Exploring Contributing Features of Pre-Graft Orthodontic Treatment of Cleft Lip and Palate Patients Using Random Forests

Zaturrawiah Ali Omar[1#], Su Na Chin[2], Albira Sentian[3], Norhafiza Hamzah[1], Fouziah Yassin[4]

1 Mathematics with Computer Graphics Programme, Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, MALAYSIA.
2 Mathematics with Economics Programme, Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, MALAYSIA.
3 Orthodontic Specialist Clinic, Queen Elizabeth Hospital, 88586 Kota Kinabalu, Sabah, MALAYSIA.
4 Physic with Electronic Programme, Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, MALAYSIA.
Corresponding author. E-Mail zatur@ums.edu.my; Tel: +6088-320000 Ext. 5750; Fax: +6088-435324.

ABSTRACT When a study faced with a limited data, it can be quite a challenge to conduct analyses that can be statistically significant. Given such circumstances, typically, researchers options would either, gathered more data (by pooling or adding artificial data), or used a better analysis algorithm. This study interest was in the latter approach looking at the implementation of a highly suggested machine learning algorithm, that is, the random forests. The primary objective was to explore features that contribute to the success of pre-orthodontic treatments for cleft lip and palate patients. Specific centres have adopted the pre-orthodontic treatments as one of the treatment protocols for the cleft lip and palate patients, before their secondary alveolar bone graft surgery. It was in the intention of the orthodontic department to achieve better patients' management with the knowledge of the contributing features, as handling these patients can be quite challenging. With only 18 datasets, the random forest out of bag error estimation (or misclassification error) was 27.78%. The error was further reduced to 11.11% when backward elimination was conducted starting with the lowest ranked variable. These leave the top four variables which were, the affected cleft palate (*acp*) either at the soft palate or hard palate or both, the ethnicity (*ethnicity*), referral age (*ageR*) and lastly, age at treatment (*ageP*). To eliminate the chances of variable selection biases, a conditional forest (*cforest*) function was conducted and the results suggested that only the affected cleft palate variable was important. Details of these explored top features are discussed further in this paper.

KEYWORDS: Random Forest; Classification; Small Dataset, Pre-graft Orthodontic Treatment; Cleft Lip and Palate

Full Article - *Computer and Information Sciences*
Received 19 January 2018 **Revised** 14 March 2018 **Accepted** 16 March 2018 **Online** 28 March 2018

INTRODUCTION

In this study, we were presented with a case of Cleft Lip and Palate (CLP) patients that went through pre-orthodontic treatments before their secondary alveolar bone graft surgery. The pre-orthodontic treatments have been adopted as one of the treatment protocols for the CLP patients, and there are studies done (Rocha *et al.*, 2012, Luque-Martín *et al.*, 2014) that show the significance of the pre-orthodontic treatments in ensuring the success of the alveolar bone graft surgery. The interest of our study was to investigate if there exist prominent features of the CLP patients that contribute to the success of the pre-orthodontic treatments. We consider a successful treatment as one that does not last more than a year based on Machos (1996), Nahai *et al.* (2005) and Luque-Martín *et al.* (2014), and this was being determined from the age a patient started the procedure till the age they went through the surgery. The current trend from the case study showed on average, a duration of more than two years of pre-orthodontic treatments per patient. Although many factors can delay the treatment length, this study mainly focused on the conditions of the CLP patients itself as the factors. With a better understanding of the contributing features could lead to better CLP patients' management and therefore improve the quality of services to these patients as well as minimises other risks that could disrupt these patients' dental health condition and their cooperation (Nahai *et al*., 2005).

In this study, however, the data collected were somewhat limited due to time and cost constraint. Therefore, the approach of this study was more of an exploratory rather than confirmatory. Somehow, this approach had led us to work done by Rodriguez-Galiano *et al.* (2012) and Jones & Linder (2016) where they were able to demonstrate random forests (Breiman, 2001) can be used as a tool for exploratory data analysis with the aids from variables ranking and partial dependence plots. Random forests is considered to be a "black box" model where its statistical significance information is hidden inside the model structure causing a misperception that it is only good for prediction but not for sound theoretical work or substantive insight (Jones & Linder, 2016). Since its introduction, however, we noticed that random forests had become one of the panacea choices for a data analyst and a study done by Delgado *et al.* (2014) (where almost every machine learning family algorithms were tested) has ranked random forests as the first algorithm that is most likely to be the best classifiers.

The advantages of random forests that requires no distribution assumptions; robust to outliers and noise data; able to handle large p and small n cases; able to handle different type of variables, seems to be fitting to be used in this study. Random forests also comes with its build in error estimation (known as out-of-bag) that is not just used to monitor error, but strength and correlation, as well as used to measure variable importance (Breiman, 2001). Unfortunately, the random forests also has its downfall. The construction of its tree predictor is based on the classification and regression tree (CART) that is known to be biased towards selection of variables with many categories, numeric variables and variables with much missing value (Strobl *et al*., 2009). Therefore, (Hothorn *et al*., 2006) has come with a conditional tree which produces a conditional forest (*cforest*) that was designed to overcome the variable selection bias inherited in random forests.

BACKGROUND THEORY

The Random Forests and CForest

The random forests is a machine learning algorithm that forms an assemble of decision trees that take the majority vote for classification. It is also applicable for regression by averaging the decision of all the trees in the forest. Random forests, however, change the way a classification or regression trees are constructed by adding a random element in the selection of variables for node splitting.

Adopting the definition from (Breiman, 2001), random forest for classification can be formally defined as a collection of a tree structured classifiers given by:

$$\{h(x,\Theta_k, k=1,....)\}, \tag{1}$$

where Θ_k are independent identically distributed random vectors, and each tree casts a unit vote for the most popular class at input x. The construction of random forest can also be broken down to as follows (Kuolis, 2003):

- At step k, $a(\Theta_k)$ is generated
- The Θ_k are i.i.d

- A tree predictor $\{h(x,\Theta_k)\}$ is then constructed from k=1 until a very large number (usually k=100) with no pruning
- After many trees have been generated, each will vote for the most popular class
- The random forest then classifies x, by taking the majority class from all the tree predictors in the forest $\{h(x,\Theta_k, k=1,....)\}$.

The Θ_k construction is based on bootstrapped sample (with replacement) where one-third of the sample is for the out-of-bag (OOB) set which will be used to estimate the prediction error (Breiman, 2001; Thanh *et.al*, 2014). Therefore, a tree predictor $\{h(x,\Theta_k)\}$ is grown with only two-thirds of Θ_k. During the growing phase, let M be the number of variables to be selected at random at each node and M << number of variables. Best split is done based on Gini split criterion through CART method without pruning (Kuolis, 2003; Strobl *et al.*, 2007).

The difference of random forest with *cforest* is the generated tree predictor $\{h(x,\Theta_k)\}$ is based on a conditional inference framework (Hothorn *et al.*, 2006) and Θ_k can be drawn without replacement, producing unbiased tree prediction, hence unbiased forest (Strobl *et al.*, 2007).

The Variable Ranking and Selection

One of the results from random forests is the variable importance that is measured using permutation importance and Gini importance. However, the Gini importance is based on Gini split that can be biased therefore the permutation importance is more reliable (Strobl *et al.*, 2009). For selecting the variables for further exploration, Strobl *et al.* (2009) suggested a conservative approach where variables that are negatives, zero or values that lie in the same range as the negative values can be excluded.

The Partial Dependence Plot

The partial dependence plot (Freidman, 2001) is a graphical depiction of how each predictor variable would affect the model's prediction. The approach is conducted by averaging all the other predictor variables except for a chosen one, and the effect of the chosen variable over a selected outcome is plotted. For a classification problem, the y-axis will show the variable's class probability, and any variable with a probability distribution of less than zero is weakly contributing.

METHODOLOGY

Sample Collection

An initial sample of 23 CLP patients' dental records was from the Orthodontic Specialist Clinic from Queen Elizabeth II Hospital, Kota Kinabalu Sabah. This data is a collection of records from January 2000 until December 2015. There are nine successes and 14 unsuccessful cases. Smaller numbers of records were partly due to the predetermined inclusion criteria that were non-syndromic, non-systemic, only complete records and only patients that were first referred at the age of 14 or below. The unsuccessful cases were further reduced to nine by eliminating the older referral age patients. The reduction was to balance the training set distribution as it helps to improve the learning algorithm performances (Weiss & Provost, 2003; Niu *et al.*, 2015), leaving the total records used to 18. Altogether, there were seven predictors and one classifier variables. The predictors comprised of two demographic variables and five CLP patients' condition variables. The details of these variables are in Table 1.

Analysis Tool and Parameters Setting

The analysis in this study was done using R version 3.3.3. The *randomForest* (Liaw & Wiener, 2002), and *party* (Hothorn *et al.*, 2006, Strobl *et al.*, 2007, Strobl *et al.*, 2008) packages were imported to enable the required functions to produce the random forests and the *cforest*. Since we want the results to be reproducible, a seed number was set to 71. The number of the tree (*ntree*) showed a stable OOB after *ntree* > 500. Therefore, *ntree* was set to 1000, and the m*try* was set to default. By default, the number of *mtry* will be square root the number of variables. The *cforest* control parameters that relate to the random forest were set the same as above with the *minsplit* and *minbucket* were set to 2. This setting was to ensure a possible split in the conditional tree as the number of data were less than the default *minsplit* value (default is 20 and seven respectively). In the variable selection process, Diaz-Uriarte (2007) has come up with a variable selection package *varSelRF*, and it has been used as a comparison in this study. The *varSelRF* random forests related parameters were set the same as that in random forests.

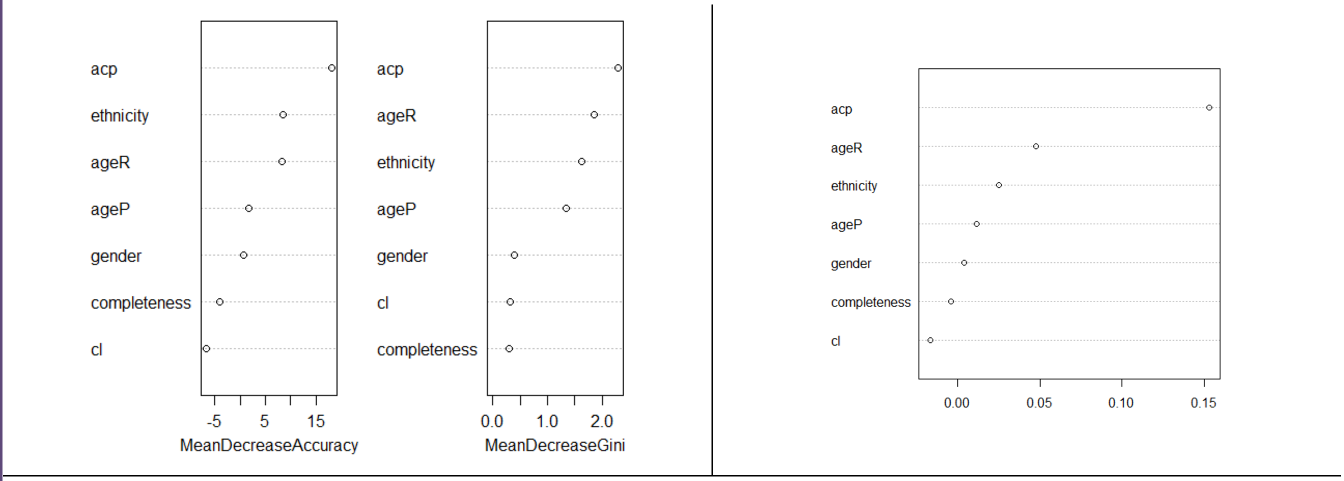
Table 1. Details of variables used for classification.

Name of variable	Description	Data Type	Categorical Value	Coded Value
Predictors				
ethnicity	The race of the patient. Since this is a sample from Sabah, the main ethnic in Sabah that is, the Kadazan, Dusun and Murut (KDM) were also included.	Categorical	Other Malay Chinese KDM	0 1 2 3
gender	The gender of the patient	Categorical	Male Female	1 0
acp	The affected cleft palate area that can reach the soft palate, or hard palate or both	Categorical	Soft & Hard Soft / Hard	1 0
cl	The cleft lateral referring to a unilateral cleft or bilateral cleft	Categorical	Unilateral Bilateral	1 0
completeness	If a cleft lip reaches to the nostril, then the cleft lateral is complete	Categorical	Complete Incomplete	1 0
ageP	An age when the patients went through the graft surgery	Continuous	Age in year	7
ageR	An age when the patients first being reviewed	Continuous	Age in year	11
Classifier				
Success	Classes that indicate the success or failure of the pre-graft orthodontic treatment	Categorical	Success Failure	1 0

RESULT AND DISCUSSION

Figure 1 shows the variable importance ranks using random forests in (a) and *cforest* in (b). At a glance, all three results were synonymous in ranking *acp* as the most important variable. The primary comparison was only between (a). Random forest – mean decrease accuracy ranking (which is based on permutation importance) and (b). *cforest* ranking. The results showed random forest had ranked *ethnicity* higher than *cforest* and all other rankings were the same. However, we can't help but notice how similar were the ranking of (a). Random forest – mean decrease Gini ranking with (b).

cforest, with only a variation in the ranking of *cl* and *completeness*. With the presence of all these variables, the random forest OOB error rate was 28.78%.



(a). (b).

Figure 1. Variables importance rank by (a). Random Forests and (b). *cforest*

The contribution each of the variables towards the Success class was further scrutinised based on the partial dependency plots in Figure 2. The plots suggested that *gender*, *cl* and *completeness* were the lowest contributors (probability ≤ 0.010) hence explaining their bottom three positions in the variables importance ranking in random forests as well as *cforest*. The plots also explained the interchangeable rank position between *ethnicity* and *ageP* (the probability ranges were the same), and why *ageR* was always ranked higher than *ageP* (probability of *ageR* > probability of *ageP*). When the variable selection was conducted using *varselRF*, the function selected *acp*, *ageR* and *ethnicity* as the most contributing. These three variables were at the top three in the variables importance ranking. When random forests was reconstructed again based on these three variables, the OOB error rate was reduced to 16.67%. Looking at the *cforest* results and applying the suggestion by Strobl, *et al.* (2009), then only *acp* was worth for further exploration. The OOB error rate for *acp* alone was also at 16.67%. We then conducted manual backward elimination starting with the lowest ranking variable and noticed that having the top three variables and *ageP* would further reduce the OOB to 11.11%.

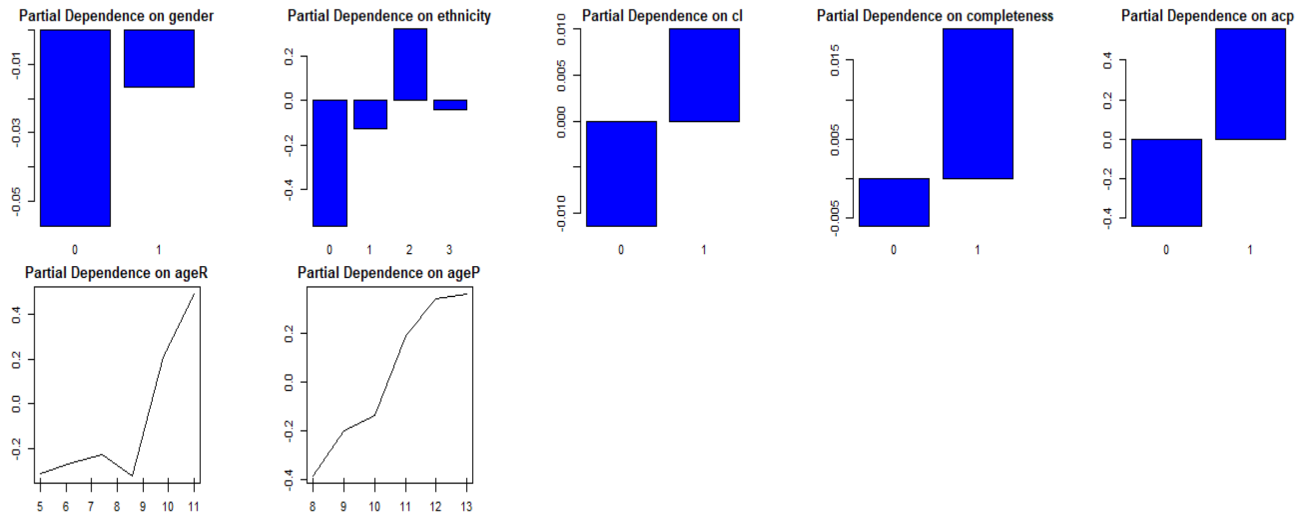


Figure 2. Partial Dependence Plot for Success class

Based on these findings, we had divided the variables into three groups labelled with high, medium and low importance as seen in Table 2. The variables of high importance are ones that would be the most important features, and in this case study, *acp* was the most important. The medium level would suggest variables that can be of importance. We considered *ageP* to be at the medium level since the inclusion of this variable would further reduce the misclassification error rate. While the least importance as those that can be eliminated.

Table 2. Variable Importance Level for pre-orthodontic Treatment.

Level	Name of variable
High	*acp*
Medium	*ageR*
	ageP
	ethnicity
Low	*cl*
	completeness
	gender

CONCLUSION

The implementation in of random forest and *cforest* for classification in this study had identified *acp* as the highly significant variable in determining the success of the pre-graft orthodontic treatments for a cleft lip and palate patients. Based on the partial dependency of *acp* (see Figure 2), patients, whose affected cleft palate reached on both the soft and hard part, were more likely to contribute to successful treatment. If *ethnicity* was to be considered as an essential feature, we can see that the Chinese would have higher chances of successful treatment. Concerning the age of the patients being first reviewed and went through the surgery, there was an increasing tendency for success after the age of 10. With these insights, the study was still far from being conclusive. However, it does give a direction to further the study by exploring more at the breakdown of the *acp* features and the treatments that were required depending on the affected cleft palate as mentioned by (Levy-Bercowski *et al.*, 2011).

REFERENCES

[1] Breiman, L. (2001). Random Forests. *Machine Learning*, **45**(1), 5-32.

[2] Delgado, M. F., Cernadas, E., Barro, S. & Amorim, D. (2014). Do We Need Hundreds of Classifiers to Solve Real World Classification Problem? *Journal of Machine Learning Research*, **15**(1), 3133-3181.

[3] Diaz-Uriarte, R. (2007). GeneSrF and varSelRF: a web-based tool and R package. *BMC Bioinformatics 2007*, **8,** 328

[4] Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**(5), 1189–1232

[5] Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A. & Van Der Laan, M. (2006). Survival Ensembles. *Biostatistics*, **7**(3), 355--373.

[6] Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.

[7] Jones, Z. & Linder, F. (2016). Exploratory Data Analysis using Random Forests. *Journal of Open Source Software*. **1**(6), 92.

[8] Kuolis, T. (2003, April). *Random Forests: Presentation Summary*. (http://sas.uwaterloo.ca/-hachipma/stat946/koulis.pdf) Accesssed on 29 August 2017.

[9] Levy-Bercowski, D., DeLeon, E., Stockstill, J. W. & Yu, J. C. (2011). Orthognathic Cleft—Surgical/Orthodontic Treatment. *Seminars in Orthodontics*, **17**(3), 197-206.

[10] Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, **2**, 18-22.

[11] Luque-Martín, E., Tobella-Camps, M.-L. & Rivera-Baró, A. (2014). Alveolar graft in the cleft lip and palate patient: Review of 104 cases. *Medicina Oral Patologia Oral Y Cirugia Bucal,* **19**(5), e531-e537

[12] Machos, C. C. (1996). Orthodontic treatment for the cleft palate patient. *Seminars in Orthodontics*, **2**(3), 197-204.

[13] Nahai, F. R., Williams, J. K., Burstein, F. D., Martin, J. & Thomas, J. (2005). The Management of Cleft Lip and Palate: Pathways for Treatment and Longitudinal Assessment. *Seminars in Plastic Surgery,* **19**(4), 275-285.

[14] Niu, S., Lan, Y., Guo, J., Wan, S. & Cheng, X. (2015). Which noise affects algorithm robustness for learning to rank. *Information Retrieval,* **18**(3), 215-245.

[15] Rocha, R., Ritter, D. E., Locks, A., Paula, L. K. & Santana, R. M. (2012). Ideal treatment protocol for cleft lip and palate patient from mixed to permanent dentition. *American Journal of Orthodontics and Dentofacial Orthopedics,* **141**(4), S140-S148

[16] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sánchez, J. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *Isprs Journal of Photogrammetry and Remote Sensing,* **67**, 93-104.

[17] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, **9**(1), 307.

[18] Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*. **8**, 25.

[19] Strobl, C., Malley, J. & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychol Methods*, **14**(4), 323-348.

[20] Weiss, G. M. & Provost, F. (2003). Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research*, **19**, 315-354.