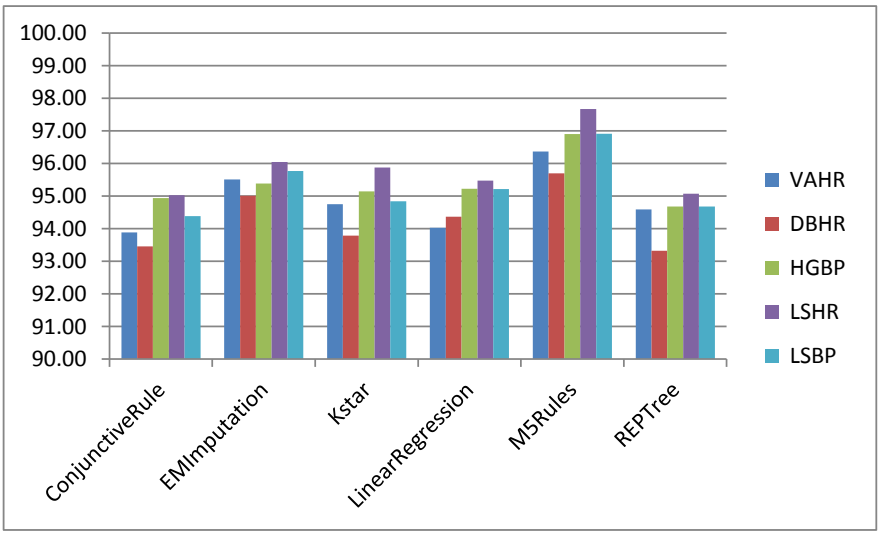

Figure 2: Accuracy of classification of CAN progression with three categories for instances with a missing Ewing feature by using base regression learners to impute missing values and then applying the Ewing formula to determine the CAN category.

tion to improve the performance of M5Rules and enhance the effectiveness of the detection of CAN and classification of CAN disease progression. Figures 3 and 4 contain the accuracies achieved by these meta regression techniques based on the best base regression method M5Rules.

Thus, our experiments have found the following best combination of meta regression techniques and the Ewing formula for classifying CAN instances with missing values. The best results have been produced by the additive regression meta learner based on M5Rules and combined with the Ewing formula. It has achieved the best accuracy of 99.78% and the average accuracy of 99.12% for two classes of CAN and, respectively, the best accuracy of 98.98% and the average accuracy of 98.31% for three classes of CAN.

To illustrate that our new method is much more effective, here we include

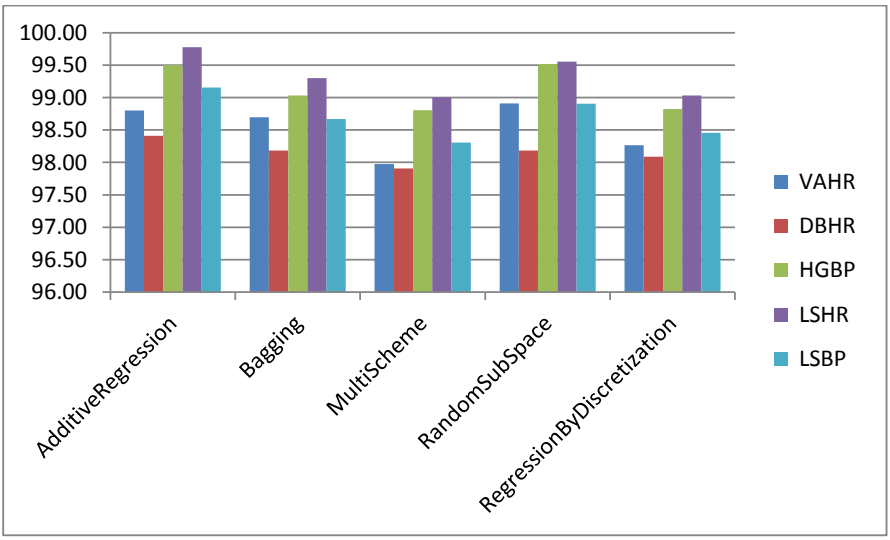

Figure 3: Accuracy of the detection of CAN with two categories for instances with a missing Ewing feature by using meta regression based on M5Rules to impute missing values and then applying the Ewing formula to determine the CAN category.

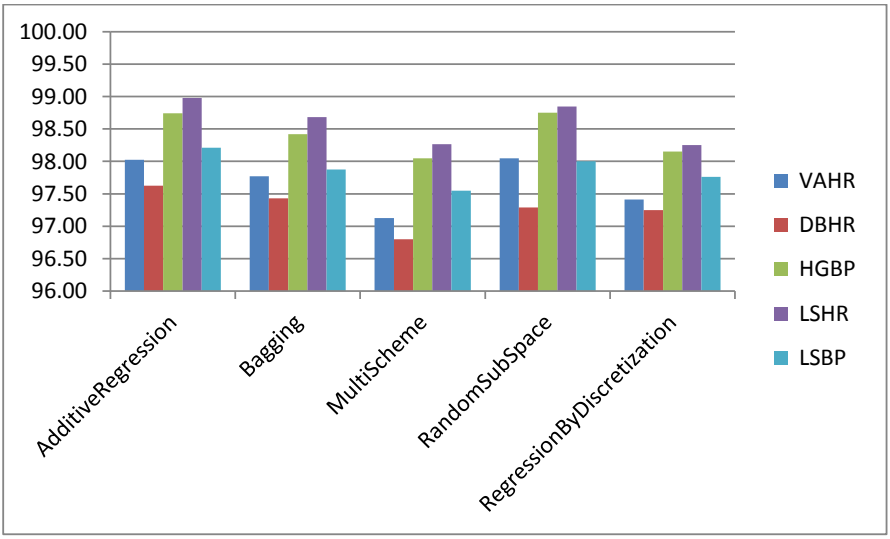

Figure 4: Accuracy of classification of CAN progression with three categories for instances with a missing Ewing feature by using meta regression based on M5Rules to impute missing values and then applying the Ewing formula to determine the CAN category.

more details on the classification schemes devised in the previous papers for the corresponding alternative sets of features considered there.

The article [11] dealt only with two classes of CAN, i.e., it handled only binary classifications. For the set of features considered in [11] the decision tree ensemble generated by Decorate based on RandomTree turned out the best. It achieved the accuracy of 94.23%. The paper [16] considered only three classes of CAN. The accuracy of the best meta classification scheme used in [16] is equal to 90.84%. Four classes of CAN and more advanced classifiers were considered in [10]. The accuracy obtained by the best classifier investigated in [10] is equal to 91.61%. Only two classes of CAN were considered in [11–14, 28]. The accuracies of the best classifiers designed in [28], [12], [13], [14] are equal to 80.66%, 94.61%, 94.84%, 97.74%, respectively. Thus, we see that our new method is more effective.

Only complete data were used in [15] and the problem of missing values was not addressed there. The paper [15] considered both two and three classes of CAN and developed multi-level classifiers that produced outcomes with the best accuracies approximately equal to the outcomes of the present article. However, the results of [15] cannot be applied to handle missing values, since all tests there used a large set of features.

Finally, let us note that the article [17] was devoted to a totally different problem of choosing an optimal order of the Ewing tests using the Optimal Decision Path Finder procedure and visual aids simplifying the selection of the next Ewing test during applications of this procedure in practice. The results obtained there for a completely different problem cannot be compared to the outcomes of our new method proposed in the present paper. Only decision trees were used in [17]

and the best accuracy achieved there is equal to 94.14%.

6. Conclusions

In this paper, we propose a new method for classifying CAN data with missing values of the Ewing features. Our experiments investigated various combinations of the base regression learners and advanced meta regression techniques with the use of the classical Ewing formula for determining the CAN category. The results of experiments show that the new method can achieve a significant improvement compared to the outcomes of all previous classification schemes.

The best results were obtained by our new method using the combination of the Ewing formula with the AdditiveRegression based on M5Rules. It has achieved the best accuracy of 99.78% for two classes of CAN, and the best accuracy of 98.98% for three classes of CAN. These outcomes are better than all previous results obtained in the literature. Another advantage of our method is that it does not require practitioners to perform more tests collecting alternative features.

7. Acknowledgements

This research was supported by a Deakin-Ballarat collaboration grant. The authors are grateful to two reviewers for comments and corrections, which have helped to improve this article.

References

[1] D. Cysarz, P. V. Leeuwen, F. Edelhuser, N. Montano, A. Porta, Binary symbolic dynamics classifies heart rate variability patterns linked to autonomic modulations, Computers in Biology and Medicine 42 (3) (2012) 313 – 318.

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

[3] I. Nejadgholi, M. H. Moradi, F. Abdolali, Using phase space reconstruction for patient independent heartbeat classification in comparison with some benchmark methods, Computers in Biology and Medicine 41 (6) (2011) 411 – 419.

[4] A. Van, V. C. Gay, P. J. Kennedy, E. Barin, P. Leijdekkers, Understanding risk factors in cardiac rehabilitation patients with random forests and decision trees, 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, 11–22, 2011.

[5] S.-N. Yu, M.-Y. Lee, Bispectral analysis and genetic algorithm for congestive heart failure recognition based on heart rate variability, Computers in Biology and Medicine 42 (8) (2012) 816 – 825.

[6] D. Zhang, J. He, J. Yao, Y. Wu, M. Du, Noninvasive detection of mechanical prosthetic heart valve disorder, Computers in Biology and Medicine 42 (8) (2012) 785 – 792.

[7] R. Pop-Busui, Cardiac autonomic neuropathy in diabetes: a clinical perspective, Diabetes Care 33 (2010) 434–441.

[8] A. I. Vinik, D. Ziegler, Diabetic cardiovascular autonomic neuropathy, Circulation 115 (2007) 387–397.

[9] D. J. Ewing, B. F. Clarke, Diagnosis and management of diabetic autonomic neuropathy, British Medical Journal 285 (1982) 916–918.

[10] H. F. Jelinek, A. Kelarev, A. Stranieri, J. L. Yearwood, Rule-based classifiers and meta classifiers for identification of cardiac autonomic neuropathy progression, Int. J. Information Science and Computer Mathematics 5 (2012) 49–53.

[11] A. V. Kelarev, J. Abawajy, A. Stranieri, H. F. Jelinek, Empirical investigation of decision tree ensembles for monitoring cardiac complications of diabetes, Int. J. Data Warehousing and Mining 9 (2013).

[12] A. V. Kelarev, R. Dazeley, A. Stranieri, J. L. Yearwood, H. F. Jelinek, Detection of CAN by ensemble classifiers based on ripple down rules, in: Knowledge Management and Acquisition for Intelligent Systems, PKAW2012, vol. 7457 of *Lecture Notes in Computer Science*, 147–159, 2012.

[13] A. V. Kelarev, A. Stranieri, J. L. Yearwood, H. F. Jelinek, Empirical study of decision trees and ensemble classifiers for monitoring of diabetes patients in pervasive healthcare, in: Proceedings of 15th International Conference on Network-Based Information Systems, NBiS-2012, 441–446, 2012.

[14] J. Abawajy, A. V. Kelarev, A. Stranieri, H. F. Jelinek, Empirical investigation of multi-tier ensembles for the detection of cardiac autonomic neuropathy

using subsets of the Ewing features, in: Proceedings of the Workshop on New Trends of Computational Intelligence in Healthcare Applications, CI-Health 2012, vol. 944 of *CEUR Workshop Proceedings*, 1–11, 2012.

[15] A. V. Kelarev, A. Stranieri, J. L. Yearwood, J. Abawajy, H. F. Jelinek, Improving classifications for cardiac autonomic meuropathy using multi-level ensemble classifiers and feature selection based on random forest, in: Data Mining and Analytics 2009, Proc. 10th Australasian Data Mining Conference, AusDM 2012, vol. 134 of *CRPIT*, ACS, Sydney, Australia, 93–101, 2012.

[16] A. V. Kelarev, A. Stranieri, J. L. Yearwood, H. F. Jelinek, A comparison of machine learning algorithms for multilabel classification of CAN, Advances in Computer Science and Engineering 9 (2012) 1–4.

[17] A. Stranieri, J. Abawajy, A. Kelarev, S. Huda, M. Chowdhury, H. F. Jelinek, An approach for Ewing test selection to support the clinical assessment of cardiac autonomic neuropathy, Artificial Intelligence in Medicine (2013) DOI:10.1016/j.artmed.2013.04.007.

[18] H. F. Jelinek, C. Wilding, P. Tinley, An innovative multi-disciplinary diabetes complications screening programme in a rural community: A description and preliminary results of the screening, Australian Journal of Primary Health 12 (2006) 14–20.

[19] D. J. Ewing, J. W. Campbell, B. F. Clarke, The natural history of diabetic autonomic neuropathy, Q. J. Med. 49 (1980) 95–100.

[20] D. J. Ewing, C. N. Martyn, R. J. Young, B. F. Clarke, The value of cardiovascular autonomic function tests: 10 years experience in diabetes, Diabetes Care 8 (1985) 491–498.

[21] R. J. Quinlan, Learning with Continuous Classes, in: 5th Australian Joint Conference on Artificial Intelligence, AI92, 343–348, 1992.

[22] G. Holmes, M. Hall, E. Frank, Generating rule sets from model trees, in: 12th Australian Joint Conference on Artificial Intelligence, AI99, 1–12, 1999.

[23] R. E. Frye, M.-H. Wu, Multichannel least-squares linear regression provides a fast, accurate, unbiased and robust estimation of Granger causality for neurophysiological data, Computers in Biology and Medicine 41 (12) (2011) 1118 – 1131.

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

[25] Y. Ren, B. Wu, Y. Pan, F. Lv, X. Kong, X. Luo, Y. Li, Q. Yang, Characterization of the binding profile of peptide to transporter associated with antigen processing (TAP) using Gaussian process regression, Computers in Biology and Medicine 41 (9) (2011) 865 – 870.

[26] W. Shoombuatong, S. Hongjaisee, F. Barin, J. Chaijaruwanich, T. Samleerat, HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees, Computers in Biology and Medicine 42 (9) (2012) 885 – 889.

[27] I. H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier/Morgan Kaufman, Amsterdam, 2011.

[28] S. Huda, H. F. Jelinek, B. Ray, A. Stranieri, J. Yearwood, 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, 297–302, 2010.

[29] A. H. Khandoker, H. F. Jelinek, M. Palaniswami, Identifying diabetic patients with cardiac autonomic neuropathy by heart rate complexity analysis, BioMedical Engineering OnLine 8 (2009) 1–12.

[30] E. M. A. Anas, S. Y. Lee, M. K. Hasan, Exploiting correlation of ECG with certain EMD functions for discrimination of ventricular fibrillation, Computers in Biology and Medicine 41 (2) (2011) 110 – 114.

[31] H. K. Chatterjee, R. Gupta, M. Mitra, A statistical approach for determination of time plane features from digitized ECG, Computers in Biology and Medicine 41 (5) (2011) 278 – 284.

[32] E. M. Dantas, M. L. Sant'Anna, R. V. Andreo, C. P. Gonalves, E. A. Morra, M. P. Baldo, S. L. Rodrigues, J. G. Mill, Spectral analysis of heart rate variability with the autoregressive method: What model order to choose?, Computers in Biology and Medicine 42 (2) (2012) 164 – 170.

[33] S. Shakibfar, C. Graff, L. H. Ehlers, E. Toft, J. K. Kanters, J. J. Struijk, Assessing common classification methods for the identification of abnor-

mal repolarization using indicators of T-wave morphology and QT interval, Computers in Biology and Medicine 42 (4) (2012) 485 – 491.

[34] F. Dorri, P. Azmi, F. Dorri, Missing value imputation in DNA microarrays based on conjugate gradient method, Computers in Biology and Medicine 42 (2) (2012) 222 – 227.