Atlantic Electronic
Journal of Mathematics
Volume **5**, Number **1**, Winter **2012**

http://aejm.ca
http://aejm.ca/rema
pp. **22–36**

DIFFERENTIATING BETWEEN SJOGREN'S SYNDROME AND DRY EYE DISEASE: AN ANALYSIS USING RANDOM FORESTS

Jessy A. Donelle*, Sunny X. Wang* and Barbara Caffery**

*Department of Mathematics, Statistics and Computer Science
St. Francis Xavier University
Antigonish, N.S. Canada B2G 2W5

**Yorkville Eye Clinic
33 Avenue Rd
Toronto, ON. Canada M5R 2G3

Abstract. To determine which non-invasive clinical tests can most easily differentiate primary Sjogrens syndrome (pSS) dry eye from non-autoimmune aqueous deficient dry eye (DE). The records of all patients seen at the University Health Network Sjogrens Syndrome Clinic from October 1992 to July 2006 were reviewed and documented. Patients were diagnosed with pSS by the AECC criteria of 2002. DE controls were non-SS patients with symptoms of dry eye and Schirmer scores less than 10 mm in 5 minutes in at least one eye. The non-parametric statistical technique, Random Forests (RF), was applied to the data set and these results were compared to the previous research results obtained by a single classification tree [Caffery et al., 2010]. Rose bengal staining of the conjunctiva and severity of the symptoms of dry eye and dry mouth were the most important non-invasive variables in differentiating pSS from DE. Random Forest analysis confirms the previous analysis of this data using single classification trees. The advantage of RF was superior accuracy when classifying data or estimating values for missing data.

1. **Introduction.** In recent years data mining techniques have been applied to a wide scope of real world problems and have had huge impacts in Statistics, Computer Science and other disciplines. Data mining can be viewed as a process in which various models, summaries, derived values, and valuable information is discovered from collections of data, and can be broken down into two main streams: directed or supervised learning and undirected or unsupervised learning [[Kantardzic, 2003], [Berry and Linoff, 2000]]. This paper employs a classification technique, Random Forests (RF) to identify which medical tests are the most important in differentiating whether a patient has primary Sjogrens's Syndrome (pSS) or non-autoimmune aqueous deficient dry eye disease (DE).

In working toward a diagnosis, clinicians prefer to make judicious use of time and costs. In the case of Sjogren's syndrome, determining which non-invasive tests are most likely to identify the disease is important as only those patients would be required to have further invasive tests which necessarily have risks and higher costs. The previous research paper, "Rose Bengal Staining of the Temporal Conjunctiva Differentiates Sjogren's Syndrome from Keratoconjunctivitis Sicca", outlined a similar goal of finding non-invasive medical tests

2000 *Mathematics Subject Classification.* Primary: 58F15, 58F17; Secondary: 53C35.

Key words and phrases. Data Mining, Random Forest, Dry Eye Disease, Sjogren's Syndrome.

The authors would like to acknowledge the funding support from NSERC and UCR (University Council for Research, St. Francis Xavier University).

but using a different data mining technique, single classification trees [Caffery et al., 2010]. This paper focuses on determining the important non-invasive medical tests but with high prediction accuracy, i.e. reduced misclassification error.

The paper is constructed as below. Section 2 gives a detailed description of the data set. The data mining techniques, classification tree and random forest, are presented in Section 3. The random forest algorithm was applied to the data set using the R software to perform the desired analysis for the given problem. Basically a random forest is a large group of unpruned classification trees created with bootstrap samples of the data set along with elements of randomness (discussed in further detail in Section 3.2). Section 4 lists all the results from RF analysis. A careful comparison between the results from RF and those from single classification trees [Caffery et al., 2010] is also presented in this section. Finally, the conclusions and discussions are given in Section 5.

2. **Background.** This section is divided into two parts. Section 2.1 gives a description of the diseases presented within the data. This corresponds to the classes of the response variable, "NAME". Section 2.2, presents a detailed description of variables contained in the data set.

2.1. **Disease Background.** There are two diseases within the data: aqueous deficient dry eye (DE) and primary Sjogren's syndrome (pSS). pSS is a systemic autoimmune disease with the hallmark presentation of dry eye and dry mouth [Fox, 1996]. The secretory glands of the eye and mouth become recognized as foreign by the immune system and are invaded by lymphocytes. Many other systems of the body may be affected by the inflammation of SS including lungs, skin and kidneys. Aqueous deficient dry eye occurs in patients without SS who have reduced tear production and symptoms of irritation and dry eyes [Lemp et al., 2007].

The diagnosis of Sjogren's syndrome is based on the 2002 American European Consensus criteria [Vitali et al., 2002]. The criteria includes (1) symptoms of dry eye, (2) symptoms of dry mouth, (3) signs of dry eye that include vital staining of the cells of the ocular surface and/or reduced tear flow as measured by Schirmer strips, (4) signs of dry mouth that include a low volume of saliva, (5) serum antibodies to nuclear proteins Ro and/or La and (6) the presence of foci of lymphocytes in the minor salivary gland biopsy. Four of the six criteria must be present for the diagnosis and at least one of those four must be either a positive blood serum or biopsy result. The diagnosis of aqueous deficient dry eye would include dry eye symptoms, ocular surface staining and low tear flow test.

2.2. **Data Set.** The data set used throughout the paper came from the University Health Network Sjogren's Syndrome clinic and included the records of patients who went to the clinic between October 1992 to July 2006 [Caffery et al., 2010]. The entire data set consisted of 378 patients and originally contained 101 variables which included the results of patients various invasive and non-invasive medical tests and their demographics.

Due to the omission of the qualitative results for the blood tests "ATA", "IgM", "IgG", "IgA" and "ANTI-MIC" listed in Table 3 from the original paper [Caffery et al., 2010], and the dental variables "missing teeth", "filled teeth", "cervical cavities", "D score", "DMF=dentate" and "missing filled" which contained a very large number of missing values, the final number of predictor variables actually used in the analysis was 89. Table 1 lists all the variables.

Among the 90 variables, "NAME" specifying which disease a patient was diagnosed with, served as the categorical response (dependent) variable, Y, and the remaining 89 became the predictor (independent) variables, $x_i,\ i = 1, \ldots, 89$. In order to do a fair comparison

TABLE 1. List of Variables

Classification of Variables	**Non-Invasive Variables**	**Invasive Variables**
Demographics	age, sex	
AECC Criteria	dry eye symptoms, dry mouth symptoms, dry eye signs (positive Schirmer or rose bengal), salivary flow positive	Biopsy, serum antibodies for Ro and/or La
Variables Associated with the AECC Criteria	severity of dry eye symptoms (0-10), how long eyes have been dry, severity of dry mouth symptoms (0-10), how long mouth has been dry, rose bengal staining score (RB) of 4/9 or greater, RB value in worst eye 0-9, Schirmer failed, i.e., 5 or less, Schirmer value worst eye, unstimulated salivary flow, salivary flow score	biopsy focus score 0-4, Chisholm Mason biopsy score 3-4 is SS, i.e., at least 1 focus in 4 mm, Ro present, La present
Systemic Autoimmune Diagnoses	mixed connective tissue (CT) disease present, CREST (calcinosis, Raynaud's, esophageal, sclerodactyly, telangiectasia) present, RA diagnosis, SLE diagnosis, PBC diagnosis	
Other Signs of Autoimmune Disease	Parotid swell, myalgia, arthralgia, fibromyalgia, lymphoma, X-ray positive	
Other Systemic Diseases	diabetes, hypothyroid	
Blood Work (quantitative)		IgG, IgM, IgA, M spike, WBC, ANA, RF, ATA, Anti Mic, TSH, AMA, SMA
Other Systemic Symptoms	dysphagia, dyspepsia, vaginal dryness, dysparunia, Raynaud's, dry skin, pruritis, rash, alopecia, photosensitive skin	
Medications by Category	diuretics, depression, anticholinergics, anti-inflammatories	
Other Eye Signs	meibomian gland dysfunction, superior limbic keratoconjunctivitis (SLK)	
Rose Bengal Stain	Worst eye (WE) RB temporal stain, WE corneal RB stain, WE nasal RB stain, each eye total of 3 areas	
Fluorescein Stain	WE temporal cornea , WE nasal cornea, WE superior cornea, WE inferior cornea, WE central cornea, each cornea by 5 quadrants, corneal stain of any kind	
Dental Information	candidiasis	

with the previous results published in the paper [Caffery et al., 2010], the analysis focused mainly on three subsets of the final 89 predictor variables: (1) all the invasive and non-invasive variables (total 89), (2) all the non-invasive excluding salivary flow (total 67) and (3) all non-invasive variables including salivary flow (total 70).

3. **Methodology.** Before introducing the Random Forest technique, a brief description of single classification tree method is presented in Section 3.1. Then, Section 3.2 will go into detail of how the Random Forest technique works and the theory behind it.

3.1. Classification Tree Analysis.

3.1.1. *Classification and Regression Trees (CART).* The CART method (also called decision trees) introduced by Breiman, Friedman, Olshen and Stone in the mid-1980s is defined to be a non-parametric, exploratory data analysis method which implements binary recursive partitioning as part of its algorithm [Sutton, 2005]. The type of decision tree that we focused on was the classification tree as the response variable Y is categorical. According to the values of the predictor variables, the observations are either sent to the left or the right child node [Yohannes and Hoodinott, 1999]. The common splitting criterion is called the Gini Index, which measures the purity of each node and the following equation (1) is the definition of the Gini Index.

$$1 - \sum_{j=1}^{J} p_j^2, \tag{1}$$

where p_j's are the class proportions and $j = 1, \ldots, J$ indexes the classes [Sutton, 2005]. In order to avoid over fitting, pruning is needed to determine an optimal tree size. The chosen pruning method was the 10-fold cross validation estimate with a cost complexity parameter and the one standard error rule.

3.1.2. *Single Tree Results.* The results presented in the paper [Caffery et al., 2010] were obtained by the CART technique using the R package "rpart". A single classification tree was conducted on each of the three variable sets mentioned in Section 2.2. They discovered that the variables determining whether a patient has the anti-RO antibody and dealing with biopsies were powerful in differentiating pSS and DE. For all the non-invasive tests, rose bengal staining was proved to be the most important. We reproduced their analysis and obtained the same conclusions. Table 2 lists the important variables for all variable sets. The sensitivity and specificity calculated from each classification tree are presented in Table 3.

TABLE 2. Important Variables: Single Classification Trees

Tree	**Important Variables**
All Variables	presence of anti-Ro immunoglobulin, biopsy score for SS, temporal conjunctival staining with rose bengal in the left eye
All Non-Invasive w/o Salivary Flow	temporal conjunctival staining with rose bengal for the worst eye, the severity of a patients dry mouth symptoms out of 10, the presence of either rose bengal staining in the worst eye or Schirmer score in the worst eye, corneal staining of any kind, the value of rose bengal staining in worst eye
All Non-Invasive w Salivary Flow	temporal conjunctival staining with rose bengal for the worst eye, the severity of a patients dry mouth symptoms out of 10, the presence of either rose bengal staining in the worst eye or Schirmer score in the worst eye, the amount of salivary flow per minute with stimulation, corneal staining of any kind

TABLE 3. Sensitivity and Specificity for Single Classification Trees

Tree	**True**	**Predicted**		**Sensitivity**	**Specificity**	**Overall Error**
		DE	pSS			
All Variables	DE	75	14	99.57	84.27	4.69
	pSS	1	230			
All Non-Invasive	DE	50	39	95.70	56.18	15.31
w/o Salivary Flow	pSS	10	221			
All Non-Invasive	DE	54	35	95.24	60.67	14.38
w Salivary Flow	pSS	11	220			

3.2. **Random Forest (RF).** Random Forest is an ensemble process that builds a large collection of de-correlated trees and then averages them to gain increased accuracy in predictions and classifications. It was proposed by Breiman in the late 1990's and can be described as a variation of the classification tree method [Roberts, 2009]. Random Forest has been substantially successful for many data sets. In this paper, the Random Forest method can simply be defined as a collection of classification trees where a final conclusion about a certain response variable is drawn based on all the trees. Random Forest follows a straightforward algorithm, which is described in Section 3.2.1. Other aspects of RF will be introduced in Section 3.2.2.

3.2.1. *The Algorithm.* Random Forest is a large collection of single decision trees that are not pruned. RF can be looked as a black box, meaning one knows the inputs and outputs

but the internal mechanisms are left unseen. In the research, each tree produced in a forest provides a classification for each data point which is usually expressed in terms of "votes". RF will then declare the class that has the majority of votes to be the classification for each observation [Breiman and Cutler, 2004]. There are three parameters that should be determined before the algorithm is implemented. They are: T, the number of trees grown the forest; n_{min}, the minimum node size of a random-forest tree, and an integer range for m_{try}, the number of predictor variables randomly selected for each split. m_{try} is usually much less than the total number of predictor variables, M, ($m_{try} << M$). The default value used in many softwares, is the integer of the positive square root of M ($int(\sqrt{M})$). Also this value is held constant throughout the entire construction process. In order to grow each tree $T_i,\ i = 1, ..., T$, the following steps are executed:

1. For $i = 1, \ldots, T$
   (a). Take a random bootstrap sample (sample with replacement) of N observations from the whole data set to form a random training data set where N is the total number of observations in the data;
   (b). Grow an unpruned classification tree from the bootstrapped data. At each split, m_{try} randomly selected predictor variables are considered and tested to identify the best split; until the minimum node size n_{min} is reached.
2. Output the ensemble of trees $\{T_i\}_1^T$.

Then, the final classification for each observation is the majority vote of all the class predictions based on $\{T_i\}_1^T$.

3.2.2. *Features of Random Forest. Out-of-bag (OOB) Error*

Unlike single classification trees in which the Cross-validation (CV) error is employed to determine an optimally sized tree, RF does not require the CV errors but instead uses the Out-of-bag (OOB) error estimate. This is one of the most important features in RF technique. The OOB error is calculated based on OOB samples. The OOB samples are the observations whose predictions are constructed by averaging only those trees corresponding to bootstrap samples in which these observations did not appear. The OOB error can be easily computed during the run for each set of OOB samples.

Variable Importance

A key objective in our research is to determine which of the variables are of greatest importance to the classification of the data. The computer software R gives two criteria of ranking the variable importance: the Gini Index and randomization. The Gini index determines the important variables based on which variables provide the largest decrease in the overall impurity averaged over all the trees in the forest [Liaw and Wiener, 2009]. Randomization on the other hand takes the OOB samples and drops them down each tree, taking vote of the correctly classified ones. This is then repeated but the values for variable i are randomly permuted. The difference in the number of votes for the unmodified values and the permuted ones is averaged over all trees. This value is then standardized by dividing by the standard error to provide a value that can be used to rank each variable. Due to its simplicity and easy implementation, the Gini Index was used to determine the important variables.

Missing Values

In RF, two techniques are available to handle the missing values: mode imputation and a combination of mode imputation and a proximity measure. The proximity measures are

obtained by running all the observations down a tree, if two fall within the same terminal node, then their proximity to one another is increased by a factor of one. This process is repeated for each tree in the forest. Once a forest is completely constructed, all the proximity measures are normalized dividing by the size of the forest [Breiman and Cutler, 2004].

Using mode imputation, for classification problems, the missing values are simply replaced by the mode value of the corresponding predictor variable. This method saves computation time but sacrifices the quality and precision of the results. However, a combination of mode imputation and a proximity measure can give much better quality estimates [Breiman and Cutler, 2004].

The second method requires two steps : (1) obtain rough estimates of the missing values by using the mode imputation; (2) then if the missing value is for a continuous predictor variable it is replaced by an average of all the non-missing values which is weighted by the proximities of the observations. If it is from a categorical variable the missing value is filled in by using the non-missing value that has the largest average proximity measure. This is usually repeated 5 times to gain a reliable overall estimate [[Liaw and Wiener, 2009], [Breiman and Cutler, 2004]].

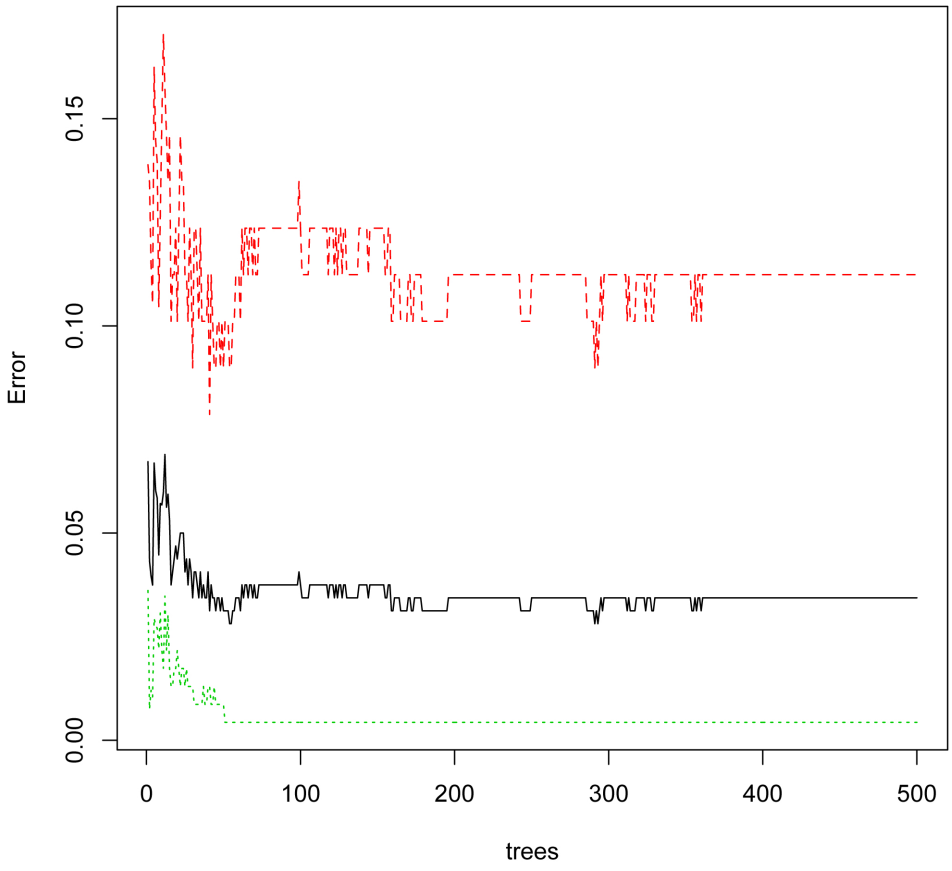
4. **Experiments and Results.** An R package, "randomForest", was used to implement the RF algorithm on the three different variable sets described in Section 2.2. They are (1) all the invasive and non-invasive variables (total 89), (2) all the non-invasive excluding salivary flow (total 67) and (3) all non-invasive variables including salivary flow (total 70).

As mentioned in Section 3.2.1, one key parameter, the number of trees grown for each forest T, is needed to be determined before the analysis is conducted. In order to figure out the most proper value of T, different values of T were exploited to create forests and the OOB error rates were calculated for each T_i. Figure 1 demonstrates the relationship between the OOB error rates and the forest size with all predictor variables. The top and bottom lines represent the OOB error rates for the DE and pSS classes respectively. The middle line is the overall OOB error rate for DE and pSS. It is clear to see that as the number of trees increases, the OOB error rate becomes stable and converges or plateaus to some value. The same experiment was also carried out for the other two variable sets and similar patterns were found. Therefore, $T = 500$ was chosen for the entire analysis.

The application of the random forest algorithm on the data produced nine different forests, three for each set of variables. This was due to the different combinations of the methods for handling missing values and the number of variables considered for each split, m_{try}. Table 4 outlines each combination implemented. Mode imputation and proximity measures are employed to fill the missing values. There are two options for m_{try}: the default value $m_{try} = |int(\sqrt{M})| = 9$ ($M = 89$ is the total number of variables) and the value chosen by the R function "tuneRF". Here, we only discuss how these combinations were applied on the variable set with all the variables. A forest was first created when the default value of $m_{try} = 9$ were employed along with the mode values (imputation) replacing the missing ones. A second forest was produced where the missing values were filled in by proximity measures, and $m_{try} = 9$. For the final combination, a m_{try} value which was twice the default value was chosen, while keeping the proximity method. This exact process was repeated on the other two variable sets. The results obtained for each set are presented in Tables 5, 6 and 7.

It was not surprising that the chosen optimal value for m_{try} in each case was the twice of the default value as the similar discovery has been found by many researchers [[Liaw and Wiener, 2009], [Breiman, 2001]]. [Breiman, 2001] stated that when selecting an optimal value

FIGURE 1. Effect on OOB error from changing the number of trees. (The top and bottom lines represent the OOB error rates for the DE and pSS classes respectively. The middle line is the overall OOB error rate.)



for m_{try} one should try half of the default, the default and the twice of the default, and then choose the best one among the three. The proximity method of filling in the missing values provided a significant improvement in the overall OOB misclassification error (a decrease between $\sim 2\%$ and $\sim 4\%$), sensitivity and specificity for all three variable sets. Moreover, the combination of the proximity method and m_{try} equal to twice the default value resulted in an even better performance for RF.

Compared to the results from a single classification tree in the paper [Caffery et al., 2010], RF significantly improved the prediction accuracy for the data set with all the variables. For the other two data sets, RF was not able to decrease the misclassification errors, so an additional tuning method was implemented (look at the single classification error rates in Table 3). [Breiman and Cutler, 2004] pointed out that it can be beneficial to utilize the important variables at each split. Therefore, in the following analysis, we only selected the possible predictor variables from a set containing only a number of variables that were deemed important, which was determined by the initial run of random forest algorithm. The R function "varSelRf" provides a convenient way to identify those important variables.

TABLE 4. Tuning Method Combinations

Variable Set	**Random Forest**	**Missing Values**	m_{try} **Value**
All Variables	1	Imputation (Mode Values)	9 (default)
(Total of 89)	2	Proximity	9 (default)
	3	Proximity	18 (tuneRF)
All Non-Invasive	4	Imputation (Mode Values)	8 (default)
w/o Salivary Flow	5	Proximity	8 (default)
(Total of 67)	6	Proximity	16 (tuneRF)
All Non-Invasive	7	Imputation (Mode Values)	8 (default)
w Salivary Flow	8	Proximity	8 (default)
(Total of 70)	9	Proximity	16 (tuneRF)

TABLE 5. Sensitivity and Specificity for Set Containing All Variables

Random Forest	**True**	**Predicted**		**Sensitivity**	**Specificity**	**Overall Error**
		DE	pSS			
1	DE	75	14	97.84	84.27	5.94
	pSS	5	226			
2	DE	79	10	99.57	88.76	3.44
	pSS	1	230			
3	DE	81	8	99.57	91.01	2.81
	pSS	1	230			

TABLE 6. Sensitivity and Specificity for Set Containing All Non-Invasive w/o Salivary Flow

Random Forest	**True**	**Predicted**		**Sensitivity**	**Specificity**	**Overall Error**
		DE	pSS			
4	DE	43	46	93.94	48.31	18.75
	pSS	14	217			
5	DE	44	45	96.54	49.44	16.56
	pSS	8	223			
6	DE	48	41	95.67	53.93	15.94
	pSS	10	221			

How many important variables should be considered? Four different choices are available based on the OOB errors given by the function "varSelRF", and are listed in Table 8. The range of the number of important variables in Table 8 was based on selection criteria. Take the third data set containing all non-invasive variables with salivary flow as an example, the range of the number of important variables can be as small as 18 and as large as 29 if the smallest OOB error is required. On the other hand, the range can be as wide as 5 to 70

TABLE 7. Sensitivity and Specificity for Set Containing All Non-Invasive w Salivary Flow

Random Forest	**True**	**Predicted**		**Sensitivity**	**Specificity**	**Overall Error**
		DE	pSS			
7	DE	49	40	90.48	55.06	19.38
	pSS	22	209			
8	DE	54	35	93.51	60.67	15.62
	pSS	15	216			
9	DE	56	33	92.64	62.92	15.62
	pSS	17	214			

when the smallest OOB+1SE is used as the selecting criterion. As before, the default value of m_{try} equals $|int(\sqrt{M_{important}})|$ ($M_{important}$ is the number of important variables chosen by the R function "varSelRF" based on different criteria), and three different values were tried: the default value, half the default value, and two times the default value. Overall, a total of thirty six forests were produced.

TABLE 8. Range of Important Variables

	Selection Criteria	
Variable Set	**Smallest OOB Error**	**Smallest OOB+1SE**
All Variables (Total 89)	15-30	5-37
All Non-Invasive w/o Salivary Flow (Total 67)	27-34	14-67
All Non-Invasive w Salivary Flow (Total 70)	18-29	5-70

By using only the important variables for each split, we were able to improve the prediction accuracy or decrease the OOB misclassification rate for two of the variable sets: the set with all the variables and the set with all the non-invasive variables with salivary flow. However, for the set containing all of the non-invasive variables without salivary flow, there was no such improvement but we were able to achieve an error equal to that of the single tree. It should be noted that for all sets the combination of using the largest number of important variables that gives smallest OOB errors and twice the default value of m_{try} gave the best performance. The random forests that achieved the smallest OOB misclassification errors were then chosen as the three final forests. The results are summarized in Table 9. For the set containing all variables, the forest was the one with 30 important variables and $m_{try} = 10$. The forest with 18 important variables and $m_{try} = 4$ was chosen for the set with all the non-invasive variables with salivary flow. Finally, for the set with all the non-invasive variables without salivary flow, the forest was the one by using 34 important variables and $m_{try} = 2$.

TABLE 9. Sensitivity and Specificity for Selected Random Forests

Random Forest	**True**	**Predicted**		**Sensitivity**	**Specificity**	**Overall Error**
		DE	pSS			
All Variables	DE	82	7	99.57	92.13	2.50
	pSS	1	230			
All Non-Invasive	DE	49	40	96.10	55.06	15.31
w/o Salivary Flow	pSS	9	222			
All Non-Invasive	DE	61	28	93.94	68.54	13.12
w Salivary Flow	pSS	14	217			

TABLE 10. Error Rate Comparison

	Single Classification Tree	**Random Forest**
All Variables	4.69%	2.50%
All Non-Invasive w/o Salivary Flow	15.31%	15.31%
All Non-Invasive w Salivary Flow	14.38%	13.12%

As shown in Table 10, with some form of tuning for RF, we were able to improve the overall OOB misclassification rates for the sets with all the variables and with all the non-invasive variables with salivary flow. Particularly, for the set with all the variables, RF gave an overall decrease of 2.19% from 4.69% to 2.5%. While, an overall decrease of 1.26% from 14.38% to 13.12% was shown for the set with all the non-invasive variables with salivary flow. For the second variable set (all non-invasive variables without salivary flow) an error of 15.31% was achieved by RF, the same as that obtained by the single classification tree. It is well-known that RF is very robust to a large data set with noise variables. This explains why RF did remarkable job compared to a single classification tree for the variable set with 89 predictors. For the other two variable sets, there are not many noise variables, i.e. all the variables are important for the classification problem. Therefore, the performance of a single classification is comparable to the results of random forest.

FIGURE 2. RF Variable Importance: All Variables (Total 89)

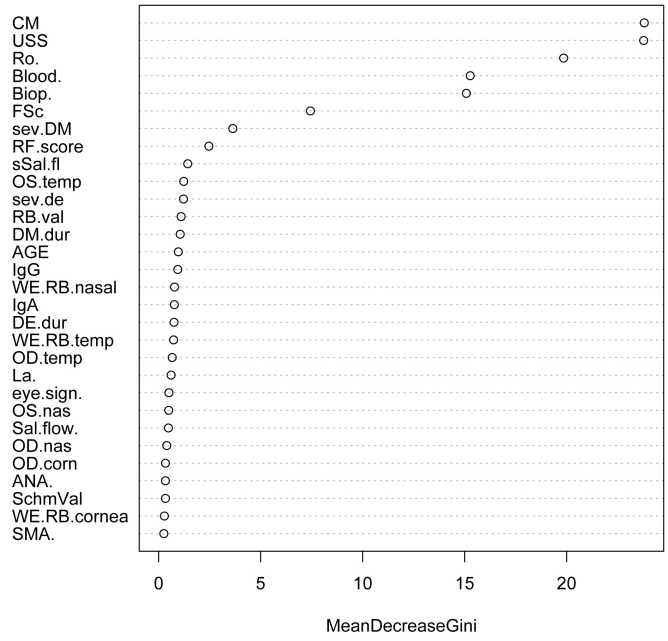
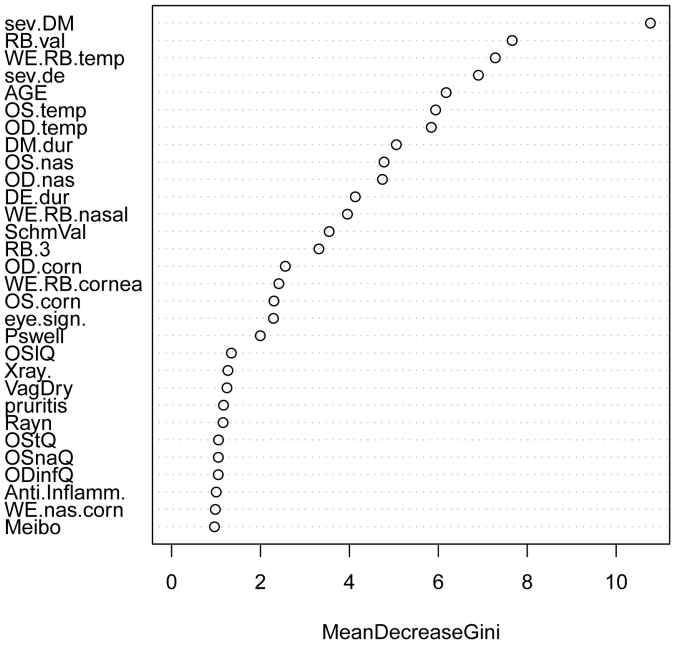
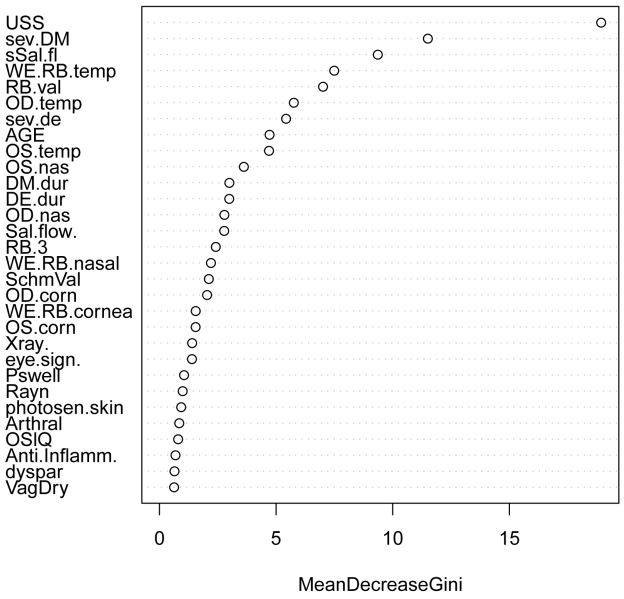


FIGURE 3. RF Variable Importance: Non-Invasive w/o SFlow (Total 67)



Figures 2, 3, and 4 are the plots presenting the ranking of variables for the three variable sets respectively. The figures also indicate how much each variable was able to decrease the

FIGURE 4. RF Variable Importance: Non-Invasive w SFlow (Total 70)



Gini Index on average. The top five variables for the first variable set with 89 variables were all in a far top-right region, and well separated from the remaining 25 variables. In terms of how well they decreased the Gini Index these five variables were all around or above fifteen. The second forest's top variables were somewhat separated from the remaining ones where they all provided decreases of at least seven with the majority of the remaining falling in a range between one and five. The final forest had only one variable that stood out, i.e. the amount of unstimulated salivary flow per minute (USS). This variable provided on average a decrease of around twenty, while the next highest decrease was only around twelve. The majority of the other variables were generally clustered together and provided decreases that were less than five. The top five important variables for all three sets are listed in Table 11.

Now a comparison can be conducted between the variables selected by RF and those by single classification trees. For the variable set with 89 variables, two of the three important variables identified by the single classification tree were also selected by RF. They are the presence of the anti-RO immunoglobulin (referred to as "Ro." in the remaining part of the paper. Similar abbreviated names will be used for other variables also.) and the biopsy score for SS (referred to as "Biop."). The one not ranked in the top of the variables by RF was whether or not a patient had signs of rose bengal staining of the temporal conjunctival in their left eye (referred to as "OS.temp"), but it was still within the first quartile of variables (ranked 16th out of a total of 89) and still considerable high. Three of the five variables that were selected by RF for the second variable set (all non-invasive w/o salivary flow) were identical tests to the ones found by the single tree. RF ranked the patients age (referred to as "AGE") and the severity of their dry eye symptoms (referred to as "sev.DE") to be in the top five important variables, while the single classification tree chose the presence of either rose bengal staining in the worst eye or Schirmer score in the worst eye (referred to as "eye.sign"), and whether or not the patient had any kind of corneal staining (referred to as "corn.stain.l"). These variables are not similar in the least. It can be noted though that

TABLE 11. Important Variables

Random Forest	**Important Variables**
All Variables	Chisholm Mason biopsy score, unstimulated salivary flow, Ro present, serum antibodies for Ro and/or La, biopsy
All Non-Invasive w/o Salivary Flow	severity of dry mouth symptoms, rose bengal value in worst eye, severity of dry eye symptoms, age of the patient, rose bengal staining of the temporal in worst eye
All Non-Invasive w Salivary Flow	unstimulated salivary flow, severity of dry mouth symptoms, salivary flow positive, rose bengal staining of the temporal in worst eye, rose bengal value in worst eye

four of the next five important variables all deal with staining of some part of a patient's eye, which are similar to the ones chosen by single trees. For the final set of variables, three of the five top ones remained the same for both methods again but the remaining two differed greatly as RF chose the variables, the value of rose bengal staining in worst eye out of 9 (referred to as "RB.val") and USS, while the single classification tree selected "eye.sign" and "corn.stain.l". RF did not even rank the "corn.stain.l" variable to be in the top thirty, and only chose "eye.sign" as the one which only provided a decrease in Gini Index of a little over one. One could see that although the final two variables selected by the single tree are not even ranked closely to the top, if you look at the next few variables beyond the top five from RF, it can be noted that they are "OS.temp" and "OD.temp". These two are qualitative variables that deal with rose bengal staining of the patients eyes, while the two chosen by the single classification tree are also qualitative ones dealing with some sort of staining.

The variables that are deemed important by random forest analysis are usually very similar to the ones chosen by the single classification tree analysis. Rose bengal staining of a patient's eye or more specifically the temporal region was always the one of the crucial non-invasive variables selected by both methods.

5. **Conclusion.** We have discovered that in all cases using the proximities of the predictor variables to fill in the missing values rather than the mode values provides smaller OOB estimates of the test set error and an increase of the specificity and sensitivity. Also a very small decrease in error can be obtained if the number of chosen predictor variables, m_{try}, is changed from the default value to the twice of the default or half of it.

Aside from a couple of variables, all the ones which were deemed as important by RF were also chosen to be equally important by the single classification trees. Particularly, both rose bengal staining and the questionnaires concerning the severity of a patients dry mouth and dry eye symptoms were crucial in differentiating pSS from DE.

The method of RF retains many advantages that single classification trees have including the ability to handle both categorical and continuous predictor variables. The main advantage of RF is the superior accuracy when classifying data or estimating values for the data in comparison to the other methods with similar purposes. Also in terms of efficiency, RF can easily deal with hundred to thousands of potential predictor variables without having

to delete any of them from the data [Breiman and Cutler, 2004]. A disadvantage of RF is computational as it requires a large amount of memory in order to properly store the large amount of trees produced by a single forest [Breiman and Cutler, 2004]. We found that a single classification tree could be grown instantaneously ($<<$ 1 second) whereas growing a RF using $T = 500$ and iterating the proximity imputation 5 times took around 12 seconds.

REFERENCES.

M. Berry and G. Linoff. *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley & Sons Inc., 2000.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman and A. Cutler. Random forests, 2004. URL `http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm`.

B. Caffery, Trefford Simpson, S. Wang, D. Bailey, J. McComb, J. Rutka, A. Slomovic, and A. Bookman. Rose bengal staining of the temporal conjunctiva differentiates sjogren's syndrome from keratoconjunctivitis sicca. *Investigative Ophthamology & Visual Science*, 51(5):2381–2387, 2010.

R. Fox. Sjogren's syndrome: Diagnostic and pathogenetic features. *Current Opinion in Rheumatology*, 8:435–438, 1996.

M. Kantardzic. *Data Mining: Concepts, Models, Methods, and Algorithms*. Wiley Interscience, 2003.

M. Lemp, C. Baudouin, M. Dogru, and G. Foulks. The definition and classification of dry eye disease: Report of the definition and classification subcommittee of the international dry eye workshop (2007). *Ocular Surface*, 5:75–92, 2007.

Andy Liaw and Matthew Wiener. *Package "randomForest"*, 2009. URL `http://cran.r-project.org/web/packages/randomForest/index.html`. R package version 3.1-29, R port by Andy Liaw.

D. Roberts. Random Forests for R, 2009. URL `http://ualberta.ca/~drr3/research/RF.htm`.

Clifton D. Sutton. Classification and regression trees, bagging, and boosting. *Handbook of Statistics*, 24:303–315, 2005.

C. Vitali, S. Bombardieri, R. Jonsson, H. Moutsopoulos, E. Alexander, S. Carsons, T. Daniels, P. Fox, R. Fox, S. Kassan, S. pillemer, N. Talal, and M. Weisman. Classification criteria for sjogren's syndrome: a revised version of the european criteria proposed by the american-european consensus group. *Annals of Rheumatic Diseases*, 61:554–558, 2002.

Y. Yohannes and J. Hoodinott. Classification and regression trees: An introduction. Technical report, International Food Policy Research Institute, 1999.

E-mail address: `x2008kyr@stfx.ca`