ForEx++: A New Framework for Knowledge Discovery from Decision Forests

Md Nasim Adnan
School of Computing and Mathematics
Charles Sturt University
madnan@csu.edu.au, zislam@csu.edu.au

Md Zahidul Islam
School of Computing and Mathematics
Charles Sturt University

Abstract

Decision trees are popularly used in a wide range of real world problems for both prediction and classification (logic) rules discovery. A decision forest is an ensemble of decision trees and it is often built for achieving better predictive performance compared to a single decision tree. Besides improving predictive performance, a decision forest can be seen as a pool of logic rules (rules) with great potential for knowledge discovery. However, a standard-sized decision forest usually generates a large number of rules that a user may not able to manage for effective knowledge analysis. In this paper, we propose a new, data set independent framework for extracting those rules that are comparatively more accurate, generalized and concise than others. We apply the proposed framework on rules generated by two different decision forest algorithms from some publicly available medical related data sets on dementia and heart disease. We then compare the quality of rules extracted by the proposed framework with rules generated from a single J48 decision tree and rules extracted by another recent method. The results reported in this paper demonstrate the effectiveness of the proposed framework.

Keywords: decision tree; decision forest; random forest; knowledge discovery; dementia; heart disease

1 Introduction

Due to their capabilities of expressing knowledge in a human-understandable form and dealing with high dimensional as well as redundant attributes, decision trees are applied in a wide range of real-world problems (Murthy (1998), Safavian and Landgrebe (1991)). Aside from those properties, decision trees are considered to be an unstable classifier as slight differentiation in a training data set can result in significant differences between decision trees obtained from the original and differentiated data sets. Interestingly, an ensemble of classifiers is found to be effective for unstable classifiers such as decision trees (Tan et al. (2005)). A decision forest is an ensemble of decision trees where an individual decision tree acts as the base classifier. The classification is performed by taking a vote based on the predictions made by each decision tree of the decision forest (Tan et al. (2005)). A decision forest in general incorporates and extends most of the capabilities of decision trees, such as a decision forest is said to be more robust to noise(s) as well as more accurate compared to a single decision tree (Bernard et al. (2008), Polikar (2006), Quinlan (1996a)).

In order to achieve better ensemble accuracy a decision forest needs both accurate and diverse (in terms of classification errors) individual decision trees as base classifiers (Adnan and Islam (2016a), Adnan and Islam (2015a), Tin Kam (1998), Polikar (2006)). An accurate decision tree can be generated by applying a decision tree algorithm such as C4.5 (Quinlan (1993), Quinlan (1996b)) on a training data set. However, a single decision tree can discover only one set of rules and thus may wrongly predict the class value of a test record which could have been predicted correctly by a more appropriate rule. A different decision tree can be obtained from a differentiated data set (such as a bootstrap sample) which may include a more appropriate rule for the given test record. If a decision forest contains a set of decision trees which are

different from each other, then some of the trees may discover more appropriate rules for a set of test records while some other trees may discover for another set of test records, resulting in better generalization performance for the forest (Adnan and Islam (2016c)). Besides, a decision forest can be seen as a pool of rules with great potential for knowledge discovery.

The main challenge of knowledge discovery from a standard-sized decision forest (usually a 100-tree decision forest (Adnan and Islam (2015b)), Adnan and Islam (2015c), Geurts et al. (2006)) comes from the enormous number of rules that it generates. For effective knowledge discovery, we need to extract accurate, generalized (meaning high in coverage), concise as well as surprising (previously unknown) rules (Geng and Hamilton (2006)). An obvious technique to extract a subset of rules is to apply some cut points based on accuracy, coverage or length of the rules. For example, we can extract only those rules that have accuracy >= 80%. However, any such cut points may react differently from data set to data set. As a result, for one data set a cut point may net more than manageable rules whereas for another data set the same cut point may acquire a few rules. To avoid such situation, a user may need to expedient on different cut points for a single variable (such as accuracy) and subsequently on all possible combinations of the variables for each data set which may not be manageable either.

There are many subforest selection algorithms that prunes a number of decision trees from a decision forest while retaining or increasing ensemble accuracy (Adnan and Islam (2016c), Lu et al. (2010), Margineantu and Dietterich (1997), Martínez-Muñoz et al. (2009), Martínez-Muñoz and Suárez (2004), Ruta and Gabrys (2005)). It was shown in Adnan and Islam (2016c)) that if the number of trees in a subforest drops considerably, the ensemble accuracy also drops significantly. It was also shown that on an average the best subforest selection algorithm amassed around 45 trees from a 100-tree Random Forest (Adnan and Islam (2016c)). Admittedly, subforest selection algorithms can shrink the size of a decision forest, yet the size remains large enough to hinder effective knowledge discovery. In addition, these algorithms do not consider any rule-level properties for selecting trees and thus trees selected by them may not contain rules that are comparatively more accurate, generalized and concise than others. Similarly, there are some algorithms that intend to increase the comprehensibility of a decision forest by representing the whole decision forest into a single decision tree. In one algorithm (Johansson et al. (2011)), the authors approximated the entire forest by selecting the nearest tree in terms of prediction results. Undoubtedly this algorithm can increase the comprehensibility of a decision forest; however cannot exploit its knowledge discovery potential.

Extraction of rules from different types of classifiers is a popular area of research; in recent years a number of rule extraction methods have been proposed (Mashayekhi and Gras (2015), Martens et al. (2008), Schmitz et al. (1999)). In Huysmans et al. (2006), the authors provided a wide range of survey on different rule extraction methods. Most of those rule extraction methods were designed for some "black box" type classifiers such as Artificial Neural Networks and Support Vector Machines (Martens et al. (2008), Schmitz et al. (1999)). On the contrary, rule extraction from decision forests remains largely ignored (Huysmans et al. (2006), Mashayekhi and Gras (2015)). Some existing rule extraction techniques (Liu et al. (2012), Mashayekhi and Gras (2015)) mainly prune forest rules in order to increase the prediction accuracy (just like pruning trees from a forest) and consequently do not solely focus on issues related to knowledge discovery.

In order to facilitate effective knowledge discovery from decision forests, recently we have proposed a data set independent framework (*ForEx*) for extracting those rules that are comparatively more accurate and generalized (high in coverage) than others. *ForEx* has been accepted in a conference (AusDM 2016) (Adnan and Islam (2016b)) but we have never submitted/published any variant of *ForEx* in any journal. In this paper, we extend *ForEx* substantially to propose *ForEx++* as follows.

- In addition to accuracy and coverage, *ForEx++* considers rule length as concise rules are said to be more comprehensible (Geng and Hamilton (2006)). On the other hand, *ForEx* does not consider rule length while extracting forest rules.

- A major limitation of *ForEx* is that its rules can be dominated by the majority class and thus may miss to net rules from *minority class(es)* (see Section 3 for details). In order to oblige the presence of rules from *minority class(es), ForEx++* considers accuracy, coverage and rule length independently for *each class* (see Section 3 for details).
- Unlike *ForEx*, *ForEx++* removes identical (exactly the same) rules from decision forests before extracting any rule. Compounded by the randomness of decision forests, rules reported in this paper are substantially different from rules reported in our conference paper (Adnan and Islam (2016b)).
- This paper involves wider experimentation and interesting knowledge/trends analysis. We apply *ForEx++* on rules generated by two different decision forest algorithms from two different publicly available medical related data sets on dementia (Oasis) and heart disease (Lichman). *ForEx++* rules are compared with the J48 (the Weka implementation of C4.5 (Quinlan (1993), Quinlan (1996b)) rules and rules extracted by a recent method (Mashayekhi and Gras (2015)). Furthermore, based on the rules extracted by *ForEx++*, both data sets are further explored using SQL in order to find interesting knowledge/trends. On the other hand, in our conference paper (Adnan and Islam (2016b)), *ForEx* was applied on the dementia data set (not on the heart disease data set) and the extracted rules were compared only with rules generated from a J48 tree (Hall et al. (2009)) (not with rules extracted by a recent method). No interesting knowledge/trends were reported in our conference paper.

The remainder of this paper is organized as follows: In Section 2 we introduce data set, decision tree, two different decision forest algorithms and the recent rule extraction method as Background Information. Section 3 explains the proposed Knowledge Extraction Framework. Section 4 provides Detailed Description of the Data Sets used and the associated Experimental Results. Finally, we offer some Concluding Remarks in Section 5.

2 Background Information

2.1 Data Set

A data set ***D*** is regarded as a two dimensional table with columns/attributes ({*A1, A2, ..., Am, C*}) and rows/records ({*R1, R2, ..., Rn*}). A data set can have two broad types of attributes: numerical (e.g. *Age* and *Income*) and categorical (e.g. *City* and *Degree*). While a numerical attribute has a natural ordering among its domain values, a categorical attribute does not. Out of all attributes, one categorical attribute is chosen to be the class attribute and all other attributes are termed as non-class attributes. A domain value of the class attribute $c_i \epsilon C$ is the class label of a record R_i and popularly termed as the class value. If minority records have a certain class value, the class value is termed as a *minority class*. Conversely, if majority records have a certain class value, the class value is termed as the *majority class*. If there are more than two class values then one of them is the *majority class* and the others are *minority classes*.

2.2 Decision Tree

A decision tree consists of nodes (denoted by rectangles) and leaves (denoted by ovals) as shown in Figure 1. The node of a decision tree symbolizes a splitting event where the splitting attribute (label of the node) partitions a data set according to its domain values. As a result, a disjoint set of horizontal segments of the data set is generated and each segment contains one set of domain values of the splitting attribute. A leaf of a decision tree represents a horizontal segment of the data set where no further splitting is carried out. In this way, all records of a training data set are distributed among the leaves.

The path from the root node to a leaf makes up a rule (i.e. pattern) that identifies a relationship between a set of non-class attributes (splitting attributes along the path) and the class values. For example, the rule for Leaf 1 is "if *Degree = Masters* AND *Income <= 85K → Lecturer*" as majority records (all four in this case) belonging to the segment represented by Leaf 1 have the class value *Lecturer*. Here, "if *Degree = Masters* AND *Income <=*" is the *antecedent* of the rule

and "*Lecturer*" is the *consequent*. A decision tree is then used to predict the class values of unseen records of a testing data set for which the class values are unknown (i.e. records are unlabeled). Based on the values of non-class attribute/s of an unlabeled record it passes through an *antecedent* to be predicted as the *consequent* (Adnan and Islam (2014)).

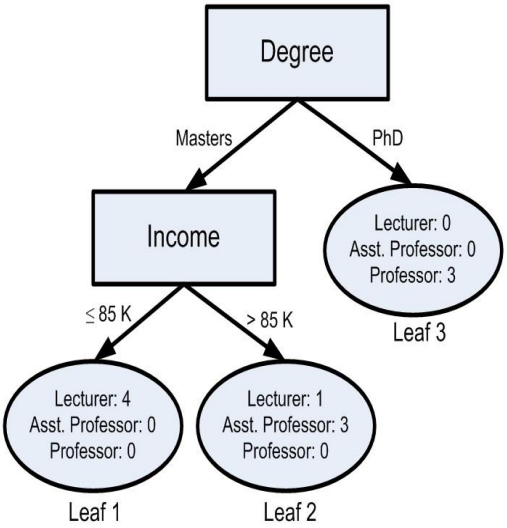

Figure 1: Decision Tree

Out of all rules obtained from a decision tree, some may have *consequent* from *minority class(es)* and some may have *consequent* from the *majority class*. In this paper, for simplification, "rules with *consequent* from *minority class(es)*" is called "rules from *minority class(es)*" and "rules with *consequent* from the *majority class*" is called "rules from the *majority class*".

2.3 Decision Forest

2.3.1 Random Forest

Random Forest (Breiman (2001)) is regarded as a state-of-the-art decision forest algorithm (Bernard et al. (2012), Bernard et al. (2008)) which is technically a fusion of Bagging (Breiman (1996)) and Random Subspace (Tin Kam (1998)) algorithms.

Bagging (Breiman (1996)) generates a new training data set $\boldsymbol{D}_i$ where the records of $\boldsymbol{D}_i$ are selected randomly from the original training data set $\boldsymbol{D}$. A new training data set $\boldsymbol{D}_i$ contains the same number of records as in $\boldsymbol{D}$. Thus, some records of $\boldsymbol{D}$ can be selected multiple times and some records may not be selected at all. Approximately, 63.2% of the original records are selected in a bootstrap sample and the remaining 36.8% records are repeated (Han et al. (2011)). Bagging generates a predefined number (T) of bootstrap samples ($\boldsymbol{D}_1$, $\boldsymbol{D}_2$, …, $\boldsymbol{D}_T$) using the above mentioned approach. A decision tree building algorithm is then applied on each bootstrap sample $\boldsymbol{D}_i$ (i = 1, 2, …, T) in order to generate altogether T number of trees for the forest.

The Random Subspace algorithm (Tin Kam (1998)) algorithm randomly draws a subset of attributes (subspace) $\boldsymbol{f}$ from the entire attribute space $\boldsymbol{m}$. $\boldsymbol{f}$ can be drawn either at the tree level or at the node level. When selected at the tree level, attributes in $\boldsymbol{f}$ remains the same for each node of a tree; on the other hand attributes in $\boldsymbol{f}$ may differ from one node to another in a tree when selected at the node level. The best attribute in $\boldsymbol{f}$ is determined to be the splitting attribute for the associated node. The Random Subspace algorithm is applied on the original training data set (not on bootstrap samples) for building decision trees.

Random Forest in its simplest form uses bootstrap sampling and node level subspacing for generating decision trees ((Breiman, 2001)).

2.3.2 Forest PA

Recently, we have proposed a new decision forest algorithm called "Forest by Penalizing Attributes (*Forest PA*)" (Adnan and Islam (2017)) that imposes penalties on attributes systematically in such a way that an attribute tested at lower level (such as in the root node) receives higher penalty (lower weight) than an attribute tested at higher level. The reason is that an attribute tested at lower level may influence more rules than an attribute tested at higher level as discussed earlier. Thus in order to discover diverse set of rules, attributes tested at lower levels are supposed to be avoided in a future tree more seriously than those attributes that are tested at higher levels. Furthermore, to increase the chance of having different weights among attributes in the same level, *Forest PA* randomly selects the weight of an attribute from the Weight-Range (WR) allocated for the attribute's level (Adnan and Islam (2017)).

Forest PA also has a mechanism to gradually increase weights (withdraw penalties) from the attributes that have not been tested in the subsequent trees. This addresses the situation where all good attributes (attributes with high classification capacity) suffer from high penalties and thus poor quality trees can be generated at some later stage of the forest building process. *Forest PA* also uses bootstrap samples of the original training data set (Breiman (1996)) to ensure further diversity.

2.4 The RF+HC method

Similar to finding the best subforest, finding the best subset of rules by exhaustive search over any criteria is an NP-Hard problem (Adnan and Islam (2016c), Mashayekhi and Gras (2015)). In order to avoid such computational burden, rule extraction problem can be solved by a heuristic or a hill climbing approach as suggested in the RF+HC method (Mashayekhi and Gras (2015)). In doing so, the RF+HC method first ranks all rules using Equation 1.

$$rule\_score = \frac{cc-ic}{cc+ic} + \frac{cc}{ic+k} + \frac{cc}{rl}$$

Equation 1

Here, *cc* is the number of records that is covered and correctly classified by a rule. *ic* is the number of records that is covered and incorrectly classified by the same rule. We see, a positive constant *k* is used at the denominator of the second term in Equation 1. Here, *k* is used to prevent the denominator to become zero when all records covered by a rule are classified correctly (i.e. *ic* = 0). In the RF+HC method, the value of *k* is set to 4 (although, any value can be used instead). Equation 1 also incorporates the influence of rule length to prioritize concise rules as they are more preferred (Geng and Hamilton (2006)). The denominator of the third term *rl* (which stands for rule length) of the equation imposes more penalties for longer rules.

Now, as suggested in the RF+HC method, a subset of rules can be selected using a heuristic or a hill climbing approach. In the heuristic approach, rules are sorted based on their individual ranks. Then, a user defined number of top rules are extracted according to the ranks. In the hill climbing approach, at first a subset of rules are probabilistically selected on the basis of their ranks (meaning a rule with higher rank has greater chance to be selected than a rule with lower rank). Then, each time one rule from the rest is added in the subset and if the change increases the overall accuracy then the change is retained (Mashayekhi and Gras (2015)).

3 The Proposed Rule Extracting Framework

3.1 Basic Concepts

The path from the root node to a leaf node makes up the *antecedent* and the majority class of the leaf node is the *consequent* of the rule in a decision tree. For example, the rule for Leaf 2 in Figure 1is "if *Degree = Masters* AND *Income >85K→ Class Value = Asst. Professor*" as majority records (Lecturer: 1, Asst. Professor: 3) belonging to Leaf 2 have the class value *Asst. Professor*. The minority record(s) are viewed as if they are misplaced under a rule. Thus, the

Accuracy (*Acc*) of the rule for Leaf 2 is: 3/4 = 0.75. All records of a data set from which a decision tree is built are distributed among the leaves. The Coverage (*Cov*) of a rule indicates the proportion of records that fall in the leaf to the total records of the data set. For example, *Cov* of the rule for Leaf 2 is: 3/11 = 0.27. *Acc* and *Cov* are independent of each other and thus no data set independent relationships exist between them.

In general, *Acc* interprets the reliability of a rule whereas *Cov* interprets its generality (Geng and Hamilton (2006)) and thus both of them are important for a rule. For example, a low-coverage rule is likely to be uninteresting from a business perspective as it may occur simply by chance even if the rule is highly reliable (Tan et al. (2005)). Hence, both *Acc* and *Cov* need to be considered when extracting a subset of rules. However, in literature we find no method considering both *Acc* and *Cov* for extracting subsets of rules from decision forests (Huysmans et al. (2006)). The recently proposed RF+HC method (Mashayekhi and Gras (2015)) considers the number of correctly and incorrectly classified records that are covered by a rule (in effect, accuracy) and the rule length for calculating a score for the rule (see Equation 1). Hence, the influence of coverage is not directly considered for calculating scores in the RF+HC method.

3.2 Priorities for the Proposed Rule Extracting Framework

3.2.1 Rule Extraction from Each Class

We understand that no data set independent relationships exist between *Acc* and *Cov*; however they are independently comparable between themselves. Based on this principle, recently we have proposed a rule extraction framework for decision forest *ForEx* (Adnan and Islam (2016b)) that finds the set of rules $\mathbf{R}_{Avg}^{Acc}$ that are more (or equally) accurate than the average accuracy Acc_{Avg} of all rules ($\{\mathbf{R} = R_1, R_2, \ldots, R_z\}$) in a forest as follows, where R_i^{Acc} is the accuracy of R_i.

$$\mathbf{R}_{Avg}^{Acc} = \{R_i : R_i^{Acc} \geq Acc_{Avg}\} | Acc_{Avg} = \frac{1}{|\mathbf{R}|}\sum_{j=1}^{|\mathbf{R}|} R_j^{Acc}$$

Equation 2

Similarly, *ForEx* (Adnan and Islam (2016b)) that finds the set of rules $\mathbf{R}_{Avg}^{Acc}$ that have more (or equal) coverage than the average coverage Cov_{Avg} of all rules in a forest as follows, where R_i^{Cov} is the coverage of R_i.

$$\mathbf{R}_{Avg}^{Cov} = \{R_i : R_i^{Cov} \geq Cov_{Avg}\} | Cov_{Avg} = \frac{1}{|\mathbf{R}|}\sum_{j=1}^{|\mathbf{R}|} R_j^{Cov}$$

Equation 3

Now, by applying both Equation 2 and Equation 3, *ForEx* is able to recognize those high quality rules that have simultaneously more (or equal) accuracy and coverage than their averages in a forest (see Equation 4).

$$\mathbf{R}_{Avg}^{Acc.Cov} = \mathbf{R}_{Avg}^{Acc} \cap \mathbf{R}_{Avg}^{Cov}$$

Equation 4

From Equation 3 and Equation 4 we understand that *ForEx* rules can be dominated by the *majority class* and thus may miss to net rules from *minority class(es)*. The reason is: the records with class values from *minority class(es)* can be far less than the records with class values from the *majority class* and thus *Cov* of rules from *minority class(es)* may fall short of Cov_{Avg}. As a result, rules from *minority class(es)* may not qualify for $\mathbf{R}_{Avg}^{Acc.Cov}$ making *ForEx* ineffective for extracting interesting knowledge from imbalanced data (see Equation 4).

As an improvement to *ForEx*, in this paper we propose *ForEx++* that obliges the presence of rules from *minority class(es)* by applying Equation 2 and Equation 3 separately on rules from each class. *ForEx++* finds the set of rules $\mathbf{R}^{Acc}_{Avg,c_k}$ that are more (or equally) accurate than the average accuracy Acc_{Avg,c_k} of all rules with the c_k class value ($\mathbf{R}_{C_k}$ = {R_h, R_i, ..., R_y}) in a forest as follows. Note that R^{Acc}_{j,c_k} represents the accuracy of the j-th rule with the c_k class value (i.e. $R_j \in \mathbf{R}_{C_k}$) and $|\mathbf{R}_{C_k}|$ expresses the number of rules with the c_k class value.

$$\mathbf{R}^{Acc}_{Avg,c_k} = \{R_i : R^{Acc}_{i,c_k} \geq Acc_{Avg,c_k}\} | Acc_{Avg,c_k} = \frac{1}{|\mathbf{R}_{c_k}|} \sum_{j=1}^{|\mathbf{R}_{c_k}|} R^{Acc}_{j,c_k}$$

Equation 5

Similarly, *ForEx++* finds the set of rules $\mathbf{R}^{Cov}_{Avg,c_k}$ that have more (or equal) coverage than the average coverage Cov_{Avg,c_k} of all rules with the c_k class value $\mathbf{R}_{C_k}$= {R_h, R_i, ..., R_y}) in a forest as follows.

$$\mathbf{R}^{Cov}_{Avg,c_k} = \{R_i : R^{Cov}_{i,c_k} \geq Cov_{Avg,c_k}\} | Cov_{Avg,c_k} = \frac{1}{|\mathbf{R}_{c_k}|} \sum_{j=1}^{|\mathbf{R}_{c_k}|} R^{Cov}_{j,c_k}$$

Equation 6

3.2.2 Consideration of Rule Length

The RF+HC method incorporates the influence of rule length to prioritize concise rules using the term $\frac{cc}{rl}$ in Equation 1. In $\frac{cc}{rl}$, rl ensures more penalty for a longer rule; however the penalty can be suppressed cc (cc is the number of records that is covered and correctly classified by a rule). For example, the score from the term $\frac{cc}{rl}$ for Rule 1 with cc = 40 and rl = 10 will be more (better) than that of Rule 2 with cc = 15 and rl = 5 even when the length of Rule 1 is double and Rule 2 is more accurate with no incorrect classification (ic = 0). In such scenarios, it is really debatable to establish any direct relationship between cc and rl without considering other related variable (e.g. ic). On the other hand, *ForEx++* prioritizes the set of concise rules $\mathbf{R}^{Len}_{Avg,c_k}$ that have less (or equal) length than the average length Len_{Avg,c_k} of all rules with the c_k class value ($\mathbf{R}_{C_k}$= {R_h, R_i, ..., R_y}) in a forest as follows.

$$\mathbf{R}^{Len}_{Avg,c_k} = \{R_i : R^{Len}_{i,c_k} \leq Len_{Avg,c_k}\} | Len_{Avg,c_k} = \frac{1}{|\mathbf{R}_{c_k}|} \sum_{j=1}^{|\mathbf{R}_{c_k}|} R^{Len}_{j,c_k}$$

Equation 7

3.3 Rule Extraction by *ForEx++*

By applying Equation 5, Equation 6 and Equation 7, *ForEx++* is able to recognize those high quality rules that have simultaneously more (or equal) accuracy, coverage and conciseness than their averages for each class in a forest (see Equation 8 and Equation 9).

$$\mathbf{R}^{Acc.Cov.Len}_{Avg,c_k} = \mathbf{R}^{Acc}_{Avg,c_k} \cap \mathbf{R}^{Cov}_{Avg,c_k} \cap \mathbf{R}^{Len}_{Avg,c_k}$$

Equation 8

$$\mathbf{R}_{ForEx++} = \bigcup_{c_k \in C, \forall c_k} \mathbf{R}_{Avg,c_k}^{Acc.Cov.Len}$$

Equation 9

3.4 Steps of *ForEx++*

The major steps of ForEx++ are organized as follows. *ForEx++* is further illustrated through ***Algorithm 1.***

Step 1: In the first step, *ForEx++* removes identical (exactly the same) rules from **R** (see **Step 1** of ***Algorithm 1***).

Step 2: In the second step, *ForEx++* selects three sets of rules from each class based on accuracy, coverage and rule length using Equation 5, Equation 6 and Equation 7. It then selects the intersection of the three sets of rules using Equation 8 (see **Step 2** of ***Algorithm 1***).

Step 3: In the third step, *ForEx++* accumulates rules for all classes using Equation 9 (see **Step 3** of ***Algorithm 1***).

Algorithm 1*: *ForEx++

input: Forest Rules **R**.
output: *ForEx++* rules $\mathbf{R}_{ForEx++}$.
begin
$\mathbf{R}_{ForEx++} \leftarrow \emptyset$;
$k \leftarrow |C|$; /* $|C|$ is the number of distinct classes */
Step 1:
R← remove_Identical_Rules(**R**);
end Step 1

Step 2:
for i =1 to k do
$\mathbf{R}_{Avg,c_i}^{Acc} \leftarrow \emptyset$; $\mathbf{R}_{Avg,c_i}^{Cov} \leftarrow \emptyset$; $\mathbf{R}_{Avg,c_i}^{Len} \leftarrow \emptyset$;
$\mathbf{R}_{C_i} \leftarrow get\_Rules\_By\_Class(\mathbf{R}, c_i)$; /* $\mathbf{R}_{C_i}$ is the set of rules with the c_i class value */
/* Implementation of Equation 5 */
for j = 1 to $|\mathbf{R}_{C_i}|$ do
$R_{C_i}^{Acc} \leftarrow R_{C_i}^{Acc} + get\_Accuracy(R_{j,C_i})$; /* R_{j,C_i} is the j-th rule of $\mathbf{R}_{C_i}$ */
end for
$Acc_{Avg,c_i} \leftarrow \frac{1}{|\mathbf{R}_{C_i}|} \times R_{C_i}^{Acc}$;
for j = 1 to $|\mathbf{R}_{C_i}|$ do
if $get\_Accuracy(R_{j,C_i}) \geq Acc_{Avg,c_i}$ then $\mathbf{R}_{Avg,c_i}^{Acc} \leftarrow \mathbf{R}_{Avg,c_i}^{Acc} \cup R_{j,C_i}$;
end for
/* Implementation of Equation 6 **Error! Reference source not found.** */
for j = 1 to $|\mathbf{R}_{C_i}|$ do
$R_{C_i}^{Cov} \leftarrow R_{C_i}^{Cov} + get\_Coverage(R_{j,C_i})$;
end for
$Cov_{Avg,c_i} \leftarrow \frac{1}{|\mathbf{R}_{C_i}|} \times R_{C_i}^{Cov}$;
for j = 1 to $|\mathbf{R}_{C_i}|$ do
if $get\_Coverage(R_{j,C_i}) \geq Cov_{Avg,c_i}$ then $\mathbf{R}_{Avg,c_i}^{Cov} \leftarrow \mathbf{R}_{Avg,c_i}^{Cov} \cup R_{j,C_i}$;
end for
/* Implementation of Equation 7 */
for j = 1 to $|\mathbf{R}_{C_i}|$ do
$R_{C_i}^{Len} \leftarrow R_{C_i}^{Len} + get\_Rule\_Length(R_{j,C_i})$;
end for
$Len_{Avg,c_i} \leftarrow \frac{1}{|\mathbf{R}_{C_i}|} \times R_{C_i}^{Len}$;
for j = 1 to $|\mathbf{R}_{C_i}|$ do

```
            if get_Rule_Length (R_{j,C_i}) ≤ Len_{Avg,c_i} then R^{Len}_{Avg,c_i} ← R^{Len}_{Avg,c_i} ∪ R_{j,C_i};
        end for
        /* Implementation of Equation 8 */
        R^{Acc.Cov.Len}_{Avg,c_i} ← R^{Acc}_{Avg,c_i} ∩ R^{Cov}_{Avg,c_i} ∩ R^{Len}_{Avg,c_i};
    end for
  end Step 2

  Step 3:
    /* Implementation of Equation 9 */
    for i = 1 to k do
        R_{ForEx++} ← ∪ R^{Acc.Cov.Len}_{Avg,c_i};
    end for
  end Step 3
  return R_{ForEx++};
end
```

4 Experiments

The experimentation is conducted on two different publicly available medical related data sets on dementia (Oasis) and heart disease (Lichman). We now provide detailed description of the data sets and their associated experimental results in the following sections.

4.1 Dementia

Dementia is a medical term linked with memory loss which is severe enough to complicate daily life (Dementia). Alzheimer's disease (AD) is the most common cause of dementia, accounting for almost 70% of all dementia cases (Dem). AD is caused by abnormal deposition of proteins in the brain that destroys cells in those areas of the brain (i.e. cerebral cortex) that are responsible for memories, thoughts, actions and personality. Unfortunately, AD is degenerative (destroys brain cells that causes to shrink the size of the brain over time), progressive (gradual decline of functioning of effected areas of the brain) and thus irreversible (Dem).Vascular Dementia (VaD) is mainly caused by full or partial blocking of arteries in the brain from deposits of fats, dead cells, and other debris that ultimately disrupts the blood flow. Vascular dementia is often related to high blood pressure, high cholesterol, heart disease, diabetes, and other related conditions. Treating those conditions can slowdown the progress of vascular dementia, but as usual the brain functions that are lost are not recoverable. Unlike in AD, in VaD the size the brain may not shrink at all (Dem).

4.2 Data Set on Dementia

In this paper, we use a data set named "OASIS: Longitudinal MRI Data in Nondemented and Demented Older Adults" that consists of a collection of 354 observations (records) on 142 subjects (patients) aged between 60 to 98 and all the observations have one of the three class values: Demented, Nondemented and Converted (Oasis). Converted class value refers to the patients that develop dementia during the period of the data collection. In addition to MRI data, the data set also includes information about patient's Gender, Age, Education Level and Socio-economic status. Some scores on dementia-related examinations are also included. InTable 1, we present the attributes of the OASIS data set and explain their meanings with their respective value ranges in detail.

Attributes	Explanation
MRI ID	The unique number of tests (354 in total).
Subject ID	The number of unique patients (142 in total). One patient may be visiting multiple times for MRI tests, so the number of MRI tests (354) is larger than the number of subjects.
Visit	Chronological visit number of a patient.
MR Delay	The delay since the last visit.
Gender	Male (M) or Female (F).
Hand	Right-Handed (R) or Left-Handed (L).
Age	Ages of the patients vary between 60 to 98.
EDUC	EDUCation level of the patients vary between 6 to 23 representing years of education.
SES	Socio-Economic Status of the patients assigned through the Hollingshead Index of Social Position. 1 representing the highest status to 5 representing the lowest status (Hollingshead (1975)).
MMSE	Mini-Mental State Examination value ranges between 0 to 30. In MMSE, a health professional asks a patient a series of questions designed to test a range of everyday mental skills. The questions mainly cover preliminary arithmetic problems, simple memory tests, and recognition of different orientations of objects. A score of 20 to 24 suggests mild dementia, 13 to 20 suggests moderate dementia, and less than 12 indicates severe dementia ((Dementia), Folstein et al. (1975)).
CDR	Clinical Dementia Rating. 0 indicates No dementia, 0.5 indicates very mild dementia, 1 indicates mild dementia and 2 indicates moderate dementia (Morris (1993)).
eTIV	estimated Total Intra-cranial Volume (in cm^3) of the brain (proportional to the size of the skull, can be obtained from MRI image) (Buckner et al. (2004)).
nWBV	normalized Whole-Brain Volume, expressed as a percent of all voxels (can be obtained from MRI image) (Buckner et al. (2005)).
ASF	Atlas Scale Factor is the volume scaling factor for brain size (proportional to nWBS and eTIV (Buckner et al. (2004)).
Distinct Class values	Three distinct class values. Demented, Nondemented and Converted.

Table 1: Description of the OASIS data set

4.3 Experimentation on the OASIS Data Set

In (Ertek et al. (2014)), the authors applied J48 (Hall et al. (2009)) on a modified version of OASIS data set in order to generate a decision tree. From the original OASIS data set, at first the authors excluded the identifier attributes ("MRI ID" and "Subject ID") and then built a preliminary decision tree. However, the first split of the tree based on the attribute "CDR" perfectly distinguished the demented patients (when CDR = 1) from others (non-demented and converted). This showed that CDR was too good attribute to be included in the analysis. Furthermore, attributes "MR Delay" and "Visit" were found strongly dependent on CDR in the preliminary tree as the following splits were based on them (Ertek et al. (2014)). Observing the "near perfect" results in the preliminary decision tree due to CDR and the inherent dependency problem of "MR Delay" and "Visit", the authors (Ertek et al. (2014)) excluded them to generate the final decision tree.

Table 2 presents the rules generated from the final J48 decision tree as reported in Ertek et al. (2014) with their respective accuracies (except we prune a few lengthy rules to contain at most

five attributes in their *antecedents* for better understanding). We also report the coverages of the rules of the final J48 decision tree even though they were not reported in Ertek et al. (2014). Demented, Nondemented and Converted are abbreviated as D, ND and C respectively to be presented in subsequent tables. Also *Antecedent, Consequent*, Accuracy and Coverage are abbreviated as ***Ante, Cons*, Acc** and **Cov** respectively.

Ante	***Cons***	**Acc**	**Cov**
If MMSE <= 26	D	94%	24%
If MMSE > 28 and Gender = F	ND	85%	36%
MMSE > 28 and Gender = M and EDUC and EDUC <= 13	ND	79%	5%
If MMSE > 28 and Gender = M and EDUC between 13 and 15	D	86%	2%
If MMSE > 28 and Gender = M and EDUC <= 13	ND	80%	5%
If MMSE > 28 and Gender = M and EDUC > 15 and ASF <= 0.93	D	80%	1%
If MMSE > 28 and Gender = M and EDUC > 15 and ASF > 0.93	ND	69%	7%
If MMSE between 27 and 28 and Gender = M and nWBV <= 0.68	ND	80%	1%
If MMSE between 27 and 28 and Gender = M and nWBV > 0.68	D	62%	8%
If MMSE between 27 and 28 and Gender = F and nWBV <= 0.71	D	63%	2%
If MMSE between 27 and 28 and Gender = F and nWBV > 0.71	ND	66%	9%

Table 2: Rules from J48 Decision Tree

Both Random Forest and *Forest PA* use bootstrap samples of a training data set. Bootstrap samples are generally used for inducing diversity among the base classifiers (Martínez-Muñoz and Suárez (2010), Quinlan (1996a)). Approximately, 63.2% of the original records are selected and the remaining 36.8% records are repeated in a bootstrap sample (Han et al. (2011)). As a result of this deviation, bootstrap samples cannot be regarded as a valid source of knowledge. Therefore, for an even comparison among J48, Random Forest and *Forest PA*, we do not use bootstrap samples for both Random Forest and *Forest PA*. As a result, Random Forest is converted to Random Subspace (RS) (Tin Kam (1998)) and *Forest PA* is converted to *Forest PA* WithOut Bootstrap Samples (*Forest PA* WOBS) (Adnan and Islam (2017)).

After generating rules from both the forests, we first remove identical rules from them and then apply *ForEx++* and the RF+HC method for extracting subsets of rules. For a fair comparison, we apply heuristic approach for the RF+HC method to select the same number of top-ranked rules (in terms of rule score) as extracted by *ForEx++*. For Example, if *ForEx++* extracts 50 rules, we select the top-ranked 50 rules for the RF+HC method.

Similarly, a 20-tree *Forest PA* WOBS generates as many as 658 rules from OASIS, a cut point with accuracy >= 95% amasses 461 rules from them, whereas *ForEx++* extracts as low as 55 rules. Thus, 55 top-ranked rules are selected for the RF+HC method from *Forest PA* WOBS. Though smaller in number, it is difficult to accommodate every *ForEx++* and RF+HC rules for detailed comparison with each other and J48 rules. Hence, we present a brief comparison among them in Table 3.

Criteria	**J48 rules**	**RS**		***Forest PA* WOBS**	
		RF+HC rules	***ForEx++* rules**	**RF+HC rules**	***ForEx++* rules**
Total Rules	11	62	62	55	55
Average Accuracy	76.73%	95.24%	97.84%	92.28%	98.06%
Average Coverage	9.09%	5.75%	7.10%	5.62%	5.50%
Average Rule Length	3.36	4.52	3.98	4.62	4.27
Distinct Root Nodes	1	3	4	3	3
Distinct Class Values	2	2	3	2	3

Table 3: Comparison among J48, RF+HC and ForEx++ rules from the OASIS data set

From Table 3, we see that J48 rules have more average *Cov* than RF+HC and *ForEx++* rules and this is due to a low number of rules (11) generated by the J48 tree from OASIS. However,

average *Acc* of J48 rules is significantly lower than that of RF+HC and *ForEx++* rules. Besides, all J48 rules start with the same attribute (MMSE) and this may restrict their ability to explore different angles for knowledge discovery.

Furthermore, none of these J48 rules have *consequent* with the "Converted" class value (a minority class in OASIS) and thus provide no knowledge about those patients who developed dementia during the period of the data collection. RF+HC rules offer a major improvement over J48 rules in terms of average *Acc* with competitive average *Cov* for both decision forests. Also, a number of different root attributes (3) are present in RF+HC rules to facilitate broader knowledge discovery. However, similar to J48 rules, none of the RF+HC rules have *consequent* with the "Converted" class value.

On the other hand, *ForEx++* rules have the highest average *Acc* with competitive average *Cov* for both decision forests and they outperform RF+HC rules in every criteria with better average *Acc*, average *Cov* (for one out of two decision forests), average *Len*, Distinct Root Nodes and Distinct Class Values. More importantly, *ForEx++* rules can accommodate *consequent* with the "Converted" class value. As *ForEx++*rules from *Forest PA* WOBS are found to be the most accurate (see Table 3), we intend to go deeper into those rules in order to discuss some of the interesting trends that exist in them as follows.

Trend 1: We know, lower the Mini-Mental State Examination (MMSE) value, higher the chance of dementia. *ForEx++* rules show us that the impact of MMSE on dementia can be influenced by other attributes. For instance, when MMSE <= 26, 94% of the patients are found to be demented (Rule *Acc*: 94%, *Cov*: 23%). When nWBV is considered with MMSE (MMSE <= 26 and nWBV <= 0.75), *Acc* increases to 97% with 21% *Cov*. Similarly, "If SES = 2 and EDUC > 12 and MMSE <= 26 then Demented (*Acc*: 100%, *Cov*: 3%)". Besides, *ForEx++* nets an interesting rule with MMSE > 27 but having *consequent* "Demented". The rule is presented as follows.

- If MMSE > 27 and Gender = M and EDUC <= 15 and Age > 74 and eTIV <= 1498.9 then D (Rule *Acc*: 100%, *Cov*: 2%).

Based on the interesting rule, we now explore the OASIS data set using SQL in order to find some contra (adjacent) rules. The attributes in the *antecedent* of a contra rule remains the same as the original rule except some of the conditions are flipped. Some of the contra rules of the interesting rule are presented in Table 4 (Flipped conditions are shown in ***Bold+Italic***).

Ante	***Cons***	**Acc**	**Cov**
If MMSE > 27 and ***Gender = F*** and EDUC <= 15 and Age > 74 and eTIV <= 1498.9	ND	73%	11%
If MMSE > 27 and ***Gender = F*** and ***EDUC > 15*** and Age > 74 and eTIV <= 1498.9	ND	91%	9%

Table 4: Some of the Contra Rules

From the top rule of Table 4 we see that when Gender = F (instead of M), patients are not demented. This implies that females are less prone to dementia than their male counterparts when all other conditions remain the same. In line with this proposition we again apply SQL on the OASIS data set and find that "If Gender = F then ND (*Acc*: 63%, *Cov*: 58%)" and "If Gender = M then D (*Acc*: 51%, *Cov*: 42%)" which further validates the proposition. The bottom rule of Table 4 indicates that females with higher educational background are very less likely to be demented.

Trend 2: Similar to MMSE, with the decrease of normalized Whole-Brain Volume (nWBV) the chance of dementia increases. From the OASIS data set (Oasis), we see that when nWBV <= 0.73, patients generally get demented. However, with lower socio-economic status (SES = 3, 4 or 5) patients can be demented with nWBV as high as 0.78. A related rule extracted by *ForEx++* is presented below.

- If SES = 4 and Age <= 72 and nWBV <= 0.78 and eTIV > 1454.86 then D (*Acc*: 100%, *Cov*: 2%).

Based on the related rule, we explore the OASIS data set using SQL and present the percentage of demented patients with high and low SES for different nWBV in Figure 2. From Figure 2 we see that the impact of nWBV on dementia can be greatly exacerbated by SES. With lower nWBV and SES, patients are more prone to dementia. Besides, patients with higher SES are less likely to be demented compared to those with lower SES when nWBV remain the same.

Trend 3: With SES = 3, 4 or 5 (meaning lower socio-economic status) we find many rules having *consequent* "Demented". However, with SES = 1 (meaning the highest socio-economic status), patients can be demented in comparatively early ages. The related rule is:

- If nWBV <= 0.73 and SES = 1 and Age <= 70 then D (*Acc*: 100%, *Cov*: 3%).

To validate the proposition we apply SQL on the OASIS data set and find that the average age to be demented for patients with higher socio-economic status (SES = 1 or 2) is 75 years whereas the average age for demented patients with comparatively lower socio-economic status (SES = 3, 4 or 5) is 77 years.

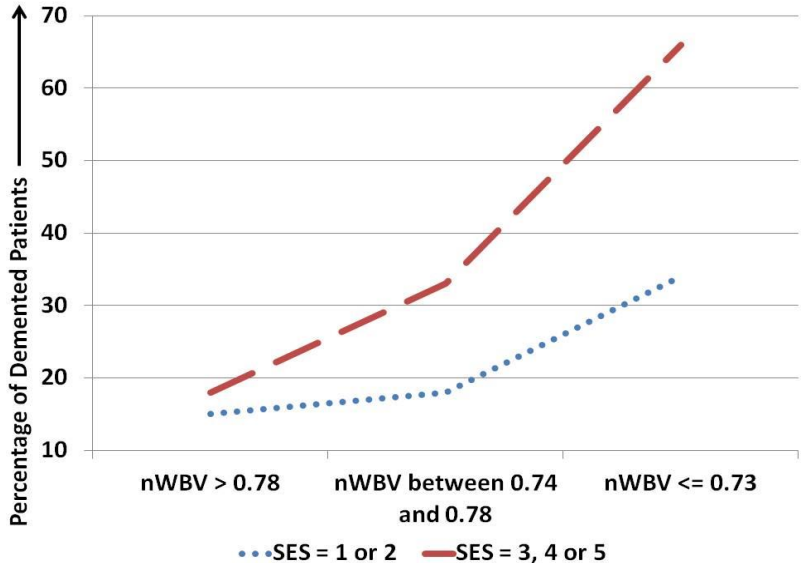

Figure 2: Percentage of Demented Patients for Different SES and nWBV

Trend 4: Similar to ***Trend 2***, with nWBV <= 0.73 patients generally get demented. However, when Education (EDUC) is high (i.e. > 15) and SES = 2 and ASF <= 1.24 and Age > 66, patients with nWBV <= 0.73 are found not demented (*Acc*: 100%, *Cov*: 5%). Thus, we reason that patients with higher educational background and socio-economic status have a lower chance to be demented than others. We then apply SQL on the OASIS data set and find the following rules (without having any dementia or MRI related attributes) to hold the same reasoning.

- If EDUC > 16 and SES in (1, 2) then ND (*Acc*: 67%, *Cov*: 24%).
- If EDUC <= 16 and SES in (3, 4, 5) then D (*Acc*: 48%, *Cov*: 45%).

Another similar rule highlighting the importance of EDUC is:

- If nWBV <= 0.73 and EDUC > 16 and ASF > 0.92 and Age between 66 and 78 then ND (*Acc*: 100%, *Cov*: 4%).

This rule indicates that even when the brain volume decreases (nWBV <= 0.73) patients may not be demented if they are highly educated (EDUC > 16). By applying SQL, we see that the average brain volume of non-demented patients with higher educational background (EDUC >16) is 0.73 and for comparatively lower educational background (EDUC <= 16) is 0.75, which verifies the indication.

From ***Trend 1*** and ***Trend 4***, it is understandable that educational background plays an important role on dementia in conjunction with Gender, SES and nWBV. We now explore the impact of different education levels on dementia independent of any other attributes and present them in Figure 3. Figure 3 clearly shows that higher the educational background, lower the chance of dementia.

Trend 5: Patients are found "Converted" when nWBV <= 0.72 even if they have the highest socio-economic status (SES = 1) (*Acc*: 100%, *Cov*: 1%). The related rule is: "If nWBV <= 0.72 and SES = 1 and Age > 70 and EDUC <= 16 then Converted. This indicates that with very small brain volume, patients can develop dementia very quickly. By applying SQL, we find that out of 37 converted patients, 22 of them have brain volume <= 0.72.

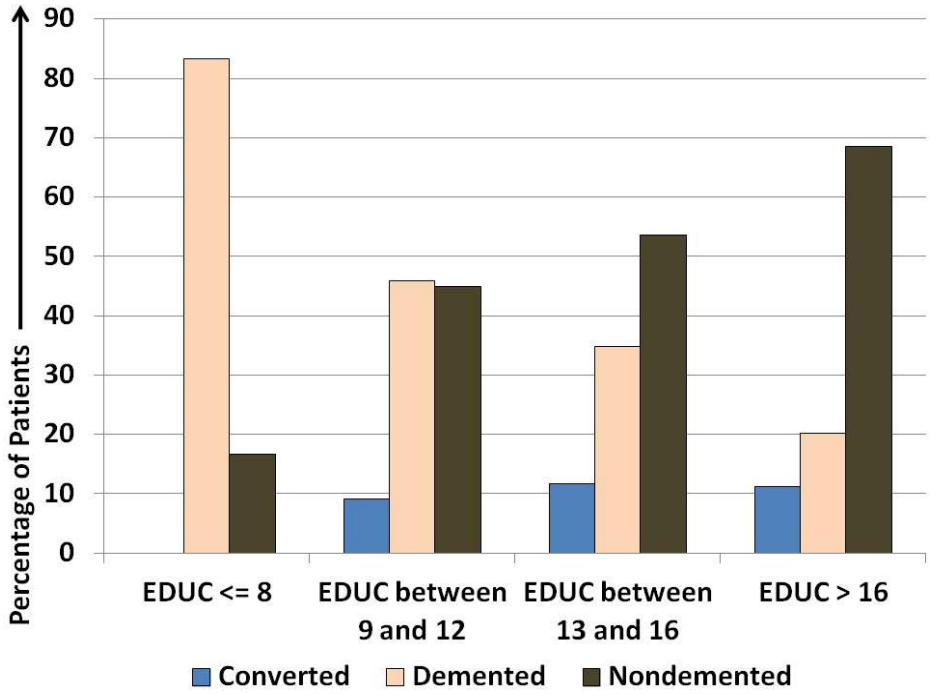

Figure 3: Percentage of Patients for Different EDUC

4.4 Heart Disease

Cardiovascular disease (or Heart disease) remains one of the biggest causes of deaths around the world. Each year about 30% of all deaths worldwide (15.6 million deaths) occur from heart diseases which is more than all communicable, neonatal, maternal and nutritional disorders combined and double the number of deaths caused by cancers (Nichols et al. (2014)). Moreover, the death from heart disease is predicted to rise approximately 23.4 million by 2030 with heart disease to remain as the leading cause of death for many years to come (Nichols et al. (2014)). Following global trends, heart disease is the first leading cause of death in Australia amounting 33.7% of all deaths (Shouman et al. (2011)). Motivated by these facts, we now facilitate exploring different aspects on heart disease from available data for better understanding the disease.

4.4.1 Data Set on Heart Disease

In this paper, we use a data set named "Statlog (Heart)" that consists of a collection of 270 observations (records) and all the observations are distributed among two class values namely Absence (A) or Presence (P) of heart disease (Lichman). In literature (El-Bialy et al. (2015)), the same data set was used for feature analysis of coronary artery diseases. The data set includes information about patient's Gender, Age, Chest Pain Type, Blood Pressure along with some medical test information. In Table 5, we present the attributes of the Statlog (Heart) data set and explain their meanings in detail.

Attributes	Explanation
Age	Ages of the patients vary between 29 to 77.
Gender	Male (M) or Female (F).
Chest Pain Type	1 for typical angina (angina is a medical term for chest pain), 2 for atypical angina, 3 for non-anginal pain, 4 for asymptomatic pain.
Resting Blood Pressure (systolic)	Resting blood pressure in mm Hg of patients when admitted to the hospital.
Serum Cholesterol	Serum cholesterol in mg/dl.
Fasting Blood Sugar	If fasting blood sugar > 120 mg/dl then 1 else 0.
Resting Electrocardiographic Results	0 for Normal, 1 for having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), 2 for showing probable or definite left ventricular hypertrophy by Estes' criteria.
Maximum Heart Rate Achieved	Maximum heart rate achieved during exercise.
Exercise Induced Angina	1 for yes, 0 for no.
Oldpeak	ST depression induced by exercise relative to rest.
Slope of the Peak Exercise ST Segment	1 for upsloping, 2 for flat, 3 for downsloping.
Major Vessels Colored	Vessels (0-3) colored by flourosopy.
Thal	A thallium stress test (in short, Thal) is a nuclear imaging test that depicts how well blood flows into the heart while a person is exercising or at rest (Nelson (2015)). 3 accounts for normal blood flow, 6 for fixed defect, 7 for reversible defect. A fixed defect refers to an area of the heart that does not get enough blood flow at rest or under stress. A reversible defect refers to an area of the heart that has good blood flow only at rest.
Distinct Class values	Two distinct class values. Absence (A) or Presence (P) of heart disease.

Table 5: Description of the Statlog (Heart) data set

4.5 Experimentation on the Statlog (Heart) Data Set

From experimental results, we see that a J48 tree generates only 19 rules from the Statlog (Heart) data set. Compared to that, a 20-tree RS generates as many as 555 rules, a cut point with accuracy >= 95% extracts 288 rules, *ForEx++* extracts as low as 59 rules. Consequently, 59 top-ranked rules are selected for the RF+HC method from RS. Similarly, a 20-tree *Forest PA* WOBS generates as many as 473 rules from the Statlog (Heart) data set, a cut point with accuracy >= 95% extracts 222 rules and *ForEx++* extracts as low as 40 rules and thus 40 top-ranked rules are selected for the RF+HC method. We now present a brief comparison among J48, RF+HC and *ForEx++* rules in Table 6.

Criteria	J48 rules	RS		*Forest PA* WOBS	
		RF+HC rules	*ForEx*++rules	RF+HC rules	*ForEx*++rules
Total Rules	19	59	59	40	40
Average Accuracy	84.39%	88.05%	94.71%	88.78%	95.07%
Average Coverage	5.26%	5.85%	13.11%	6.96%	16.37%
Average Rule Length	3.51	3.8	3.2	3.7	3.2
Distinct Root Nodes	1	6	7	5	6
Distinct Class Values	2	2	2	2	2

Table 6: Comparison among J48, RF+HC and ForEx++ rules from the Statlog (Heart) data set

From Table 6, we see that both RF+HC and *ForEx*++ rules outperform J48 rules in almost every criterion. Between RF+HC and *ForEx*++ rules, we see that *ForEx*++ rules prevail quantitatively with better average *Acc*, average *Cov*, average *Len* and Distinct Root Nodes even when Distinct Class Values remain the same. As was done before, we now discuss some of the trends that exist in *ForEx*++ rules from *Forest PA* WOBS as they are found to be the most accurate (see Table 6). The trends are discussed as follows.

Trend 6: All 9 rules with Thal = 3 have the *consequent* "Absence" (Average *Acc* and Average *Cov* of these rules are 95% and 24% respectively). This indicates that when there is normal blood flow in the heart, there is lower chance of having heart disease.

All 5 rules with Thal = 7 have the *consequent* "Presence" (Average *Acc*: 99%, Average *Cov*: 17%). This indicates that when the heart is not supplied with enough blood flow under stress, heart disease may occur. In this case, with Thal = 7 the heart has good blood flow at rest and thus the defect is reversible (meaning the heart muscles are still salvageable and a surgical procedure may prevent a future heart attack).

Trend 7: All 11 rules with Chest Pain Type = 2 or 3 have the *consequent* "Absence" (Average *Acc*: 92%, Average *Cov*: 16%). This means that patients with atypical or non-anginal pain have lower chance of having heart disease.

All 9 rules with Chest Pain Type = 4 have the *consequent* "Presence" (Average *Acc*: 96%, Average *Cov*: 15%). This means that patients with asymptomatic pain (who have had previous heart attack or diabetes) are especially at risk of developing a silent ischemia.

Trend 8: All 7 rules with Major Vessels Colored = 0 have the *consequent* "Absence" (Average *Acc*: 93%, Average *Cov*: 26%), which indicates that patients with no coronary arteries being blocked may not suffer from heart disease. All 12 rules with Major Vessels Colored = 1 or 2 have the *consequent* "Presence" (Average *Acc*: 98%, Average *Cov*: 8%), which indicates that if there are blockage or narrowing of some of the coronary arteries, there is a high chance of having heart disease.

Trend 9: All 4 rules with Age <= 59 have the *consequent* "Absence" (Average *Acc*: 95%, Average *Cov*: 24%). All 4 rules with Age between 59 and 63 have the *consequent* "Presence" (Average *Acc*: 98%, Average *Cov*: 9%). This indicates that higher the age, higher the chance of heart disease as ageing causes cardiac output to slow down and makes atherosclerosis (hardening and narrowing of the arteries) to deter enough blood circulation in the heart.

Trend 10: All 4 rules with Exercise Induced Angina = 0 have the *consequent* "Absence" (Average *Acc*: 94%, Average *Cov*: 21%). This indicates that if the heart gets enough blood flow under stress, there is lower chance of having heart disease (also see Trend 6).

Three out of 4 rules with Exercise Induced Angina = 1 have the *consequent* "Presence" (Average *Acc*: 95%, Average *Cov*: 12%). This indicates that if the heart does not get enough blood flow under stress then there is high chance of having heart disease (also see ***Trend 6***). The exceptional rule is:

- If Major Vessels Colored = 0 and Exercise Induced Angina = 1 and Oldpeak <= 1.5 and Resting Blood Pressure (systolic) <= 156 then A (*Acc*: 88%, *Cov*: 6%).

In the exceptional rule, we see that there is no blockage in the coronary arteries (Major Vessels Colored = 0) and thus the pain may not originate from the heart.

Trend 11: 2 out of 3 rules with Serum Cholesterol <= 304 have the *consequent* "Presence". The exceptional rule is:

- If Exercise Induced Angina = 0 and Major Vessels Colored = 0 and Gender = F and Serum Cholesterol <= 304 then A (*Acc*: 100%, *Cov*: 1%).

The exceptional rule indicates that females may withstand higher Serum Cholesterol level due to their hormonal protection. Based on the indication, we apply SQL on the Statlog (Heart) data set and find that the average Serum Cholesterol level for females with heart disease is 290 and without heart disease is 257.2. On the other hand, the average Serum Cholesterol level for males with heart disease is 249.8 and without heart disease is 233.7.

5 Conclusion

In this paper we propose a new, data set independent framework (*ForEx++*) for extracting a subset of decision forest rules that are comparatively more accurate, generalized and concise than others. We apply *ForEx++* on rules generated by Random Subspace and *Forest PA* WithOut Bootstrap Samples from two different publicly available medical related data sets on dementia and heart disease. We also apply a recent rule extraction method on rules generated from the same setup. Compared to J48 rules, rules extracted from *ForEx++* and the existing method are far more accurate and also can accommodate different angles for knowledge exploration. In head-to-head, *ForEx++* outperforms the recent method in almost all criteria. Also, the knowledge discovery potential of *ForEx++* is found to be impressive.

We have found some limitations of *ForEx++* that are described in the following.

1. Rules extracted by *ForEx++* may not be very diverse; meaning similar rules may be selected from each class.
2. Rules extracted by *ForEx++* may not cover the entire data set.

In future, we intend to solve these problems and also apply *ForEx++* on different types of data sets.

Acknowledgement: The OASIS data set (Oasis) is a contribution from the following grants: P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382 and R01 MH56584.

References

Adnan, M. N., & Islam, M. Z. (2014). "ComboSplit: Combining Various Splitting Criteria for Building a Single Decision Tree". *The International Conference on Artificial Intelligence and Pattern Recognition* (AIPR2014), 17-19 November, 2014 Kuala Lumpur, Malaysia. 1-8.

Adnan, M. N., & Islam, M. Z. (2015a). "Complement Random Forest". *The 13th Australasian Data Mining Conference*. Sydney, Australia.

Adnan, M. N., & Islam, M. Z. (2015b). "Improving the Random Forest Algorithm by Randomly Varying the Size of the Bootstrap Samples for Low Dimensional Data Sets". *23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, (ESANN 2015), 22-24 April, 2015 Bruges, Belgium. 391-396.

Adnan, M. N., & Islam, M. Z. (2015c). "One-Vs-All Binarization Technique in the Context of Random Forest". *23rd International Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, (ESANN 2015), 22-24 April, 2015 Bruges, Belgium. 391-396.

Adnan, M. N., & Islam, M. Z. (2016a). "Forest CERN: A New Decision Forest Building Technique". *20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part I*. Cham: Springer International Publishing.

Adnan, M. N., & Islam, M. Z. (2016b). "Knowledge discovery from a data set on dementia through decision forest". *The 14th Australasian Data Mining Conference*. Canberra, Australia.

Adnan, M. N., & Islam, M. Z. (2016c). "Optimizing the number of trees in a decision forest to discover a subforest with high ensemble accuracy using a genetic algorithm". Knowledge-Based Systems*,* 110**,** 86-97.

Adnan, M. N., & Islam, M. Z. (2017) "Forest PA: Constructing a Decision Forest by Penalizing Attributes used in Previous Trees". *Expert Systems with Applications*, Accepted.

Bernard, S., Adam, S., & Heutte, L. (2012). "Dynamic random forests". *Pattern Recognition Letters,* 33**,** 1580-1586.

Bernard, S., Heutte, L., & Adam, S (2008). "Forest-RK: A New Random Forest Induction Method". *ICIC (2)*. Springer.

Breiman, L. (1996). "Bagging predictors". *Machine Learning,* 24**,** 123-140.

Breiman, L. (2001). "Random forests". *Machine Learning,* 45**,** 5-32.

Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., & Snyder, A. Z. (2004). "A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume". *Neuroimage,* 23**,** 724-38.

Buckner, R. L., Snyder, A. Z., Shannon, B. J., LaRossa, G., Sachs, R., Fotenos, A. F., ... & Mintun, M. A. (2005). "Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory". *J Neurosci,* 25**,** 7709-17.

Dementia [Online]. Available: http://www.dementia.com/causes.html [Accessed 15 March, 2016 2016].

El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). "Feature analysis of coronary artery heart disease data sets". *Procedia Computer Science,* 65**,** 459-468.

Ertek, G., Tokdil, B., & Günaydın, İ. (2014). Risk Factors and Identifiers for Alzheimer's Disease: A Data Mining Analysis. *In:* PERNER, P. (ed.) *Advances in Data Mining. Applications and Theoretical Aspects: 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014. Proceedings*. Cham: Springer International Publishing.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). ""Mini-mental state": a practical method for grading the cognitive state of patients for the clinician". *Journal of psychiatric research,* 12**,** 189-98.

Geng, L., & Hamilton, H. J. (2006). "Interestingness measures for data mining: A survey". *ACM Computing Surveys,* 38**,** 9.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). "Extremely randomized trees". *Machine Learning,* 63**,** 3-42.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). "The WEKA data mining software: an update". *ACM SIGKDD explorations newsletter*, 11(1):10-18 doi:10.1145/1656274.1656278.

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers Inc.

Ho, T. K. (1998). "The random subspace method for constructing decision forests". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 20**,** 832-844.

Hollingshead, A. B. (1975). *Four-factor index of social status*. New Haven, CT: Yale University.

Huysmans, J., Baesens, B., & Vanthienen, J. (2006). "Using rule extraction to improve the comprehensibility of predictive models". *FETEW Research Report,* 1-55 doi:10.2139/ssrn.961358.

Johansson, U., Sönströd, C., & Löfström, T. (2011). "One tree to explain them all". *IEEE Congress of Evolutionary Computation (CEC)*, 5-8 June 2011 2011. 1444-1451.

Lichman, M. *UCI Machine Learning Repository* [Online]. Available: https://archive.ics.uci.edu/ml/index.html [Accessed 15 March, 2016 2016].

Liu, S., Patel, R. Y., Daga, P. R., Liu, H., Fu, G., Doerksen, R. J., & Wilkins, D. E. (2012). "Combined rule extraction and feature elimination in supervised classification". *IEEE transactions on nanobioscience*, 11(3):228-236 doi:10.1109/TNB.2012.2213264.

Lu, Z., Wu, X., Zhu, X., & Bongard, J. (2010). "Ensemble pruning via individual contribution ordering". *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. Washington, DC, USA.

Margineantu, D. D., & Dietterich, T. G. (1997). "Pruning Adaptive Boosting". *Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc.

Martens, D., Huysmans, J., Setiono, R., Vanthienen, J., & Baesens, B. (2008). "Rule extraction from support vector machines: an overview of issues and application in credit scoring". *Rule Extraction from Support Vector Machines*. Springer Berlin Heidelberg.

Martínez-Muñoz, G., Hernández-Lobato, D., & Suárez, A. (2009). "An analysis of ensemble pruning techniques based on ordered aggregation". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 31**,** 245-259.

Martınez-Munoz, G., & Suárez, A. (2004). "Aggregation ordering in bagging". *International Conference on Artificial Intelligence and Applications (IASTED 2004)*.

Martínez-Muñoz, G., & Suárez, A. (2010). "Out-of-bag estimation of the optimal sample size in bagging". *Pattern Recognition,* 43**,** 143-152.

Mashayekhi, M., & Gras, R. (2015). "Rule Extraction from Random Forest: the RF+HC Methods". *28th Canadian Conference on Artificial Intelligence, Canadian AI 2015, Halifax, Nova Scotia, Canada, June 2-5, 2015, Proceedings*. Cham: Springer International Publishing.

Morris, J. C. (1993). "The Clinical Dementia Rating (CDR): current version and scoring rules". *Neurology,* 43**,** 2412-4.

Murthy, S. K. (1998). "Automatic construction of decision trees from data: A multi-disciplinary survey". *Data mining and knowledge discovery*. 2, 345-389.

Nelson, J. (2015). *Thallium Stress Test* [Online]. Available: http://www.healthline.com/health/thallium-stress-test [Accessed 15 November, 2016 2016].

Nichols, M., Townsend, N., Scarborough, P., & Rayner, M. (2014). "Cardiovascular disease in Europe 2014: epidemiological update". *European Heart Journal,* 35**,** 2950-2959.

Polikar, R. (2006). "Ensemble based systems in decision making". *IEEE Circuits and systems magazine,* 6**,** 21-45.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*, Morgan Kaufmann Publishers Inc.

Quinlan, J. R. (1996a). "Bagging, boosting, and C4.5". *Proceedings of the thirteenth national conference on Artificial intelligence - Volume 1*. Portland, Oregon: AAAI Press.

Quinlan, J. R. (1996b). "Improved use of continuous attributes in C4.5". *Journal of artificial intelligence research*. 4, 77-90.

Ruta, D., & Gabrys, B. (2005). "Classifier selection for majority voting". *Information Fusion,* 6, 63-81.

Safavian, S. R., & Landgrebe, D. (1991). "A survey of decision tree classifier methodology". *IEEE transactions on systems, man, and cybernetics,* 21, 660-674.

Schmitz, G. P., Aldrich, C., & Gouws, F. S. (1999). "ANN-DT: an algorithm for extraction of decision trees from artificial neural networks". *IEEE Transactions on Neural Networks,* 10, 1392-1401.

Shouman, M., Turner, T., & Stocker, R. (2011). "Using decision tree for diagnosing heart disease patients". *Proceedings of the Ninth Australasian Data Mining Conference - Volume 121*. Ballarat, Australia: Australian Computer Society, Inc.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*, Addison-Wesley Longman Publishing Co., Inc.

What is Dementia? [Online]. Available: http://www.alz.org/what-is-dementia.asp [Accessed 15 March, 2016 2016].

What is OASIS? [Online]. Available: http://www.oasis-brains.org/ [Accessed 15 March, 2016 2016].