Scandent tree: a decision forest based classification method for multimodal incomplete datasets

by

Soheil Hor

B.Sc., Isfahan University of Technology, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Biomedical Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2016

Abstract

Incomplete and inconsistent datasets often pose difficulties in multimodal studies. A common scenario in such studies is where many of the samples are non-randomly missing a large portion of the most discriminative features. We introduce the novel concept of scandent decision trees to tackle this issue in the context of a decision forest classifier. Scandent trees are decision trees that optimally mimic the partitioning of the data determined by another decision tree, and crucially, use only a subset of the feature set. We use the forest resulting from ensembling these trees as a classification model. We test the proposed method on a real world example of the target scenario, a prostate cancer dataset with MRI and gene expression modalities. The dataset is imbalanced with many MRI only samples and few with MRI and gene expression. Using scandent trees, we train a classifier that benefits from the large number of MRI samples at training time, and of the presence of MRI and gene expression features at the time of testing. The results show that the diagnostic value of the proposed model in terms of detecting prostate cancer is improved compared to traditional methods of imputation and missing data removal.

The second major contribution of this work is the concept of tree-based feature maps in the decision forest paradigm. The tree-based feature maps enable us to train a classifier on a rich multimodal dataset, and use it to classify samples with only a subset of features of the training data. This has important clinical implications: one can benefit from an advanced modality to train a classifier, but use it in a practical situation when less expensive modalities are available. We use the proposed methodology to build a model trained on MRI and PET images of the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and then test it on cases with only MRI data. We show that our method is significantly more effective in staging of cognitive impairments compared to a model trained and tested on MRI only, or one that uses other kinds of feature transform applied to the MRI data.

Preface

The concept of scandent trees was originally introduced by the author in a paper published in Medical Imaging and Computer Assisted Interventions conference (MICCAI-2015) titled as "Scandent tree: a random forest learning method for incomplete multimodal datasets". This paper is the source of most of the material used in chapter 3.

Majority of the work discussed in chapter 4 is extracted from a paper conditionally accepted to the MICCAI special issue on Medical Image Analysis (MedIA) journal under the title of "Learning in data-limited multimodal scenarios: scandent decision forests and tree-based features".

The contribution of the author is development and evaluation of the techniques proposed in these publications and was performed under supervision of Dr. Mehdi Moradi.

This study has been performed as part of an ethics certificate approved by UBC research ethics board under the title of "Computational multimodal radiologic profiling as a prognostic bio-marker for individualized prostate cancer therapy" with UBC CREB number of H14-00359. Under supervision of Dr. Peter Black and Dr. Mehdi Moradi.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Funding from Canadian Institutes of Health Research (CIHR, Operating Grant) and Natural Sciences and Engineering Research Council of Canada (Discovery Grant) is acknowledged. Prostate imaging and genomic data were obtained at Vancouver General Hospital and UBC Hospital with approval from Clinical Research Ethics Board and informed patient consent. I would like to acknowledge Drs. Peter Black, Larry Goldenberg, Piotr Kozlowski, Jennifer Locke, Silvia Chang, Edward C. Jones, Ladan Fazli, all from UBC/VGH; Dr. Elai Davicioni, Christine Buerki, Heesun Shin, and Zaid Haddad from GenomeDx Biosciences Inc.

ADNI Disclosure: Data collection and sharing for parts of this work was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging

at the University of Southern California.

Dedication

First and foremost, I want to dedicate this thesis to my parents and my sisters for being the most supportive family one could possibly have. Thank you for supporting me with love and believing in me all through my life.
I also want to thank my supervisor, Dr. Mehdi Moradi for his kind supervision through this study. During the past two years, his kind comments, his guidance and support kept me motivated. Thank you for your trust and your patience.
Finally, I want to thank all my friends who made my life away from home more tolerable. Thank you for your continuous help and encouragement, when I needed it the most.

Chapter 1

Introduction

1.1 Motivation

In recent years there has been a surge of interest in multimodal data analysis. Different modalities provide researchers with complementary information about diseases and provide the means for more accurate detection and staging. This can be valuable in the case of progressive illnesses such as Alzheimer's disease and certain kinds of cancer. Simultaneous analysis of multiple modalities could also help us discover novel relations between different modalities, such as understanding the relationship of molecular changes caused by a disease and its imaging signature when both genetics and imaging data are available. Given these potential advantages, there has been a trend of merging different modalities in biomedical studies. For instance the Alzheimer's Disease Neuroimaging Initiative (ADNI), a six year $65 million study, has focused on using medical imaging modalities like Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) together with genetics and other clinical biomarkers for gaining better understanding of Alzheimer's Disease and its progression.

Acquiring multimodal data is generally more costly and time consuming than a single modality. As a result, multimodal datasets usually have valuable features, but a small set of samples with all features. This makes it difficult to build classifiers with large training data for highly multimodal protocols. For instance, in the case of the ADNI dataset, nearly half of the patients are missing the PET data. PET imaging is expensive and requires the use of radioactive tracers. As a result, a large number of patients only receive MRI scans, despite the fact that PET imaging provides unique brain functional information by quantification of the cerebral blood flow, metabolism, and receptor binding, which are not measured with MRI. This is a common scenario in dealing with multimodal data. So designing a computational model that can be trained on both MRI and PET data (multimodal data), but be deployed in clinical settings where only MRI (single modal data) is available, is a valuable contribution in this area, provided that the model outperforms one that is solely trained on MRI data.

Another common scenario is the case of a new multimodal research protocol including at least one component which is only obtained in the course of the study itself. An example of this scenario is our study to understand the relationship of molecular signature of prostate cancer with the imaging signature of the disease obtained through multiparametric MRI (mpMRI). The hope is that the simultaneous analysis of the molecular and imaging data can provide clues towards building a reliable and affordable clinical staging test. Since prostate cancer is a multifocal disease with tumors at different stages in each foci, this study requires tissue samples for molecular analysis that are obtained from a specific area with known spatial registration to the MRI images. The steps taken to acquire this data are not part of the clinical routine and the pace of data acquisition is slow. On the contrary, we have access to hundreds of samples with only mpMRI data and known histology. In this scenario, we are building a computational model that would be studied on the mpMRI+genomics (multimodal) data. Here, we may benefit from a computational framework that can utilize the rather large single modal dataset during training, but be able to handle the multimodal data at the testing stage.

In this thesis we present solutions, within the context of decision tree/forest paradigm of learning to address the problems posed in the two scenarios described above. Recent relevant work includes an investigation of the applications of imputation methods for dealing with missing values in the ADNI dataset [6]. The results show that joining a multimodal dataset with a single modal dataset by imputation of the missing values improves the classification accuracy, compared to training a classifier on either the single modal or the available multimodal data. In our current work, we intend to go beyond the paradigm of imputation. This is due to the fact that multimodal studies do not necessarily hold the usual assumptions in imputation that only a small number of data points are missing at random. We intend to deal with situations where blocks of data are missing together and the missing values are not spread randomly.

One trend in dealing with block wise missing values in multimodal datasets is separately modeling different blocks of data and then joining the resulting models by using a merging classifier or an ensembling method. One of the most successful attempts in this field is applying multi-source learning techniques for dealing with block wise missing data in ADNI [38, 39, 42]. The incomplete Multi-Source Feature learning method (iMSF) proposed by Yuan *et al.*, models different blocks of data with similar feature sets as different tasks and learns a joint model by imposing a sparse learning regularisation on these tasks [42]. The authors also propose a different approach by using

a model score completion scheme. This method is based on training independent classifiers on different blocks of data, and then using the prediction scores calculated by each classifier as a new presentation of the data that can then be imputed using conventional imputation techniques. A recent paper by Yu *et al.*, proposes a new method based on Multi-task Linear Programming Discriminant (MLPD) analysis [41]. This method formulates the problem as a multi-task learning scenario in a fashion similar to the iMSF method but does not constraint all of the tasks to share the same set of features, allowing joint learning of a more flexible model.

As a limitation to these studies, the training and testing datasets are assumed to have the same distribution and feature sets. Recently, Cheng *et al.*, addressed this issue and proposed a method for multimodal data analysis based on multimodal manifold-regularized transfer learning method [9]. This method enables using data from different domains together with unlabeled data for multimodal classification. This work uses a kernel based data fusion approach and includes a sparsity constraint in order to deal with the high dimensionality issue.

In this thesis we address the same limitations reported in [9], but with different assumptions that fit our scenarios. We don't assume that there is unlabeled data available. We do assume that the feature set of the test data is a subset of the training data. For instance, in case of the ADNI dataset, we assume that the training dataset consists of a set of samples with both MRI and PET data (although incomplete) but the test sample only consists of MRI data. This scenario is aimed at enabling the use of multimodal datasets for training of a classifiers that requires only a subset of modalities for testing.

Another important issue in multimodal classification is the high dimensionality that poses difficulties in feature selection and classifier building. The majority of the methods in the literature use the multi-kernel SVM framework for multimodal classification and need to impose sparse conditions on the multimodal feature set in order to avoid over-fitting [9, 20, 43].

However by working within the decision tree/forest paradigm we can benefit from its embedded way of dealing with high dimensional data through feature bagging [4]. Another motivation for the use of decision forest paradigm is that it provides the ability to morph the treatment of missing data within the framework of learning to maximize the classification performance. This area of work has seen significant contributions in recent years. These include the state of the art imputation methods embedded in the classification and regression tree (CART) algorithm and C5.0 algorithm for decision tree growth [28, 29] and in Random Forests (rfImpute) [4] which we discuss in

more detail in the next chapter.

1.2 Objective

The objective of this thesis is two fold:

First, developing a classifier in the context of decision forests that benefits from a large single modal dataset and a small multimodal dataset at the time of training, but is tested on multimodal data. This is motivated by the work in the area of prostate cancer detection and staging.

Second, developing a method based on the decision forest classifier that can benefit from a multimodal dataset together with a single modal dataset at the time of training, but is tested on single modal data. This is motivated by the work in the area of Alzheimer's Disease detection and staging.

1.3 Contributions

This thesis reports two specific contributions:

- *First*, introducing the concept of scandent trees, a novel forest-based method that can leverage one or more single modal datasets in order to enhance a multimodal forest. To our knowledge, this is the first decision-forest based algorithm specifically designed for this purpose. We provide results for different scenarios by simulation of the missing value problem on publicly available benchmark datasets. We also compare the scandent tree method to different state of the art methods for missing value imputation on a prostate cancer dataset which is a real-world example of the target scenario.

- *Second*, we develop the idea of scandent tree-based feature transforms to solve the problem of missing data in the single modal testing scenario. This problem has many clinical applications in areas where expensive research protocols meet the realities of clinical practice and high cost. Here, the assumption of a multimodal dataset with block wise missing values remains. However, there is no multimodal assumption about the test set. Using the proposed approach on the ADNI dataset, we show that we can use MRI and PET data for training a classifier that only requires the MRI data for the prediction of different stages of Alzheimer's disease. We show that the inclusion of the PET data at the time of training results in an improved classification

accuracy, even though the test cases are not subjected to PET imaging. We also examine the proposed method in different scenarios by simulation of the single modal and multimodal datasets on publicly available benchmark datasets. To our knowledge this is also the first method based on decision forests that has been designed specifically for this purpose.

1.4 Organization of the thesis

In this chapter, we discussed the advantages and limitations of multimodal studies and the importance of developing a new approach based on the decision forest classifiers leveraging multimodal datasets with block-wise missing data for either single modal or multimodal classification tasks. The remainder of this thesis is organised as follows:

- In Chapter 2 we explain the basics of the state of the art methods for growing decision trees and decision forests. Then we present a review of conventional methods for handling missing data, including general purpose imputation methods and the methods specific to tree-based classifiers.

- Chapter 3 describes the concept of scandent trees for multimodal classification and provides experimental and simulation results as proofs of concept.

- Chapter 4 introduces the tree-based feature transforms and applications of the scandent tree model for single modal classification. This chapter also provides simulation results together with experimental results.

- Finally, chapter 5 provides the conclusions of the thesis and a discussion about the limitations of this study and the potential for future work.

Chapter 2

Background

2.1 Introduction

In order to gain a better understanding of the missing value handling problem in tree-based classifiers, we first review tree-based classifiers. In the first section of this chapter, two of the most well known algorithms for growth of decision trees named, Classification And Regression Trees (CART) and C5.0 algorithms are introduced and explained in detail. Then an introduction to the concept of random forests and their theory of operation is presented. In a separate section, the challenges present in handling the problem of missing values in a general context are discussed in detail and a brief introduction to different general approaches to handle data with missing values is provided. In the next section of this chapter, we investigate the state of the art imputation methods which are specifically designed for tree-based classifiers. Detailed information about three of the state of the art embedded imputation methods for CART, C5.0 and random forests (rfImpute) is also in this section.

2.2 Decision trees and decision forests

2.2.1 Classification and regression decision trees

In this section the tree growth algorithm known as "Classification And Regression Trees" or in short, the CART algorithm will be explained. This algorithm was first introduced by Breiman*, et al.* in 1984 [5]. A CART tree is a binary tree that is grown based on an iterative process of finding the binary split point that gives the maximum purity gain at each node and using this division point to split each node to two child nodes. Let Y be the dependent variable or outcome class that can be ordinal categorical, nominal categorical or a continuous number. If Y is categorical with k classes, its class takes values in $C = 1, 2, \ldots, k$. Lets also define F as the set of features describing the data. Each feature (predictor) can also be ordinal categorical, nominal categorical or continuous. Assuming this notation, the

growing process of a CART tree can be explained as described below.

Tree growing process

As it was mentioned, the CART tree algorithm is an iterative process applied to each node of the tree starting from the root (the node with all of the available samples). The aim of this algorithm is to find the best split defined as the split in data that can result in the maximum purity in the child nodes. In the basic implementation of CART it is assumed that the splits are univariate. Meaning that each split only depends on the value of one feature and one feature only. If F is the set of all available features and f_i is a nominal categorical feature in F with k_i different categories, there exist $2^{k-1} - 1$ possible splits for this feature. If f_i is an ordinal categorical or continuous variable with k_i different values there are $k_i - 1$ different possible splits for f_i. The iterative growth algorithm of CART for each node can be simplified as:

- **Find each predictor's best split.** In case of continuous or ordinal categorical features, first sort the samples by the given feature. Then for each possible split from smallest to largest form the two child nodes. Given the child nodes one can calculate the splitting criterion or purity function for each split. The splitting criterion will be defined later. Choose the split with the highest purity. For each nominal categorical feature, examine all the possible subsets of the categories to find the purest split. For the sorted predictor, go through each value from top to examine each candidate split.
- **Find the node's best split.** Among the splits selected for each feature in previous step, select the one with maximum purity (the one that maximises the splitting criterion).
- **Split parent node to child nodes.** Use the best split found in previous step to divide the parent node to the child nodes.
- **Iterate.** Until the criteria for maximum depth of the tree has not reached, apply the same algorithm to each child node.

Splitting criteria and impurity measures

For any given node t, and a given split s, the best split is the split that maximizes the splitting criterion $\Delta i(s, t)$. Which corresponds to a decrease in impurity i.

For classification tasks (categorical Y), there are three splitting criteria defined for CART algorithm: Gini, Twoing, and ordered Twoing.

Let us define $P(t)$ and $P(j,t)$ as the probability that a sample belongs to node t and the probability that a sample in class j be in node t respectively. We can estimate these probabilities by :

$$P(j,t) = \frac{\pi(j)N_{w,j}(t)}{N_{w,j}}, \tag{2.1}$$

$$P(t) = \sum_j P(j,t), \tag{2.2}$$

where $\pi(j)$ is the prior probability of the outcome class j and by definition

$$P(j|t) = \frac{P(j,t)}{P(t)} = \frac{P(j,t)}{\sum_j P(j,t)}, \tag{2.3}$$

and

$$N_{w,j}(t) = \sum_{h(t)} w_n f_n I(y_n = j), \tag{2.4}$$

in which w_n and f_n are the case weight and frequency weight associated with sample n, $h(t)$ is the set of samples at node t and $I(j_1 = j_2)$ is the identifier function resulting in 1 when j_1 and j_2 are equal and is 0 otherwise.

Gini criterion

The Gini impurity measure at node t is defined as:

$$i(t) = \sum_{i,j} C(i|j)P(j|t)P(i|t), \tag{2.5}$$

in which $C(i|j)$ is the cost of classifying a sample to class i given that it belongs to class j assuming that $C(i|i)$ is equal to zero. Using the Gini impurity measure we can define one of the most well known splitting criterion's, the Gini decrease of impurity, defined as:

$$\Delta i = i(t) - P_L i(t_L) - P_R i(t_R). \tag{2.6}$$

in which P_L and P_R are the probabilities that a sample is sent to the left child node or the right child node respectively and can be defined as:

$$P_L = \frac{P(t_L)}{P(t)}, P_R = \frac{P(t_R)}{P(t)}. \tag{2.7}$$

It should be noted that if user specified costs are involved, altered priors can be used instead of the empirical estimations. In this case the altered prior can be defined as :

$$\pi^{'}(j) = \frac{C(j)\pi(j)}{\sum_j C(j)\pi(j)}, \tag{2.8}$$

in which

$$C(j) = \sum_i C(j|i). \tag{2.9}$$

Twoing criterion

This Criterion is actually a goodness measure not an impurity measure. So it should be maximised for each split. The Twoing criterion can be defined as:

$$\Delta i(s,t) = P_L P_R \Big[\sum_j |p(j|t_L) - P(j|t_R)| \Big]^2 \tag{2.10}$$

Ordered Twoing criterion

In the case of ordinal categorical outcome classes, the ordered Twoing criterion is the purity measure of choice. The algorithm to calculate this measure is as follows:

- First separate the class $C = \{1, \ldots, k\}$ of Y into two complementary super-classes C_1 and C_2 such that C_1 is of the form $C_1 = \{1, 2, \ldots, k_1\}$ in which, $k_1 \in (1, 2, \ldots, k-1)$.
- Using the purity measure $i(t) = p(C_1|t)P(C_2|t)$ find the split that maximises the Twoing criterion shown in equation 2.10 on C_1.
- Find the super class C_1 that results in best split (maximum gain in Twoing measure).

Continuous dependent variable

For continuous outcome variables (regression trees) a splitting criterion similar to the equation 2.6 can be used. However, the impurity measure used in this equation is different. A usual choice for the impurity measure is the Least Squares Deviation (LSD) measure which can be described as:

$$i(t) = \frac{\sum_{h(t)} w_n f_n (y_n - \overline{y}(t))^2}{\sum_{h(t)} w_n f_n}, \tag{2.11}$$

in which

$$P_L = \frac{N_w(t_L)}{N_W(t)}, P_R = \frac{N_w(t_R)}{N_W(t)}, \tag{2.12}$$

and

$$\overline{y}(t) = \frac{\sum_{h(t)} w_n f_n y_n}{N_W(t)}, \tag{2.13}$$

where

$$N_w(t) = \sum_{h(t)} w_n f_n. \tag{2.14}$$

Stopping rules

Stopping rules determine if the growth algorithm should continue on dividing each child node into two other nodes or should it set the final nodes as leaves. The stopping roles can be set based on the application. But the ones usually used in CART are as follows:

- If a node becomes completely pure, the tree growth stops. Meaning that if in classification trees, all cases in a node belong to the same outcome class, or in regression trees, all the cases in one node have the exact same number as the outcome variable, the node will not be split any further.
- If all samples in a node have the exact same values for all of the features, the node will not be split any further.
- If the tree depth reaches the predefined maximum limit set by the user, the tree growth will stop.
- If the sample size at a node is less than the minimum threshold set by the user the tree growth at this node will stop.
- If in case of splitting the parent node into child nodes, the child node sample size would be less than the minimum threshold set by the user, the parent node will not be split.

- If for the best split possible at node t, the splitting criterion ($\Delta I(t)$) is less than a user specified minimum purity gain, the node will not be split.

2.2.2 C5.0 decision trees

The C5.0 algorithm is based on the Iterative Dichotomiser 3 (ID3) algorithm first introduced by Ross Quinlan in 1986 [27]. Similar to the CART algorithm, this algorithm is based on iterative splitting of the sample space into smaller nodes. However, unlike the CART algorithm, the first versions of the ID3 algorithm only supported categorical features. Later improvements resulted in C4.5 algorithm which is the most well-known variation of the ID3 algorithm and in addition to supporting continuous variables, has several advantages over its ancestor, ID3. In this subsection we start with a brief explanation of the ID3 algorithm, then we introduce the C4.5 algorithm and finally the latest version of this family of algorithms, C5.0, is explained.

ID3 algorithm

Similar to the CART algorithm, the Iterative Dichotomiser 3 (ID3) algorithm grows a tree using a top-down greedy search through all the possible splits of training data for each feature at each node. However it uses information gain as the measure of goodness for each split. Information and entropy are measures in information theory that can directly be used as impurity measures for tree growth algorithms. In information theory, entropy of a sample set S that consists of c different classes can be defined as:

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2(p_i), \tag{2.15}$$

in which p_i is the probability that sample s belongs to class i and can be estimated by the proportion of samples in class i relative to the whole population. Given that in the above equation the logarithm function is in base 2, the unit for entropy is Bits. In this equation if the probability of a class is too small ($p_i \doteq 0$) or p_i is too large ($p_i \doteq 1$) the entropy measure becomes very small. So the entropy measure gives smaller values for pure subsets of samples. In other words, the more uniform the probability distribution between classes becomes, the larger the entropy measure. If we define information simply as "lack of entropy", information gain can be de-

fined as an splitting criteria that uses entropy as the impurity measure. The information gain can be defined as :

$$\Delta I(S,F) = Entropy(S) - \sum_{f_i \in F} \frac{S_{f_i}}{S} Entropy(S_{f_i}) \tag{2.16}$$

In which F is the set of all features of samples in S and the sum on f_i is over the all the values in F. Assuming this function as the splitting criteria, the algorithm to grow an ID3 tree is as follows:

- Given the samples in node t, calculate the splitting criteria (information gain) for all the features,
- Select the feature with the largest information gain (f_t) and the relative splitting point as the optimum division points,
- Use the optimum division points to split the samples at node t to two child nodes,
- If none of the stopping rules are true, continue growth of the tree on the child nodes using the remaining features.

The minimum set of stopping rules for ID3 algorithm are:

- If all the cases in one node are from the same class (entropy=0),
- If the Information gain is 0 or smaller than a pre-defined threshold,
- If the number of remaining features is 0.

It should be noted that other limits similar to the ones explained for the CART algorithm can be put on the number of samples at each node or the total depth of the tree in order to penalize over-fitting.

C4.5 algorithm

The C4.5 is an improved version of the ID3 algorithm with three major improvements:

- It can handle continuous features as well as categorical features. C4.5 extends the ID3 algorithm by putting a threshold on the continuous features and calculation of the Information gain based on the selected threshold,

- It has an embedded method for handling missing values (this method will be explained in detail later),
- It can incorporate user-defined weights for importance or cost of the features,
- It has an embedded method for pruning the tree.

More information about this method can be found in a recent paper by Quinlan [28].

C5.0 algorithm

C5.0 algorithm is the latest version of the tree growth algorithms of this family which has several advantages over the previous implementations, including:

- More computationally efficient implementation in comparison to C4.5 resulting in faster performance,
- A more memory efficient implementation,
- Results in smaller trees in comparison to previous versions which usually results in smaller probability of over fitting,
- Supports boosting. The C5.0 algorithm uses a process similar to adaboost [15]. In this process, first a conventional C5.0 tree is grown. Then the weights for each sample are calculated and subsequent iterations are used to build weighted trees and rule-sets. Then these rule-sets and trees are used to generate class probabilities. Finally the average of these class probabilities is reported as the final prediction,
- Supports using different weightings for samples and different misclassification measures,
- Supports winnowing, an embedded feature selection method that is particularly useful if the number of features are large but sample size is relatively small. This process is done as follows: First the samples are randomly split in half and one conventional C5.0 forest is grown using the first half of data. The effect of removing each feature on the performance of the first tree is determined using the other half of the data. If there is a feature that it's removal does not increase the total error rate, that feature will be removed from the set of features used for growth of the final tree.

2.2.3 Decision forests

Decision trees offer a fast and easy-to-interpret classifier for simple classifications tasks but generally result in weak classifiers that easily over-fit, especially if the sample size is small. An idea to increase the classification power of decision trees and avoiding over-fitting at the same time is to ensemble decision trees and form a decision forest. Ensembling is known as a very effective method that can improve the performance of any single classifier. In a similar way, the ensemble of decision trees as a forest is expected to always outperform a single decision tree. The idea of decision forests became popular after a paper by Ho*, et al.* in 1995. In [18] they show that if trees of a forest are trained on a set of randomly selected features, the accuracy of the forest grows by increasing the number of the trees of the forest. This observation that forests get more and more accurate as the model grows (gets more complex) is in direct contrast to the common belief that complexity of a classifier can only be increased to a certain point before it is reduced do to over-fitting. The key for this unique advantage of random forest over other classifiers is in randomising the basic classifiers to gain an ensemble of independent estimators of the outcome label. More detail on the importance of randomisation in random forests can be find in a paper by Kleinberg*, et al* [22].

The current state of the art decision forest method is based on the algorithm proposed by Breiman*, et al.* [4]. In this work Breiman uses two main tricks in order to ensure a fully randomised forest. First, randomly selecting a "bag" of samples for each tree, introduced for the first time by Breiman*, et al.* And second, randomly sampling a subset of features for each tree, introduced by Ho*, et al* in [19]. Breiman also introduced methods for calculating a feature importance measure based on a forest by calculating a distance measure between samples and calculation of error rates based on out of the bag samples. In the original implementation of random forest by Breiman, the decision trees were grown using an algorithm similar to the CART algorithm and the randomisation was introduced to each tree using the bagging and randomised feature selection methods. There exist different implementations of random forest that have put more emphasis on the randomisation of the base classifiers. For instance, an idea introduced by Dietterich*, et al.* in [12] is to ignore the optimum division step in CART algorithm and choose a random split for the data at each node of each tree. This results in a faster classifier but because it also results in weaker base trees it may limit the performance of the final forest.

Algorithm

Decision forests use the general technique of bagging or bootstrap aggregation with replacement to re-sample n_b number of bags. Each one with the same size as the original training set. For instance if the training set consists of n samples $S = \{s_1, s_2, \ldots, s_n\}$ described by k features $F = \{f_1, f_2, \ldots, f_k\}$ and an outcome class $C = \{c_1, c_2, \ldots, c_n\}$, each bag of samples will also consist of n samples randomly chosen (with replacement) from the same sample set. We hereby name these samples as S_b. These samples are described by the same set of features and corresponding outcome classes (C_b). After bagging samples, each bag is used to train a decision tree that is grown solely based on S_b and uses F to predict C_b. In the testing phase, the predictions of these trees can be merged either by majority voting or by averaging the probabilities using the following equation:

$$P(C) = \frac{1}{n_b} \sum_{b \in B} \hat{p}(S_b, C) \tag{2.17}$$

In which $P(C)$ is the probability of class C predicted by the whole forest and $\hat{p}(S_b)$ is the probability of class C predicted by the tree trained on bag b.

If grown too deep, each decision tree will over-fit to the training data and result in a low-bias but high-variance classifier. The boot strap and ensemble trick does not have any effect on the bias but reduces the variance of the final classifier as the number of the trees grow. However, this is true with the assumption that the resulting decision trees are uncorrelated. Otherwise the resulting trees will be very similar and averaging the similar trees does not have significant effect on the variance of the final classifier. The number of the bags and corresponding trees (n_b) is a parameter ranging from a few hundreds to several thousand trees depending on the nature of the dataset and the sample size and can be optimised by cross validation or by observing the out-of-bag error. The decrease in out of bag error seems to diminish after the number of the trees in the forest becomes large. The random forest growth algorithm also randomly selects a subset of features (for classification, usually $\sqrt{k}$, in which k is the total number of features) to grow each tree of the forest. This process helps in building a larger number of uncorrelated trees in the random forest.

Variable importance measure

Breiman also introduces a novel way to measure importance of each feature embedded in the random forest. The first step in order to measure importance of a feature is to grow a random forest using the training data available. Then the out-of-bag errors for each data point is calculated and averaged over the whole forest. Now for each feature f, we permute values of this feature from the whole data set (replace it with a randomly generated value in the same range). Then we re-calculate and average the out-of-bag error. The importance of each feature is the difference in the total out-of-bag error before and after permutation normalised by standard deviation of these differences. The features that result in larger differences are ranked as the most important features. This method has a few drawbacks, for instance this method is biased to give higher importance to categorical variables that have many different levels in comparison with features with fewer levels. A way to overcome this problem is to use methods like growing unbiased trees [30] or partial permutations [1].

2.3 Handling missing values

Missing data is a well known problem that if not handled correctly, can significantly affect the accuracy of any statistical inference performed on a dataset. This problem might be caused by human factors during the data acquisition stage. For instance patients that drop out of the study before the data acquisition is complete. Or it can be a natural possible state for the target variable. For instance the age of the spouses of patients in case of single patients. To decide how to handle missing data, it is helpful to first learn about the usual assumptions about missing data. The missing value scenarios can be divided into four general types or classes:

- **Missing completely at random.** A variable is missing completely at random if the probability of a data point being missing is the same for all samples. For example, if the decision that each patient should or should not undergo a specific clinical exam is taken by generating a random number or rolling a dice. If data is missing completely at random, then throwing out cases with missing data does not bias the statistical inference.

- **Missing at random.** Most of the time data points are not missing completely at random. For instance, the probability that a patient is required to undergo additional examinations might be taken based on

his preliminary clinical test results. As a general assumption in this scenario, it is assumed that the probability that a variable is missing depends only on the available information. Thus, if preliminary data for a patient consists of age, sex and race, then it is assumed that the probability that each patient will undergo more examinations is solely dependant on these fully available parameters. In this case it is reasonable to assume a model for this process. One example is assumption of a logistic regression model, where the outcome variable equals 1 for observed cases and 0 for missing. In this scenario, any data point can be removed from the study as long as it does not affect the assumed model for probability of the data-point being missing.

- **Missing that depends on missing parameters.** In this scenario not only the data is not missing at random, the probability that each data point is missing depends on the variables that are also missing. For instance, an example of this scenario is when the probability that a cancer patient is sent for an MRI scan is dependent on ultrasound pre-screening results that are not available at the analysis time. Another familiar example from medical studies is that if a particular experiment causes discomfort, a patient is more likely to drop out of the study. These data points are not missing at random (unless discomfort is measured and observed for all patients). If the data points are not missing at random, they must be explicitly modeled, otherwise adding bias to the statistical inference will be inevitable.

- **Missing that depends on the missing value itself.** This scenario makes handling missing values difficult not only because the data points are not missing at random, but also because the probability of missing a data point depends on the data point that might be missing. For instance in case of heart disease patients, if the blood pressure of the patients is saved only if it is outside of the normal range, we are dealing with missing values that are also determining the probability of them being missing. Another example is the case of censoring of data. For example in case of financial surveys, people with very high earnings may be less likely to report their actual salary. In the extreme case (for instance, if all participants earning more than $100k a year refuse to report their earning) a large part of the dataset will be missing and the probability of lack of the income variable depends on the income itself.

It is correct that when data is not missing at random specially when it depends on the missing values themselves, it is hard to compensate for the bias introduced into the inference algorithm. However, this bias can be mitigated by using the available variables. For instance, available features like age or sex or preliminary clinical data can be used in order to guess whether a patient's blood pressure would be out of the standard range and if it is higher than average or lower. Or in case of financial surveys, it can be assumed that if a person has higher education and is above a certain age he or she will have higher income. We then can use that information to compensate for relative bias in inference. These methods can not yield in a good estimation of the missing values but can certainly help in achieving a better model of the blanks in data.

It should be mentioned that it is in general impossible to prove whether data is missing at random or not. If it is, the process will be simple because we can model the probability of data being missing based on the available features, if the data is not missing at random we try to add as many related parameters as possible to the model so the missing at random assumption becomes reasonable. For instance, it is correct that the probability of missing the blood pressure is dependant on itself, but because the blood pressure and heart rate are indirectly related, adding the heart rate to the model can help in modeling the probability of missing of the blood pressure values.

This approach helps in reducing the bias caused by removal of the data points from a study. The next step will be to fill-in the missing data points or removing them in a way that has the least negative effect on the final classifier used on the final completed dataset. This approach is the basis of the imputation methods introduced in literature. We first investigate the general imputation methods that do not assume any specific model in data other than the missing at random assumption. We also introduce the regression based methods that treat each missing variable as a regression target, for instance for linear regression. These methods can model complex relationships in data but are still independent of the final classifier used on the imputed data so might not yield in the best classification performance. We also investigate imputation methods that are proposed specifically for our target classifiers: decision trees and forests.

2.3.1 Data removal methods

The simplest and maybe most common approach in dealing with missing data is simply ignoring the part of data that is missing one or more of the features. This can be done in two general ways, removing columns (features)

or rows (samples) from the data matrix.

Sample removal

This approach is most useful in cases when the number of samples with one or more missing values is small or a set of samples are missing a large set of features. This method may decrease the accuracy of the final classification because it reduces the total sample size but lets us train a more complex model because it preserves all of the features.

Feature removal

In contrast to the sample removal approach, the feature removal approach is usually selected when almost all of the samples are missing a set of specific features. This method decreases the accuracy of the final classifier by removing some of the potentially useful features but it is the simplest method that can preserve the sample size of the dataset.

2.3.2 General imputation methods

Zero

A simple approach to the imputation problem is to just replace all the missing values with a constant value: Zero. This approach originates from the natural assumption that in the absence of input signal in a sensor, the recorded value should be zero. This method sometimes significantly outperforms the data removal methods because it preserves both samples and features but obviously it introduces a bias towards smaller values.

Random guess

Another approach is replacing the missing values with a random guess in the acceptable range of the missing value. This approach matches a scenario in which a data acquisition system yields in white noise in the absence of input signal. This method also makes use of all data by filling in the missing values. Therefore, it may perform significantly better than data removal methods. As an advantage of this method, it guarantees that no unwanted correlation will be added between different features. In other words, the features that are statistically independent will stay statistically independent after the imputation. This is a necessity that the constant value replacement methods usually do not guarantee.

Replacement with mean

Another well known method that is frequently used for handling missing data in large datasets is replacement by the mean value. This approach originates from the statement that without any other knowledge, the average of the feature over the whole population is the best estimation of the real value and yields in the smallest average error. Assuming that the sampling population is balanced and the final classifier uses the deviation from mean as an error measure (which is the common scenario in regression models), this method is the simplest method that can introduce a minimum bias into the dataset. In case of a normalised dataset with mean value of 0 and standard deviation of one which is the usual scenario in many data analysis problems, this method is the same as the zero imputation method.

Replacement with median

For very large datasets, this method performs the same as the mean replacement method. However, in smaller datasets it results in a more robust solution which is less dependent on the outliers.

KNN imputation

The mean and median replacement methods might result in a reasonable estimation of the expected value for missing values over the whole population but do not necessarily give the best local estimation. The K-Nearest-Neighbors (KNN) method is based on replacement of missing values locally with mean or median of the neighbor data points selected via a distance function. Although KNN method seems like a very simple and naive approach, it has some advantages comparing to some advanced imputation methods. For instance, it can predict both discrete attributes (the most frequent value among the k nearest neighbours) and continuous attributes (the mean among the k nearest neighbours), so there is no necessity for creating a predictive model for each attribute that has missing data. But as a major drawback, it has to compute distance to all the samples in the dataset for each sample that needs to be imputed. In large datasets this becomes a major issue. Even with this drawback, the KNN method is used in many medical data analysis studies (for instance [3] and [21]) and many general data analysis applications as a simple and robust imputation method.

Regression based imputation methods

Another approach to the missing value problem which is specifically useful when the missing values are distributed only between a small set of features is using regression for imputation. These methods try to predict missing values of each feature using the other available features. These methods also make use of all the samples and all the features in order to model and predict missing values in each feature so they may be more effective than methods which only use local data for estimation of each missing value (like KNN). However, assuming a wrong model between features may have unexpected effects on the final classification. For instance, using a linear model for regression may introduce correlations between different features that did not previously exist in the data. The regression methods also can not guarantee the same error rate for all the predictions because missing values in each feature are predicted using a potentially very different set of samples depending on the distribution of missing values among features and samples. Although these weaknesses limit the power of regression methods in many applications, these methods are among the most popular methods used in literature especially in medical applications (for instance [8], [34], [2]).

2.4 State of the art tree-based imputation methods

The imputation methods explained in the previous section can be used with any type of classifier including decision trees or random forests. But it should be noted that the aim of imputation is not finding the best estimation of the missing values. It is finding a set of values that cause the minimum reduction in the accuracy of the final classifier. With this goal in mind, the imputation methods that are designed to work with a specific classifier, in our case the tree-based classifiers, are the best choice. The three most well known methods for this purpose are introduced in this section.

2.4.1 CART embedded imputation method: surrogate divisions

The missing data problem in CART algorithm can be divided into two smaller problems. First, finding the optimum division points and second, assigning samples to child nodes at each division. Let us re-examine the

information gain equation used for finding the optimum division point:

$$\Delta I = I(t) - p(t_L)I(t_L) - p(t_R)I(t_R) \tag{2.18}$$

In which I is the impurity function and P is the probability that a sample in the parent node belongs to the corresponding child node. Considering that $I(t)$ is calculated on the parent node, the missing features do not have any effect on this term of the equation above. However, the left and right child node impurities and the probability that a sample belongs to these nodes is affected by the missing features. In the CART algorithm for every given feature, tested for the impurity gain, these values are calculated using the samples that are not missing that particular feature.

Given that the division points are calculated using the explained method, the problem of assigning samples with missing features to each child node remains unsolved. The method used in the CART algorithm for this purpose is "surrogate divisions". This method proposes that for each optimal division which is found using the method explained above, we find a "surrogate" division that uses one of the other features to split the data in a similar way as the original optimal division. In other words, given a known division of the complete samples in the parent node, lets grow a tree of length one that can split the samples into the same child nodes using a different feature. Suppose there are n predictors $(x_1, x_2 \dots x_n)$ included in the CART analysis, lets assume that there are missing values only for one of the features. In this case, x_1 which happens to be the best predictor chosen to define the optimal split. The split necessarily defines two categories for x_1. This x_1 feature now actually becomes a binary response variable that splits the data into two classes, left and right nodes. Then a tree of depth one (a single split) is grown that uses $x_2 \dots x_n$ as potential splitting variables and x_1 as the response variable. The next step is to rank the $n-1$ possible predictors by the proportion of cases that are inevitably misclassified. The surrogate splits that do no better than the marginal distribution of the missing feature are ignored and removed from the list of surrogate divisions. The best split based on this ranking is then used to divide the samples with missing values into the child nodes. In other words, the class predicted by the surrogate split is then used in order to split the data similar to the original division when x_1 is not available. If a sample is missing the optimum division feature (x_1 in this example) we then use the best surrogate division instead. If the best surrogate division is also missing, we use the second best and so on. If none of the features are available, the sample is assigned to the child node that the majority of samples have been directed to. In the implementation

of the CART algorithm in R language (rpart package [33]) there are three ways to deal with missing data:

- Display only. Samples with missing values are completely ignored and are not passed to deeper nodes in the tree.
- Use surrogates. Split subjects with missing values according to the surrogate divisions, if all of the surrogates are missing, ignore the observation.
- The same as the second option, but if all surrogates are missing, assign the samples with missing value to the child node with the majority of complete samples.

In practice when a small portion of data is missing completely at random (MCAR) or missing at random (MAR), this method provides a robust and effective solution. However, if a large portion of the data is missing, consecutive surrogate divisions will be very likely to completely mis-guide the decision tree. Another issue that is very common in scenarios with very small sample size is skewed data. As mentioned in [17] this problem can be made worse by this imputation method. These weaknesses motivate us to examine other methods for imputation of the missing values, for instance the embedded method of C5.0 decision trees.

2.4.2 C5.0 embedded imputation method

C5.0 is the new version of the C4.5 algorithm which is one of the most well-known decision tree growth methods used today. Besides the basic differences between a C5.0 decision tree and a CART decision tree, the two algorithms basically use the same criteria for finding the optimum division points. However, unlike the CART algorithm that simply ignores the missing values in calculation of the information gain in equation 2.18, the C5.0 method uses a modified version of the information gain equation as it can be seen bellow.

$$\Delta I = \frac{N - N_0}{N} \Delta I(t - t_0). \tag{2.19}$$

In which N is the total sample size, N_0 is the number of samples that are missing the feature tested by the C5.0 algorithm and $\Delta I(t - t_0)$ is the impurity decrease assuming that only the samples that possess the respective feature are present. In simple words, the C5.0 algorithm calculates the impurity decease in the same fashion as the CART algorithm does, but

it assigns a smaller weight to the attributes that are missing from a large portion of data. As the second difference to the CART algorithm, C5.0 does not use surrogate variables in order to assign samples with missing values in parent node to the corresponding child nodes. Instead, the samples are fragmented into fractional cases and then assigned to the corresponding child nodes. For instance, if child node i has N_i samples, the missing samples in child node i will have weights equal to $N_i/(N-N0)$. These weights will then be used in order to weight the class probabilities in the leaves and calculate the probability of each class at each leaf.

The method used to handle missing data during the training phase of the C5.0 algorithm is straightforward. However, handling the missing values in the testing phase is a different story. Let P be the classification result of the test case Y using a C5.0 tree named T. There are three possible scenarios:

- If T is leaf (a tree of depth 0), P is found by the relative frequency of training cases that belong to the leaf.
- If T is a tree of depth one or more and all the features used as division points in T are available for Y, P is found by the relative frequency of training cases that belong to the same leaf as Y.
- Otherwise, all the possible outcomes of the decision tree (all the leaves that Y might belong to) are are explored and combined probabilistically, giving:

$$P = \sum_{i=1}^{k} \frac{N_i}{N - N_0} P_i, \tag{2.20}$$

  in which P_i is the probability of each class given that Y belongs to leaf i. N is the total training set sample size, N_i is the number of corresponding training samples at leaf i, N_0 is the number of training samples at leaf i with missing value (each N might be fractional).

When the probability P for each class of the outcome is calculated, the class with the largest probability is chosen as the classification result. The C5.0 algorithm has a more sophisticated approach for dealing with missing data problem. Nonetheless the approach is designed for a single decision tree. In comparison to a random forest, decision trees are very prone to over-fitting. In the next section, we investigate the embedded imputation method for random forests, rfImpute.

2.4.3 Decision forest embedded imputation method: rfImpute

There are two main methods for dealing with missing values embedded in the implementation of random forests by Breiman, *et al.* [4]. First is the rough-fix method. This method is a simple and naive approach which uses a technique similar to the median imputation explained in previous section to estimate the missing values for continuous variables and assigns the majority class for the categorical variables.

The state of the art embedded imputation method for random forests is rfImpute. This method starts with the rough-fix method as an initial estimation of the missing values, grows a random forest based on the imputed dataset, calculates the proximity matrix based on the resulting forest and then updates the missing values relative to the proximity matrix. This process is repeated for the new imputed values for a few iterations. In order to explain this method in more detail, let's first see how the proximity matrix is formed. The proximity matrix is an intrinsic measure of similarity between samples in a forest based on the number of times that two samples land in the same leaf in each tree of the forest. The values in the proximity matrix are calculated as follows:

Given that all the samples are run down each tree of the forest, for each tree in the forest add 1 to the proximity measure between i and j if they both land in the same leaf. Then divide the whole matrix by the number of the trees in the forest and set the main diagonal of the matrix to 1. This matrix is embedded within the random forest growth algorithm and as Breiman, *et al.*, mentioned in [4], the values $1 - prox(i, j)$ can be interpreted as squared distances in an Euclidean space of high dimension. Each row of this proximity matrix is then used in order to calculate an estimate of each missing value based on weighted average of the corresponding feature in other samples. For the categorical features, each class is weighted relative to the proximity measures and the most probable class is chosen. This process is usually repeated for 5 to 6 iterations. Although the method explained is iterative and therefore slow, it has proven to be an effective imputation method designed specifically for random forests.

2.5 Summary

In this section we provided a detailed review of the most common tree growth algorithms, an overview of different types of missing data, and some general purpose imputation methods. Then we focused on the imputation methods

designed specifically for tree-based classifiers like decision trees and random forests.

In the work presented in this thesis, we select the state of the art imputation method embedded in decision forests (rfImpute) as a natural choice for a baseline imputation method. In the multimodal test scenario of prostate cancer we compare the proposed method with two data-discarding methods in which we simply drop one or the other dataset (the single modal forest and the multimodal forest), two forest-based imputation methods (C5.0 forest and rfImpute) and two other general purpose imputation methods, namely replacing the missing values with zero, and replacing with the weighted average value of the K nearest neighbors (KNN) [3].

Chapter 3

Scandent tree: a forest based method for multimodal classification

3.1 Introduction

Missing data is a well known problem in data analysis and machine learning, however missing data handling in a multimodal study is more challenging because it might not hold some of the basic assumptions in usual missing data scenarios. For instance, the usual assumption in missing value handling is that only a subset of features are randomly missing from a subset of samples (typically 10% to 30% of data). However, in multimodal scenarios it is common that a large set of samples is missing a large set of features, often belonging to the same modality.

Incomplete multimodal datasets are common in biomedical experiments, where a usual scenario is to examine new protocols or new modalities and the relationship between them. For instance, joint analysis of medical imaging modalities and genetic biomarkers is an attractive subject for biomedical research. However, since clinical analysis of both imaging and genetic biomarkers is not common, it is hard for biomedical researchers to build large datasets of this type in a time and cost effective manner. This makes dealing with very valuable but very small datasets a common scenario.

On the other hand, some of the modalities used in a multimodal study might be part of the standard clinical procedures separately. For instance, although it is impractical and expensive to do MRI and genetics analysis on each and every patient that visits a clinic, the MRI data of a large number of patients is easily accessible from medical imaging archives of hospitals. So the missing value problem in multimodal scenarios can usually be formulated as merging a small valuable multimodal dataset with a large but single modal dataset for enhancement of a multimodal data analysis task.

In this chapter we intend to introduce a new method named “scandent

trees" that is based on decision forests and is specifically designed for this task. In this chapter first we explain the concept and implementation of the proposed method in detail then evaluate the proposed method in different sample sizes and feature set sizes by simulation of an incomplete multimodal scenario using publicly available datasets. Finally, we examine the proposed method on a real incomplete multimodal dataset, a joint analysis of MRI and genetics features for prostate cancer detection.

3.2 Method

3.2.1 Mathematical formulation

Let us assume that the training data consists of at least one single modal dataset defined as $S = (s_1, \ldots, s_{N_s})$ (in which each s_i is a single modal sample) and at least one multimodal dataset defined as $M = (m_1, \ldots, m_{N_m})$ (in which each m_i is a multimodal sample). These two datasets are described respectively by the single modal feature set F_s and the multimodal feature set F_m, where $F_s \subset F_m$. We do not set conditions on the feature or sample sizes but in practical scenarios, usually the multimodal dataset has fewer samples ($N_m < N_s$). Also the single modal set is missing some of the more discriminative features. In this section we aim to train a classifier using both S and M that can predict the outcome class C, for any test data described by F_m. In other words, we want to make use of a single-modal dataset for optimisation of a multimodal decision forest.

3.2.2 Intuition

Assuming the decision tree model for a classification task, two main sources of error can be imagined. First is the error caused by in-efficient partitioning of the sample space by the decision tree. And second, the error in estimation of the outcome class probabilities at each leaf. As an advantage of having all the important features, decision trees formed by the multimodal dataset are expected to partition the feature space very effectively. However, because of the low multimodal sample size, the estimation of the outcome probability at each leaf may not be accurate. The proposed method tries to reduce the prediction error at each leaf of the multimodal tree by using single modal samples that are likely to belong to the same leaf.

In order to find these single modal samples, a feature space partitioning algorithm is needed that can simulate the feature space division of the target multimodal tree on the single modal dataset. The proposed method is to

grow single modal trees (trees that only need the available feature set) that mimic the feature space division structure of the multimodal decision tree (a tree that needs all of the features for classification). Although it can not be guaranteed that such tree exists, we try to provide an estimation of the division boundaries of a multimodal tree by breaking it into smaller sub-trees and trying to estimate simple division boundaries of these trees using the set of available features.

This technique is expected to be more effective than the imputation methods because it eliminates the need to know the exact value of each missing feature and only relies on predicting the feature-space partition that each sample belongs to. This can be a much easier task.

By using such a technique we expect that we can use high-level relationships between modalities in order to merge datasets. For instance, we know that in usual scenarios the values of two different modalities like MRI and PET might not be predictable by each other. However, assuming that they are not statistically independent, the local trees might be able to avoid prediction of the exact values and instead translate the knowledge of the missing modality in form of given that a patient belongs to a partition of the PET feature space, what is the probability that the patient belongs to a partition in MRI feature space?. For instance, if a patient is similar to another patient in PET space, is it likely that these patients are also similar in MRI space ?. We are trying to show that sometimes, the answer to this question is enough for a more accurate classification and we wont need to predict exact values of a modality by the other one.

Growing a tree that follows the structure of another tree from the root to the top brings analogy to the behaviour of "scandent"trees in nature that climb a stronger "support" tree. Considering this analogy, the proposed method can be divided into three basic steps: First, division of the sample space by a multimodal decision tree, called "the support tree". Second, forming the single modal trees that mimic the structure of the support tree, called "the scandent trees". And third, leaf level inference of outcome label C, using the multimodal samples in each leaf and the single modal samples that are most likely to belong to the selected leaf.

3.2.3 Support tree

The first step in the proposed method is growing a decision tree to predict the outcome class based on the multimodal dataset. This tree can be one of the trees in a decision forest or an individual tree grown using any of the well known methods, such as C4.5 [28] and CART [33]. The method used

in this paper for growth of the support tree is based on the implementation of CART algorithm in the package "rpart" in R language [32].

Assuming that the tree is grown and optimized using the multi-modal dataset M, there are two steps that might be the source of classification error in the tree: Division of sample space at inner branches, and majority voting at the leaves. The sample space division requires sufficient sample size at each division point which becomes an issue as the tree gets deeper. However, ensembling within the forest paradigm compensates for occasional incorrect divisions at inner branches, leaving majority voting at the leaves as the critical step to get a precise estimation of probability of the class label. This error can be compensated for by the scandent trees.

3.2.4 Scandent trees

The second step is to form the scandent trees which enable the assignment of single modal samples to the leaves of the support tree. The process of feature space division in the support tree can be considered as grouping the multimodal data set M to different multimodal subsets. Let us define the subset of the samples of M in the i_{th} node as M_i and the feature used for sample space division at node i as f_i. For any arbitrary choice of node j, and it's immediate parent node i, we define node j as a 'link node' if f_i belongs to a different feature set from f_j, or if node j is either the root node or a leaf. In other words, node j is a link node if and only if :

Node j is the root node,

or

Node j is a leaf node,

or

$f_j \in F_s \quad$ and $\quad f_i \notin F_s$

or

$f_j \notin F_s \quad$ and $\quad f_i \in F_s$

Intuitively, the link nodes are the nodes that mark the roots and leaves of the largest sub-tree that uses only one modality for feature space partitioning. For each division node i in the set of the link nodes of the support tree, there exists a set of nearest child link nodes and child leaves $j_1, j_2, \ldots j_{ki}$. We define T_i as an optimum tree that can divide the set of multimodal samples at node i (M_i) to the set of multimodal samples at each child node (M_j) using the feature set F_s. The pseudo-code for forming a scandent tree is as follows:

For each link node i in the support tree,
{
For each sample n in M_i and each node j in set of
nearest child link nodes and child leaves of node i
{
if $n \in M_j$,
$C'_{i,n} = j$

}
Grow T_i, as optimum tree that for each sample n in M_i,
predicts $C'_{i,n}$ using only F_s.
}

The above algorithm forms local trees T_i for each node i that divide M_i to the child subsets M_j, using only the single modal features F_s. Here $C^{'}$ is a new categorical label-set defined for the corresponding local tree. For each sample in the parent node, the $C^{'}$ is assigned in a way that the samples belonging to a specific child node j are mapped to the same category within $C^{'}$.

For each node i, if $f_i \in F_s$, then T_i is expected to divide M_i to the child subsets (M_j) with perfect accuracy. But if $f_i \notin F_s$, then T_i will be optimized to form the smallest tree that can divide the sample space in a similar manner to the support tree. Using T_i's for feature space division at each node, we can form a new tree that consists of the same link nodes as the support tree but only uses features of a single modal (F_s) for sample space division, we name this single modal tree, a scandent tree. Since T_i's are single modal trees, they can be used to predict the probability that each single modal sample s belongs to link node j, calculated by:

$$p(s \in Node_j) = p(s \in Node_j | s \in Node_i) p(s \in Node_i)$$

in which $Node_i$ is the parent link node of $Node_j$, the term $p(s \in Node_j | s \in Node_i)$ is estimated by the corresponding sub-tree T_i and $p(s \in Node_i)$ is calculated by recursion.

This method is expected to be generally more accurate than direct estimation of the leaves by other single modal classifiers. Because the scandent tree only has to predict the division boundaries for features that do not belong in F_s and other divisions will be perfectly accurate.

Given the small multimodal sample size, the local trees could be prone to over-fitting if only the few samples in the corresponding link nodes are

used for training T_i's. We overcome this problem by using all of the available multimodal samples (M) for training of each local tree by running the whole multimodal training set through the corresponding sub-tree of the support tree. This will give each sample in the multimodal dataset a label from the set $C^{'}$. This method adds more multimodal samples to the parent link node (M_i) and each child link node (M_j). This results in better estimation of T_i. We found that using this trick adds to the robustness of the scandent tree method. Figure 3.1 shows a simple diagram of the proposed method.

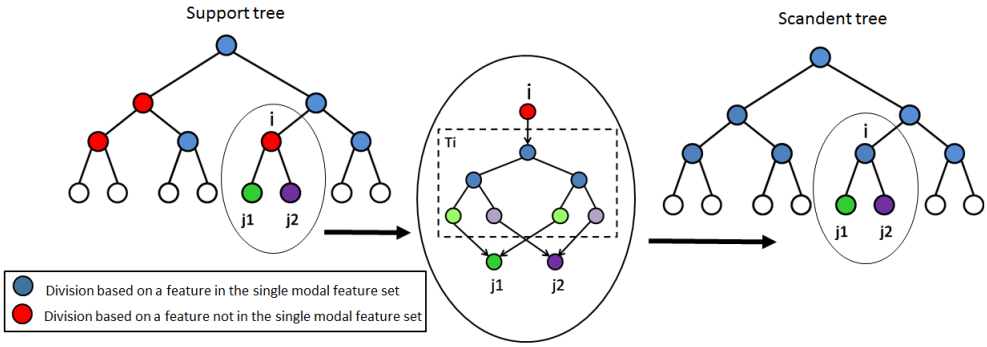


Figure 3.1: Diagram of the proposed method for growing the scandent trees

3.2.5 Leaf level inference

The standard method for leaf level inference is majority voting. However, if there are a large number of single modal samples misplaced by the scandent tree, they might flood the original multimodal samples. The proposed method is weighted majority voting by non-uniform re-sampling from each leaf i and then calculating the probability of outcome C using the resampled data. We define the re-sampling probability of each sample x in leaf i as:

$$p(x)_{re-sample,i} = \begin{cases} 1/\mathrm{N}, & x \in M_i \\ p(x \in Leaf_i)/N, & x \notin M_i \quad \& \\ & p(x \in Leaf_i) > q \\ 0, & x \notin M_i \quad \& \\ & p(x \in Leaf_i) < q \end{cases}$$

In which q is the selected minimum threshold for the probability that

a single modal sample belongs to the selected leaf i, and N is the total number of samples in leaf i (single modal and multimodal). As q value increases, the probability that a misplaced sample is used in the leaf level inference is reduced. This may increase the accuracy of the majority voting but increasing q will also reduce the number of single modal samples at each leaf resulting in low precision of the probability estimation. This trade-off is more evident at the two ends of the spectrum, for $q = 1$ the tree will be the same as the support tree which suffers from low sample size at the leaves. For $q = 0$ all the single modal samples will be used for inference at each leaf.

The optimization of the q parameter for each leaf is essential for optimal performance of the resulting tree. This can be done by cross validation over the multimodal dataset, using out of the bag samples in case of a decision forest. Using non-uniform re-sampling instead of majority voting ensures that the single modal samples at the leaves are randomized. This randomization is critical because the single modal samples are not randomly selected in the scandent tree growth algorithm and without re-sampling, there is a possibility that many of the scandent trees in the resulting forest are not independent. This would violate one of the basic requirements of tree ensembling in a decision forest.

Although the proposed algorithm is explained only for one single modal dataset, the same method can be applied on different single modal datasets using the same support tree. As a result, the proposed framework can be used flexibly when different subsets of features are missing.

3.2.6 Implementation

For building the support trees, we randomly bagged two-third of the multimodal samples and randomly selected the square root of the dimension of the multimodal feature set as the feature bag. This bootstrapping and bagging phase is done separately for each of the outcome classes to ensure balanced class labels. Then the scandent trees are formed and for each leaf of each support tree in the forest the q parameter is optimised using the corresponding out of the bag samples.

After growing and optimizing each of the trees, the probability of outcome class C is calculated by averaging the corresponding probabilities of all trees in the forest. We use the R package "rpart" [32] both for growing each support tree and each of the local single modal trees (T_i's). This package uses internal cross validation to form the optimal tree. But for the purpose of controlling the bias-variance of the resulting forest, the depth of support tree is limited by controlling the minimum of samples needed for each divi-

sion. The depth of T_i's in each scandent tree is optimized by internal cross validation.

3.3 Evaluation

We simulate the missing data scenario using three publicly available datasets. First is a dermatology dataset which is multimodal in nature, therefore we only need to simulate the size of the single modal and multimodal datasets. Second is a heart disease dataset which is formed by subsets that are partially overlapping in terms of features. However, it does not match the definition of a multimodal dataset in the sense that the missing features do not belong to separate modalities. For this dataset we simulate modalities by intentional feature removal from one of the subsets. The third dataset is a breast cancer dataset which is neither multimodal nor multi-source, so we should simulate both the modalities and the single modal and multimodal sub-sets by random sampling and feature removal. A brief description of each dataset and evaluation method is presented below.

3.3.1 Evaluation using benchmark datasets

Dermatology dataset

In order to test the performance of the proposed method for different sample sizes we simulate an incomplete multimodal dataset by discarding a set of features from a complete multimodal dataset. Because the dermatology dataset is a complete multimodal dataset, we have the opportunity to simulate the target incomplete multimodal scenario by intentionally removing one of the modalities from a set of randomly selected samples.

This set is a natural example of a multimodal dataset publicly available through the University of California Irvine (UCI) database [23]. This dataset consists of two distinct feature sets, an easily accessible feature set obtained during clinic visit of each patient (for instance, age of each patient) and a harder to access feature set that is acquired by further histopathological tests in a laboratory (eg. Melanin incontinence observed in skin samples). The total sample size of this dataset is 357, the clinical feature set size is 12, while the size of the histopathological feature set obtained in the laboratory is 22. The outcome class is the diagnosis of one of six dermatology diseases.

For this simulation we examine the classification task of one disease class (Seborrheic Dermatitis) vs other classes. We examine the performance of the proposed method for different multimodal sample sizes by assuming that the

histopathological features are missing from a subset of samples. We change the size of the multimodal sub-set (the set with no missing features) and report Area Under Curve (AUC) as a function of the multimodal sample size in comparison to the state of the art imputation method for random forests (rfImpute). A list of clinical and histopathological features used in this study is shown in the Table 3.1. The outcome class is selected from the dermatology diseases shown in the same table.

Table 3.1: List of features and the outcome classes, dermatology dataset

Histopathological features	clinical features	dermatology diseases
Melanin incontinence	Erythema	Psoriasis
Eosinophils in the infiltrate	Scaling	Seboreic dermatitis
PNL infiltrate	Definite borders	Lichen planus
Fibrosis of the papillary dermis	Itching	Pityriasis rosea
Exocytosis	Koebner phenomenon	Cronic dermatitis
Acanthosis	Polygonal papules	Pityriasis rubra pilaris
Hyperkeratosis	Follicular papules	
Parakeratosis	Oral mucosal involvement	
Clubbing of the rete ridges	Knee and elbow involvement	
Elongation of the rete ridges	Scalp involvement	
Thinning of the suprapapillary epidermis	Family history	
Spongiform pustule	Age	
Munro microabcess		
Focal hypergranulosis		
Disappearance of the granular layer		
Vacuolisation and damage of basal layer		
Spongiosis		
Saw-tooth appearance of retes		
Follicular horn plug		
Perifollicular parakeratosis		
Inflammatory monoluclear inflitrate		
Band-like infiltrate		

Heart disease dataset

This set consists of data from two different studies reported in [11]. This dataset is a natural example of a complete dataset accompanied by a similar large dataset with non-random missing features. One set (data from the Hungarian Institute of Cardiology) is missing two out of 14 features. We use this as the single modal dataset in our experiments while the complete set (the Cleveland dataset) is used as the multimodal dataset. In real world problems, such as our prostate cancer study, the single modal dataset is missing some of the most discriminative features. To simulate this condition we used a classical random forest feature ranking approach. Moreover we

study the effect of decreasing the number of features in the single modal dataset on the overall performance by sweeping from 12 to two features, always removing the most discriminative ones. The multimodal dataset in this experiment (the Cleveland dataset) consists of 303 samples. 100 samples were randomly separated and used as the test data. The remaining samples were used as the multimodal data for training the support trees. We experimented with scenarios that included 10% to 90% of this data in training of the support trees. More information about the features used in this experiment can be found in Table 3.2

Table 3.2: The feature set of the heart disease dataset

Features/Variables	Explanation
age	Age in years
sex	sex (1 = male; 0 = female)
cp	chest pain type: Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	resting electrocardiographic results: Value 0: normal Value 1: having ST-T wave abnormality Value 2: showing left ventricular hypertrophy by Estes' criteria
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment Value 1: upsloping Value 2: flat Value 3: downsloping
ca	Number of major vessels (0-3) colored by flourocsopy
thal	3 = normal; 6 = fixed defect; 7 = reversible defect
outcome class	Diagnosis of heart disease (Angiographic disease status) Value 0: < 50% diameter narrowing Value 1: > 50% diameter narrowing

Breast cancer dataset

This is a complete set with 569 samples [37]. This dataset consists of 30 features describing the nucleus properties of a breast cancer cell. The scenario of multimodal and single modal datasets was simulated with sampling.

We change the size of the single modal feature set and report the AUC as a function of the number of single modal features. In a fashion similar to the heart disease experiment, we simulate the feature quality disadvantage of the single modal dataset by removing the top ranked features of this dataset. More information about this feature set can be found in Table 3.3

Table 3.3: The feature set of the breast cancer dataset

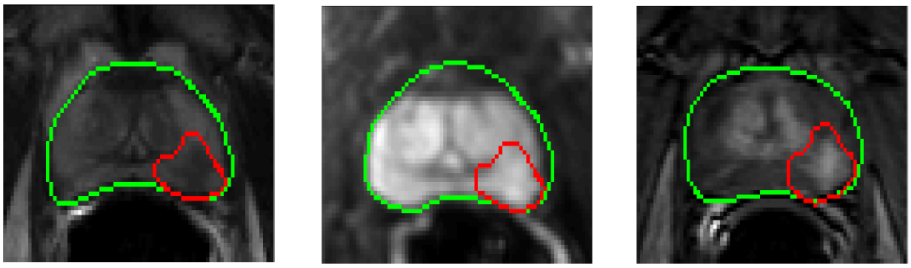
Features/variables	Explanation
radius	Mean of distances from center to points on the perimeter
texture	Standard deviation of gray-scale values
perimeter	
area	
smoothness	Local variation in radius lengths
compactness	$perimeter^2/area$ - 1
concavity	Severity of concave portions of the contour
concave points	Number of concave portions of the contour
symmetry	
fractal dimension	"coastline approximation" - 1
outcome class	Cancer diagnosis:
	B:Benign
	M:Malignant

3.3.2 A real scenario: prostate cancer dataset

This set consists of a small genomics+MRI prostate cancer dataset ($N_m = 27$) accompanied by a relatively large MRI only dataset ($N_s = 428$). The single modal dataset consists of five multi-parametric MRI features from dynamic contrast enhanced (DCE) MRI and diffusion MRI on a 3 Tesla scanner. We used the apparent diffusion coefficient (ADC) and fractional anisotropy (FA) from diffusion MRI, and three pharmacokinetic parameters from DCE MRI: volume transfer constant, k^{trans}, fractional volume of extravascular extracellular space, v_e, and fractional plasma volume v_p [16, 25].

This data is from patients undergoing radical prostatectomy at Vancouver General Hospital and has been collected with informed consent, and with the approval of the Research Ethics Board of the Vancouver General Hospital. Imaging is performed a week before the surgery. After the surgery, the prostate specimens were processed with wholemount cuts that matched the slices in the MRI scans. A cutting device and the procedure described in [13] ensured that the cuts matched the MRI slices. An experienced pathologist outlined the area of the tumor/normal from wholemount histopathology slides ([16, 25]).

The histopathology slide of each patient was then registered to a T2-weighted image and corresponding DTI and DCE-MRI slices ([16, 24, 25]). This registration lets us find the Region Of Interest (ROI) in the DTI and DCE MRI slices that co-respond to the tumor(please see Figure 3.2). We took the average of the quantitative MRI features (apparent diffusion coefficient, ADC, fractional anisotropy, FA, volume transfer constant, k^{trans}, fractional volume of extravascular extracellular space, v_e, and fractional plasma volume, v_p) as the MRI features for each ROI.



(a) (b) (c)

Figure 3.2: Registration example: (a) T2-weighted , (b) DTI and (c) DCE-MRI slice. The green contour represents the boundaries of the prostate gland. The red contour represents the mapped tumor ROI([16, 25]).

The tissue samples were then obtained by needle biopsy from the corresponding formalin-fixed paraffin-embedded (FFPE) tissue blocks and RNA was extracted and purified from these samples. The expression level of 39 genes that form the most recent consensus on the genetic signature of prostate cancer for patients with European ancestry as reported and maintained by National Institute of Health [26] were used as features (please see Table 3.4). Each of the selected genes were mapped to the closest probe location on an Affimetrix Exon micro-array and the gene expression at each of these locations were selected as a genetic feature in our study. Figure 3.3 shows a heat-map of the gene expression data for all the patients in our study. The dendrogram shown in this figure shows clustering of samples and genes based on the gene expression.

We have 27 samples with gene expression data and registered imaging data (14 normal, 13 cancer) from 21 patients. The evaluation of the proposed method on this small dataset was carried out in a leave one out scheme. Each time, the support trees were trained using 26 samples, with all the single modal data samples and features used for forming the scandent trees.

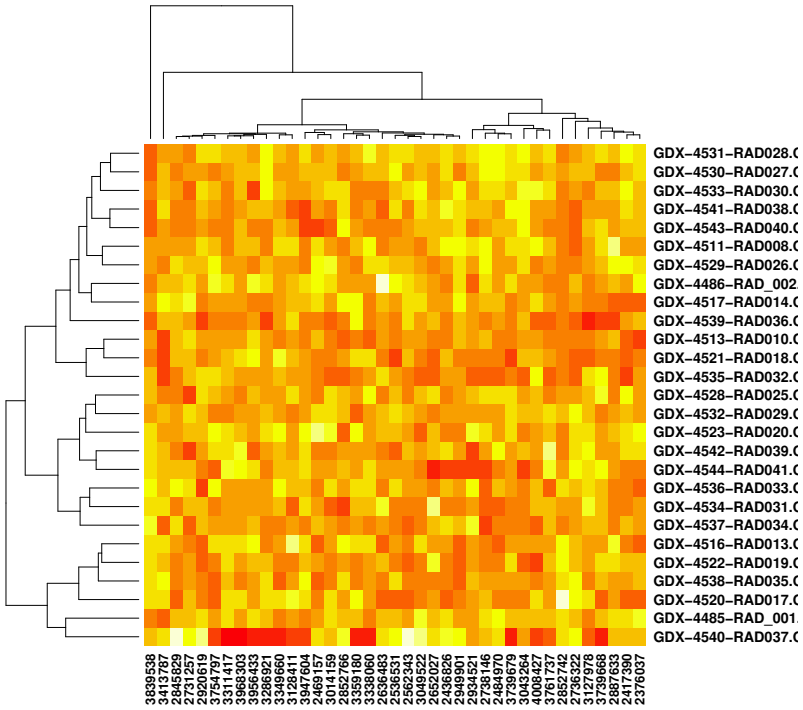


Figure 3.3: Gene expression heat-map of the probes corresponding to the selected genes for each patient. Each row presents a sample. Each column presents a gene expression feature. The vertical dendograms show clustering of samples. The horizontal dendograms show clustering of features. Sample clustering correctly clusters each patient. This shows that the gene expression profiles are mostly patient-specific. Although all of the selected genes are known to be biomarkers of prostate cancer, neither correlations between features nor cancer-related patterns are visible.

Table 3.4: List of the genes used in the prostate cancer study

Probe ID	Gene name	Probe ID	Gene name
2376037	MDM4	3947604	BIK
4008427	NUDT11	3956433	CHEK2
2436826	KCNN3	2887633	BOD1
2562343	GGCX	3968303	SHROOM2
3128411	EBF2	2920619	ARMC2
3761737	ZNF652	3286921	08-Mar
3754797	HNF1B	2934521	SLC22A3
2852766	AMACR	2852742	AMACR
2731257	AFM	2652027	CLDN11
2736322	PDLIM5	2949901	NOTCH4
3127978	NKX3-1	3043264	JAZF1
2484970	EHBP1	3349660	HTR3B
2845829	TERT	3359180	TH
3739668	VPS53	3739679	VPS53
2738146	TET2	3014159	LMTK2
2536531	FARP2	3338060	MYEOV
3839538	KLK3	3049522	TNS3
2417390	CTBP2	2469157	GRHL1
3311417	CTBP2	2636483	SIDT1
3413787	TUBA1C		

3.4 Simulation and experimental results

In this section we provide simulation and experimental results to verify the performance of the method explained in the previous sections. The experimental results are on a prostate cancer dataset which is an example of a real multimodal incomplete dataset which consists of a small but multimodal dataset with genetics and MRI and a larger single modal dataset with only MRI. This dataset is a perfect example of our target scenario but because of the small sample size of the multimodal dataset, it is hard to verify that the proposed method outperforms the state of the art in imputation of random forests.

In order to evaluate the performance of the proposed method in different scenarios we also used three publicly available datasets. We simulated the missing value scenario for different sample sizes of the multimodal dataset and different feature set sizes of the single modal dataset. These datasets are real biomedical datasets and some are even multimodal in nature (for instance the dermatology dataset). However, we call the results on these datasets, "simulation results" because these datasets are complete and the missing values are intentionally removed from each dataset to simulate different missing value scenarios.

3.4.1 Simulation results

Dermatology dataset

Figure 3.4 shows the AUC of the proposed method and the state of the art embedded imputation method of random forests(rfImpute) in different multimodal sample sizes. It can be seen that the proposed method outperforms the rfimpute method especially in smaller samples sizes. For instance when the sample size is as small as 51, simulations on this dataset resulted in average AUC of 0.97 for the proposed method and AUC of 0.92 for the rfImpute method. This is because in smaller samples sizes the imputation method is forced to predict a large number of missing values using a very limited set of available samples and this bias introduced by the imputation in the training set may mis-guide the random forest classifier. The scandent tree method enhances the performance of the random forest leaf by leaf, using cross validation instead of trying to predict the missing values. Therefore, it is expected to be less vulnerable to mis-classifications.

Because the dermatology dataset is multimodal in nature, we can not examine the performance of the proposed method when a larger number of

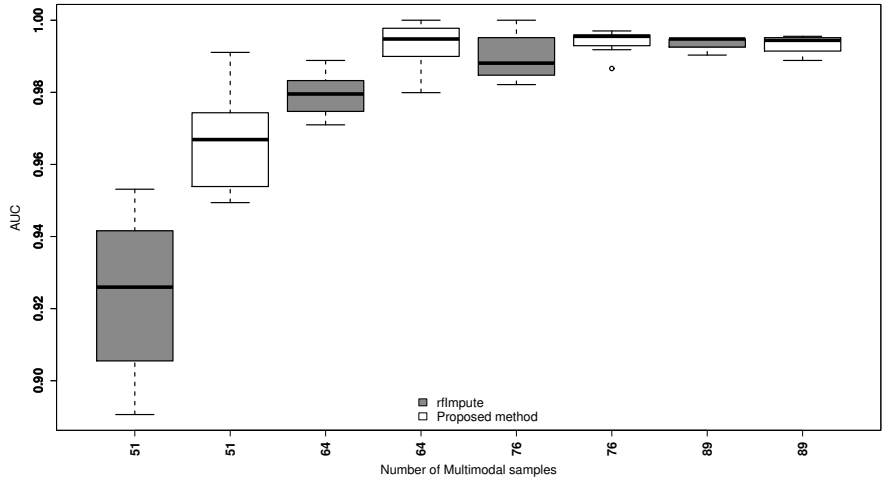


Figure 3.4: AUC vs multimodal sample size for the dermatology dataset (each box shows variation of AUC values for different randomised training and test sets)

features are missing. In the next section we use the heart disease dataset for this purpose.

Heart disease dataset

Figure 3.5 shows the AUC of the proposed method and the rfImpute method for different multimodal sample sizes. Each box in this figure shows AUC values for different single modal feature set sizes and a fixed multimodal dataset sample size. The expected upward trend in AUC *vs.* multimodal sample size is evident and it can be seen that the proposed method outperforms the rfImpute method especially in smaller samples sizes. For example, when only 14 multimodal samples are available, the rfImpute method results in a mean AUC of 0.91 whereas the proposed method delivers an AUC of 0.94. As the number of multimodal samples increases to 112, the performances increase for rfImpute and scandent tree to 0.96 and 0.97, respectively. In other words, the scandent tree approach has a clear advantage when the dataset with multimodal data is significantly smaller.

Figure 3.6 shows the AUC of the proposed method and the rfImpute method for different single modal feature set sizes. Each box shows changes of AUC for different sample sizes at a fixed feature set in the single modal

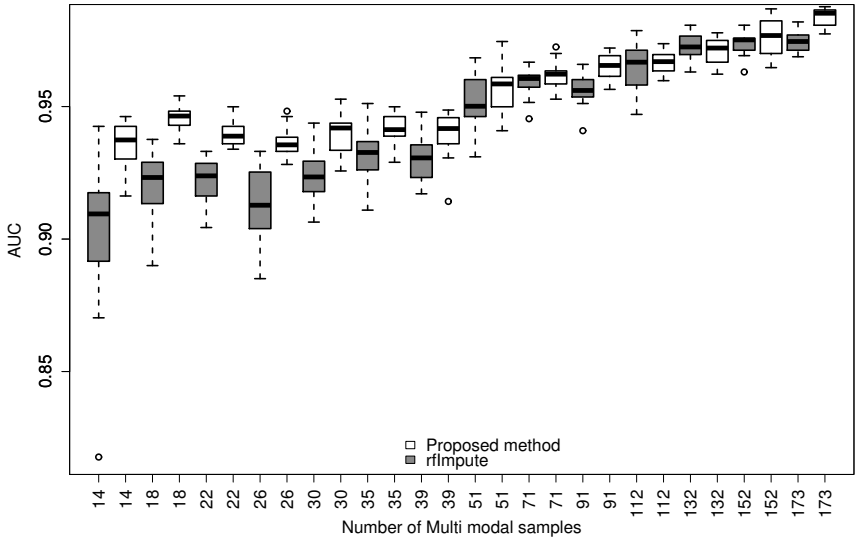


Figure 3.5: AUC vs multimodal sample size for heart disease dataset (each box shows AUC values for different single modal feature sets)

data. Smaller variances of the boxes for the proposed method, especially in smaller feature set sizes, show that the proposed method is on average less sensitive to the multimodal sample size especially when the single modal dataset has a large number of missing features. For example, at feature vector size of 2 for the single modal dataset, the performance of rfImpute varies from 0.88-0.98, whereas scandent tree shows a performance range of 0.93-0.98. This stable behavior is due to the unique ability of the scandent trees to predict division points for missing features that only conditionally depend on the available features.

Breast cancer dataset

Figure 3.7 shows the AUC for the proposed method and the state of the art imputation method of random forests as a function of multimodal sample size. Each box in this figure shows AUC of the two methods for a fixed multimodal sample size and different single modal feature set sizes. Smaller variances of the boxes for the proposed method, especially in smaller sample sizes, show that the proposed method is on average less sensitive to the single modal featureset size. For instance, at sample size of 393 for the multimodal dataset, the performance of rfImpute varies from 0.978-0.989, (more than

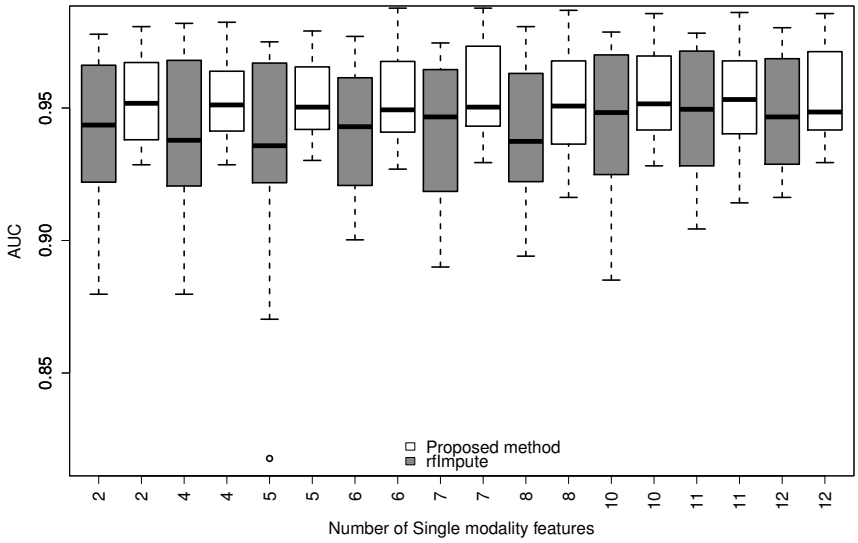


Figure 3.6: AUC vs single modal feature set size for heart disease dataset (each box shows AUC values for different multimodal sample sizes)

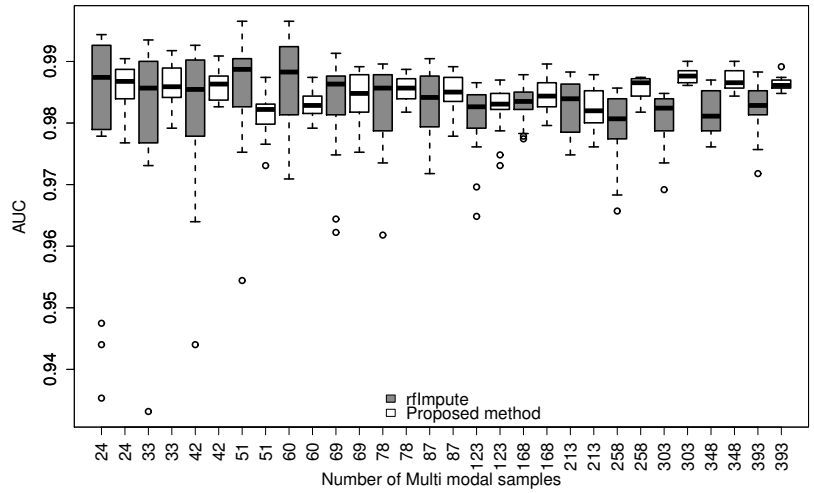


Figure 3.7: AUC vs multimodal sample size for breast cancer dataset (each box shows AUC values for different singlemodal feature sets)

1%) whereas scandent tree shows a performance range of 0.986-0.988 (less than 0.3%).

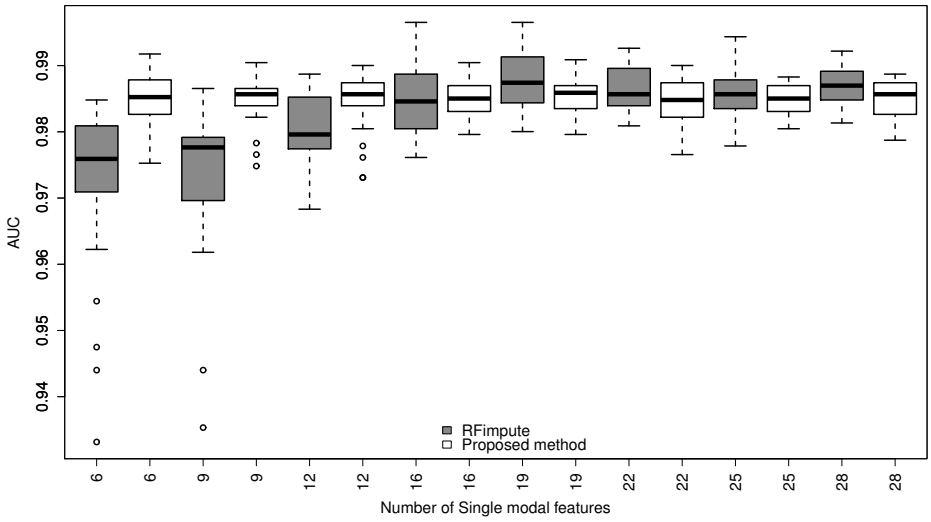


Figure 3.8: AUC vs single modal feature set size for breast cancer dataset (each box shows AUC values for different multimodal samples sizes)

Figure 3.8 shows the AUC of the proposed method and the rfImpute method for different single modal feature set sizes. Each box shows changes of AUC for different sample sizes at a fixed feature set in the single modal data. It can be seen that the proposed method outperforms the state of the art imputation method of random forests (rfImpute) when a large number of features are missing from the single modal dataset. It also has similar performance when the single modal feature set is almost complete. Smaller variances of the boxes for the proposed method, show that the proposed method is on average less sensitive to the multimodal sample size especially when the single modal dataset has a large number of missing features.

3.4.2 Experimental results: prostate cancer dataset

Figure 3.9 shows the AUC obtained on this data, for detection of prostate cancer, for several experiments, namely from left to right the bars show the distribution of AUC areas for 1) a multimodal decision forest that simply ignores the existence of archival imaging data, 2) our proposed scandent tree approach to use the archival data to improve the performance of a forest trained and testes on multimodal data, 3) the standard rfImpute method

applied at the forest level to include the single modal data in training, 4) the standard C5.0 method applied at trees level, 5) training and testing a tree using only the single modal features of the multimodal set, 6) KNN imputation, and 7) zeroing of the missing feature values.

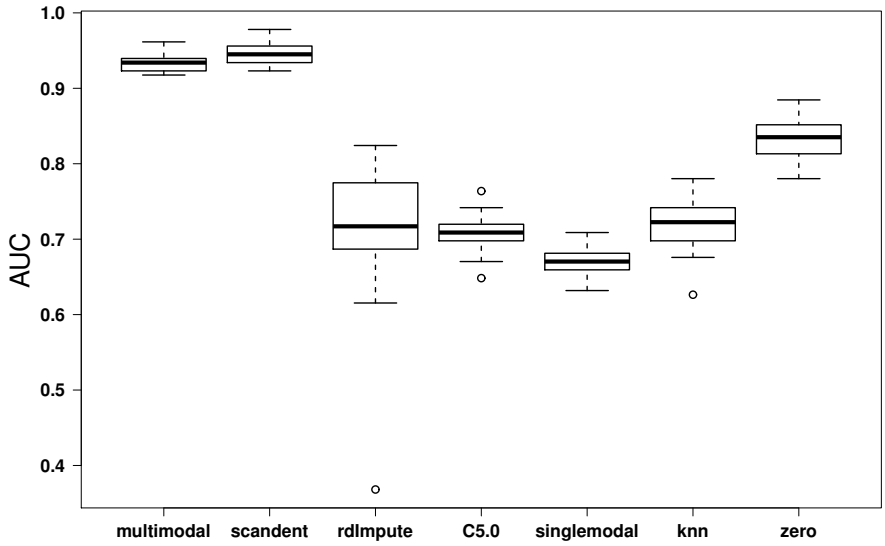


Figure 3.9: AUC for multimodal classification task obtained with different strategies for handling the missing data issue, prostate cancer dataset

It can be seen that the multimodal forest is performing significantly better than the single modal forest even though the sample size of the single modal dataset is significantly larger than the multimodal dataset. This suggests that the missing modality, in this case the genetic features, is far more discriminative than the shared modality, MRI. The imputation methods outperform a single modal forest, but they fail to outperform the multimodal forest. This shows that even the state of the art imputation methods may misguide the decision forest when a large portion of data is missing, to the extent that a simple imputation method like zero replacement outperforms the state of the art imputation approaches.

In order to measure the statistical significance of the difference between AUC values of different methods we use one of the most well-known univariate tests, student's t-test. The t-test measures the difference between two populations relative to their variances. This test is commonly used

when the sample size is small and the variance of the two populations are unknown. The outcome of a t-test is the p-value: the probability that the Null hypothesis (in this case, having similar distributions for AUC values) is true. In other words, the p-value is the probability that we observe a result similar to what we are observing or more extreme, assuming that the two distributions are equal. In our experiments we measure the significance of results by comparison of the AUC values resulting from 100 random runs of different methods.

In case of the proposed method, scandent forest, the significant advantage over a single modal forest, and each of the imputation methods is evident. Moreover, the proposed method does not introduce bias into the prediction like the other imputation methods and as a result, it outperforms both the multimodal forest and the single modal forest. However, because the shared modality is significantly less discriminative than the missing modality, the improvement in performance is small (mean AUC of 94% for the scandent forest and 93% AUC for the multimodal forest), although it is statistically significant (a two sample t-test resulted in $p < 0.01$).

3.5 Summary

In this section we introduced the novel concept of scandent trees, single modal trees that enable a conventional random forest trained on a small multimodal dataset to leverage a single modal dataset. Also a detailed explanation of the proposed method and implementation was presented.

We evaluated the proposed method using three publicly available datasets by simulation of different incomplete multimodal scenarios. We also compared the proposed method with the state of the art imputation methods on prostate cancer dataset as a real incomplete multimodal dataset. Using the dermatology and heart disease datasets we showed that the proposed method outperforms the state of the art imputation method of random forests if the multimodal dataset (the block of the dataset that has all of the modalities) is very small in comparison with the whole dataset. Using the breast cancer dataset we showed that the proposed method outperforms the state of the art imputation method for random forests if a large number of features be missing from the single modal dataset. Moreover, we showed that on all of the datasets, in comparison with rfImpute, the proposed method is in general less sensitive to multimodal sample size and single modal feature set size.

The experimental results on the prostate cancer dataset show that the

proposed method significantly outperforms the well known imputation methods, even the state of the art embedded imputation methods of random forests (rfImpute) or C5.0 trees. Because in this study the missing modality (genetic features) are far more discriminative than imaging features, a multimodal classifier which is equivalent of using the sample removal method to handle the missing values was the best method among the conventional imputation methods. The proposed method outperformed the multimodal classifier. This improvement was small, but statistically significant.

Chapter 4

Tree-based feature transforms: applying scandent tree model for single modal classification

4.1 Introduction

In the previous chapter we focused on a scenario in which a small but valuable multimodal dataset could be merged with a large and easy to access dataset in order to improve performance of a multimodal classifier. We discussed that this is a common scenario in biomedical data analysis laboratories or in data analysis scenarios that deal with very novel or special multimodal datasets. However, from clinical point of view the reverse problem is more attractive: leveraging a multimodal dataset in order to enhance a classifier trained and tested on a single modality.

One of the well-known applications of multimodal data analysis is in grading of generative diseases like cancer or Alzheimer's disease. Each modality provides us with unique information about the patient and is able to compensate for weaknesses of other modalities. However, the fact that many modalities are very expensive or are simply not feasible to use for every patient, limits the clinical applications of multimodal data analysis.

In case of Alzheimer's disease, the state of the art medical imaging modalities used are MRI and PET scan. The information provided by a PET scan is more useful than MRI because it provide information about cerebral blood flow, metabolism, and receptor binding but PET imaging is expensive and requires the use of radioactive tracers. As a result, a large number of patients only receive MRI scans. For example, in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study which is one of the largest multimodal studies of Alzheimer's disease worldwide, nearly half of the patients are missing the PET data. So a very valuable contribution in this

field would be to use a multimodal dataset together with a single modal dataset in order to train a classifier that only needs one of the modalities. This means, training a classifier that uses MRI and PET data to train a classifier that only needs MRI data for Alzheimer's disease grading.

In a fashion similar to the previous chapter, we benefit from the embedded way of decision forests for dealing with high dimensional data through feature bagging [4]. Another motivation for the use of decision forest paradigm is that it provides the ability to morph the treatment of missing data within the framework of learning to maximize the classification performance.

In this chapter we focus on applications of tree-based feature maps in the scandent tree model. A disadvantage of decision forests compared with SVM is the lack of an embedded framework for kernel-based feature transformation in the case of forests. Using multi-kernel approaches, researchers have devised solutions for incorporation of various modalities in the SVM context.

Tree-based feature transforms have recently received some attention. For example, a recent work by Cao *et al.,* [7] uses stacked decision forests. This method is based on using the probability values estimated by trees in a random forest as a feature vector, and using this feature vector for training of an enhanced decision forest, potentially together with the original feature set. Inspired by the applications of multi-kernel SVMs in multimodal data analysis, we apply this concept of tree-based feature maps, for multimodal data analysis.

In this chapter we intend to develop the idea of scandent tree-based feature transforms to solve the problem of missing data in the single modal testing scenario. This problem has many clinical applications in areas where expensive research protocols meet the realities of clinical practice and high cost. Here, the assumption of a multimodal dataset with block wise missing values remains. However, there is no multimodal assumption about the test set. To solve this problem, we use the idea of tree-based feature transforms along with the scandent tree. This combination allows us to use tree-based feature transforms built on one modality to transform the features from a different modality. Using this approach, we use MRI and PET data in the ADNI dataset and train a classifier that only requires the MRI data for the prediction of different stages of Alzheimer's disease. We show that the inclusion of the PET data at the time of training results in an improved classification accuracy, even though the test cases are not subjected to PET imaging. In order to test the performance of the proposed classifier in different scenarios, we also simulate the missing target scenario by intentionally removing a set of features from two publicly available benchmark datasets,

The dermatology dataset and The breast cancer dataset introduced in the previous chapter.

4.2 Method

To obtain a forest that transfers the value of the multimodal dataset into a single modal environment, it is tempting to simply replace all the trees in a forest trained on the available multimodal training data with their corresponding scandent trees. However, this approach fails due to bias and the fact that many of the multimodal divisions of support trees might not be predictable by the single modal feature set.

Instead, we choose an approach inspired by the use of decision trees as feature maps. For this we start with growing a scandent forest using the method explained in the previous chapter. However, instead of directly using the scandent trees, we use the set of local trees (T_i's) from all the scandent trees of a multimodal forest as tree-based "feature-maps" or "tree-based feature transforms". Each T_i is a single modal tree which maps F_s to a new space defined by the corresponding C' set. This means that each T_i yields a categorical feature to describe each sample. Then we use the single modal dataset with the extended feature set, including the original and these tree-based features, to grow an improved single modal forest. Note that trees trained on single modal features can be directly used as categorical or continuous (similar to [7]) feature transformers. In the current work, however, we use the scandent subtrees to link two inconsistent datasets.

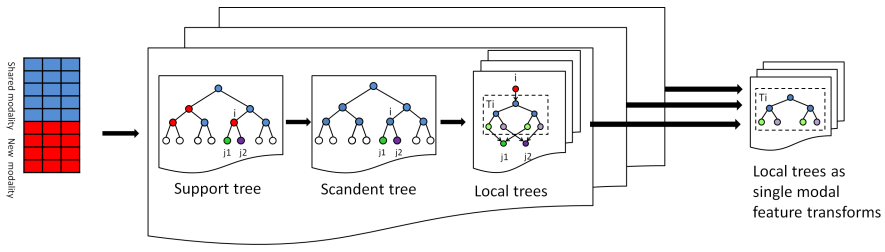


Figure 4.1: Extracting tree-based feature transforms from the scandent tree model

This method has a few advantages compared to the conventional method for forming a single modal decision forest or directly using the scandent trees as a new set of trees in a single modal decision forest. First, because at each split of each tree in the single modal forest, the tree growth algorithm

searches for the best division feature among both the original single modal features and the new features generated by the local trees (T_is), the resulting tree is expected to be more accurate than both the scandent tree and the tree grown using only the original single modal features. Second, although the T_is are formed by a small multimodal dataset, the feature selection criteria (Gini impurity or information gain) is calculated based on the large single modal dataset. In other words, the single modal forest uses the features inspired by the multimodal forest, but it is completely randomized and optimized based on the larger single modal dataset.

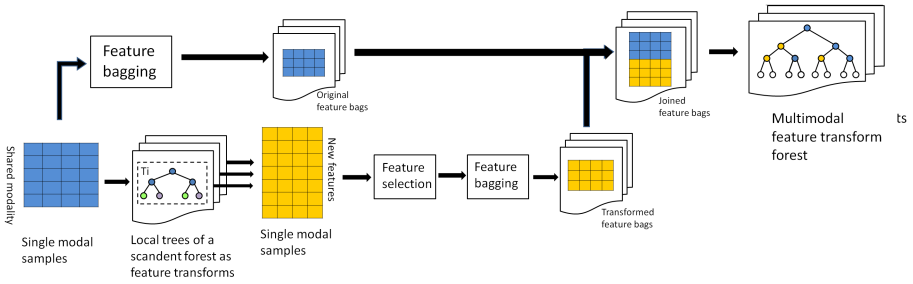


Figure 4.2: Diagram of the proposed method for training the "multimodal feature transform" forest

4.2.1 Implementation

The first step is to grow a multimodal forest and the related scandent trees using the method explained in the previous chapter. Then the local trees (T_is) are extracted from each tree and each T_i is used as feature generators for the single modal dataset. Given that each T_i is a single modal classifier, it can assign labels relative to the local class labels ($C^{'}$) to each single modal sample. The resulting labels are used as new categorical features which can be calculated for any test data using the corresponding T_i. We then use a conventional decision forest growth method similar to what was explained in the previous chapter to grow a forest using this set of new features together with the original single modal feature set.

It should be mentioned that because the local trees are trained using the small multimodal dataset, many of the generated features might not be useful for the single modal decision forest. Considering the large number of local trees in a random forest, this can flood the original single modal features. So we filter the new features by a conventional feature selection algorithm, namely based on the feature importance measure in a decision

forest. We apply feature bagging separately to the set of the original single modal features and the new features, and then merge them together to form the feature bag used for each single modal tree. Diagrams of the proposed method for extracting the tree-base feature transforms from the scandent tree model and using that for single modal classifier design are shown in figures 4.1 and 4.2.

4.3 Evaluation and results

In this section we evaluate the proposed method for single modal classification task. First, we examine the proposed method in different simulated scenarios and for different single modal sample sizes using two of the datasets used in previous chapter: the dermatology dataset and the breast cancer dataset. Then we test the performance of the proposed method on a real dataset that exactly matches our scenario, the ADNI dataset.

4.3.1 Evaluation using benchmark datasets

For this simulation task we should simulate the same two simulation parameters that were used in the previous chapter, with the difference that instead of the multimodal sample size that is assumed to be sufficiently large in this scenario, we examine the effect of the single modal sample size of the training set. The dermatology dataset introduced in the previous chapter is multimodal in nature so it is the perfect benchmark for testing the single modal sample size. However, because the feature sets for each modality is fixed, it is not meaningful to simulate the single modal feature size using this dataset. Instead, we use the breast cancer dataset introduced in the previous chapter in order to simulate different feature set sizes for single modal dataset.

Effect of the single modal sample size: dermatology dataset

It is informative to investigate performance of the proposed method for single modal classification tasks as a function of the single modal sample size. We design an experiment similar to the previous chapter using the the dermatology dataset but with the assumption that the test set is also missing the most discriminative modality. For this experiment we first randomly select 100 samples from the dermatology dataset as a test set. Then we discard the histopathological features from this set. Moreover, we form a multimodal dataset with a fixed size (40 percent of the remaining samples

in this experiment) and use it to extract the proposed feature transforms. Finally we evaluate performance of a single modal forest with and without the new features in different single modal sample sizes.

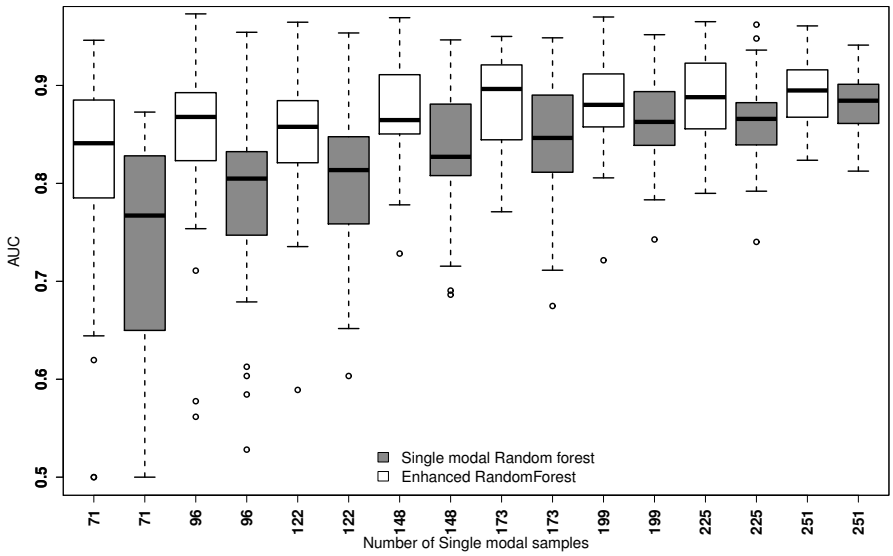


Figure 4.3: AUC vs single modal sample size for dermatology dataset

Figure 4.3 shows AUC of a single modal random forest and the enhanced forest trained on the dermatology dataset for different sample sizes. It can be seen that the enhanced forest can improve the single modal forest especially when the single modal dataset is small. For instance at sample size of 71 the single modal classifier gives AUC of 0.78 while the enhanced forest gives AUC of 0.84. This improvement is also evident in larger sample sizes but is less significant (AUC of 0.87 and 0.85 for the proposed method and a single modal forest). The dermatology dataset is multimodal in nature, so the sizes of the single modal and multimodal feature sets are fixed. On the other hand the breast cancer dataset does not have a predefined single-modal feature set so it can be used for simulation of different single modal featuresets.

Effect of the single modal feature set size: The breast cancer dataset:

For this simulation we first randomly select 100 test samples. We then change the single modal feature set size from 5 to 29 (the total feature set

size is 30) removing the most discriminative features at each step. Then we observe AUC as a function of feature set size for a single modal forest with and without the use of the tree-based feature transforms. The tree-based feature transforms used in each iteration were trained using 30 samples randomly selected from the multimodal dataset. We found that the size of the multimodal dataset in the breast cancer dataset does not have much effect on the tree-based feature transforms once the multimodal sample size is more than or equal to 30. Figure 4.4 shows the AUC of the conventional single modal forest and the proposed method for different single modal feature sets.

The expected upward trend in AUC *vs.* single modal feature set size is evident, it also can be seen that the proposed method outperforms a conventional single modal forest method especially in smaller feature set sizes. For example, when only the 5 least discriminative multimodal features are available, the conventional single modal random forest results in a mean AUC of 0.77 whereas the proposed method delivers an AUC of 0.83 while in larger feature sets the AUC of the two methods methods is almost equal. This is because at feature set sizes of larger than 20, the AUC of a single modal forest without using the tree based feature transforms is almost 1. This eliminates the need for any more improvement on the conventional classifier.

The breast cancer dataset and the dermatology dataset are complete datasets that provided us the opportunity to test the performance of the proposed method in different simulated scenarios. Furthermore we test the performance of the proposed method on a real incomplete multimodal dataset, the ADNI dataset.

4.3.2 A real scenario: ADNI dataset

We test the proposed single modal classification method on a dataset from Alzheimer's Disease Neuro-imaging Initiative (ADNI) database. The ADNI was launched in 2003 as a public-private partnership, led by Dr. Michael W. Weiner. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers Disease (AD). The ADNI study is an example of a multimodal scenario in which a large portion of samples are missing one of the modalities. In this chapter we take the samples that come from patients with both MRI and PET scan as multimodal dataset ($N_m = 218$)

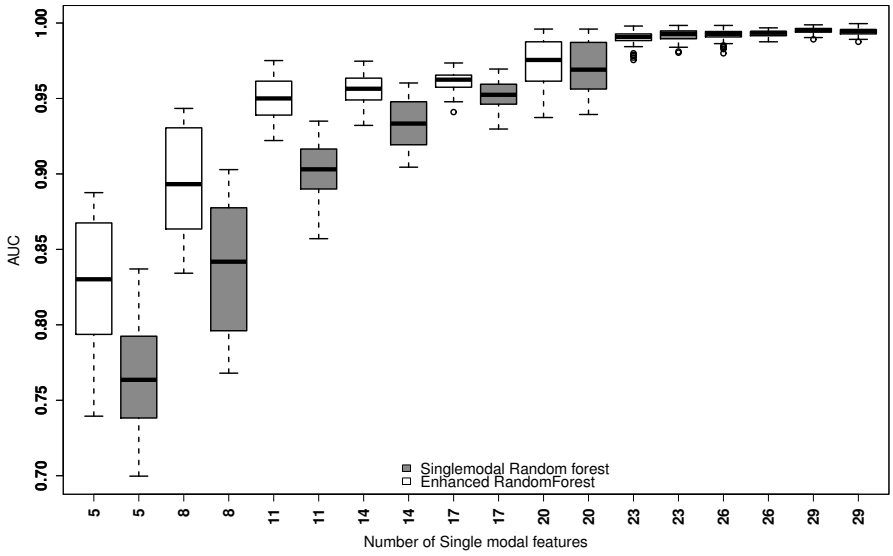


Figure 4.4: AUC vs single modal feature set size for breast cancer dataset (each box shows AUC values for different multimodal sample sizes)

accompanied by a relatively large single modal dataset ($N_s = 508$) consisting of patients with only MRI data. This includes the MRI data from the 218 multimodal samples.

The single modal dataset consists of MRI volume measurements of six ROIs in the human brain (ventricles, hippocampus, whole-brain, entorhinal, fusiform and mid-temporal) and intra-cranial volume (ICV) in mm^3. The multimodal feature set consists of the same MRI features together with two additional PET scan features, FluoroDeoxyGlucose (FDG) measurement and AV45 uptake measurement. The outcome labels include cognitively normal patients (NL), patients with confirmed dementia (AD) and patients with mild cognitive impairment (MCI). The MCI group can be divided into progressive (pMCI) that eventually converts to dementia and stable (sMCI). In this chapter we assume a maximum of 36 month conversion time for the MCI class to be considered pMCI.

The distribution of different outcome classes in the two datasets is as follows: for the normal class we have 178 samples in the single modal dataset versus only 18 samples in the multimodal dataset, for the dementia class we have 108 single modal samples versus 29 multimodal samples, for the sMCI

class we have 126 single modal versus 144 multimodal samples and for the pMCI class we have 96 single modal samples versus 27 multimodal samples. In other words, not only the multimodal dataset is much smaller than the single modal dataset, it also does not have the same distribution of outcome classes. This makes the data fusion between the two datasets extremely difficult with traditional approaches. We examine the performance of the proposed method by reporting AUC for three classification scenarios: NL versus pMCI, sMCI versus AD. and sMCI versus pMCI.

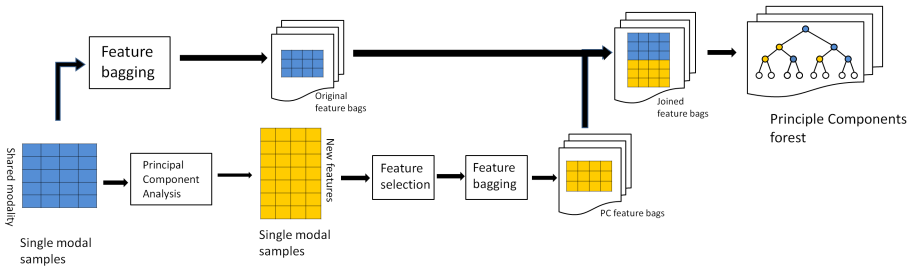


Figure 4.5: Diagram of the method used for forming the PC forest

We take the performance of a single modal forest trained solely on the original MRI feature set as the baseline. We then compare performance of this baseline with a forest enhanced using the principal component features (PC forest) as shown in Figure 4.5, with a similar forest trained using MRI-based transformed features shown in Figure 4.6, and a transformed feature set extracted using both MRI and PET using scandent tree approach shown in Figure 4.2. It should be noted that all of these classifiers are designed for the single modal classification task, meaning that they only need the original MRI feature set for classification but may use the other modalities (PET in this example) for better feature transform design in the training phase.

5-fold cross-validated ROC curves of the baseline single modal forest, PC forest, single modal feature-transform forest, and the scandent tree multimodal feature transform forest for NL vs. pMCI classification task are shown in Figure 4.7.

As it can be seen in Table 4.1, the forests grown based on the tree-based feature transforms significantly outperform the baseline single modal forest and the PC forest. The difference between the baseline and the feature transform methods is statistically significant (p=0.01) for the single modal feature transforms and (p=0.002) for multimodal feature transforms.

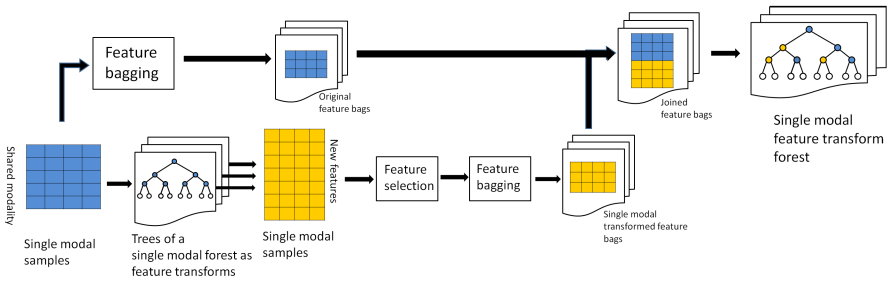


Figure 4.6: Diagram of the method used for forming a forest based on single modal tree-based feature transforms

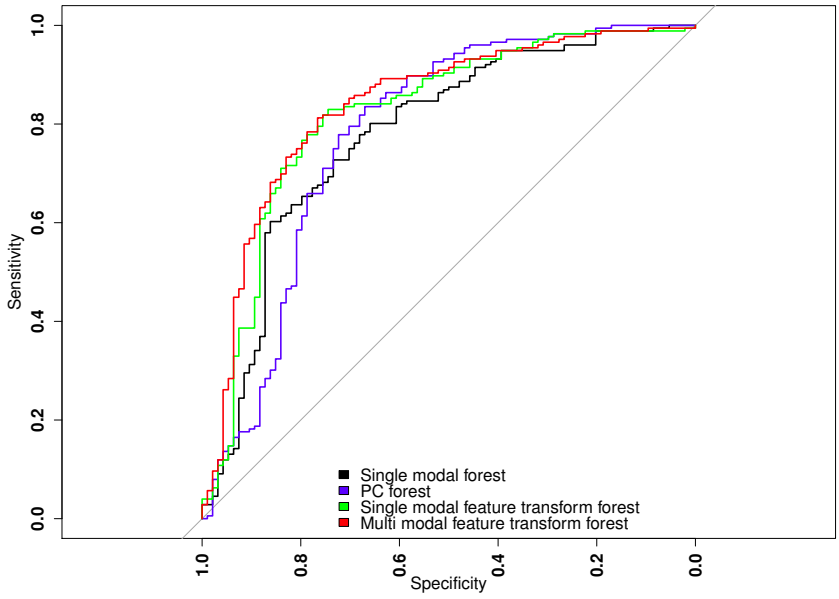


Figure 4.7: ROC curve for NL vs. progressive MCI classification, single modal classification task, ADNI dataset

Table 4.1: Accuracy (Acc), Sensitivity (Sens), Specificity (Spec) and Area under ROC curve (AUC) of the proposed methods and the baseline forest for the NL vs. pMCI single modal classification task, ADNI dataset

	Acc	*Sens*	*Spec*	*AUC*
Single modal forest	0.744	0.663	0.791	0.779
PC forest	0.774	0.878	0.747	0.781
Single modal feature transform	0.781	0.691	0.844	0.819
Multimodal feature transform	0.788	0.747	0.805	0.837

Table 4.2: Accuracy (Acc), Sensitivity (Sens), Specificity (Spec) and Area under ROC curve (AUC) of the proposed methods and the baseline forest for the sMCI vs. AD single modal classification task, ADNI dataset

	Acc	*Sens*	*Spec*	*AUC*
Single modal forest	0.731	0.824	0.699	0.814
PC forest	0.752	0.758	0.748	0.836
Single modal feature transform	0.782	0.734	0.863	0.868
Multimodal feature transform	0.795	0.737	0.897	0.892

However, the improvement in the performance achieved by the PC-based features is not statistically significant (p-value = 0.92). The multimodal feature transforms are more effective compared to the single modal feature transforms. This difference is significant (p=0.04).

Another classification problem worth investigating is discrimination of samples with stable MCI from dementia cases using the MRI feature set. Figure 4.8 and Table 4.2 show ROC curves and performance measures of the enhanced and baseline forests for this classification task.

It can be seen that similar to the NL vs. pMCI task, the forests enhanced by the new feature sets are outperforming the baseline single modal forest. The improvement observed in the PC forest is more significant than the previous task but it still can not be considered statistically significant (p-value=0.08). On the other hand the proposed tree-based feature transform methods significantly outperform the baseline methods with p-values of 0.0001 and 3.698e-07 for the single and multimodal feature transforms,

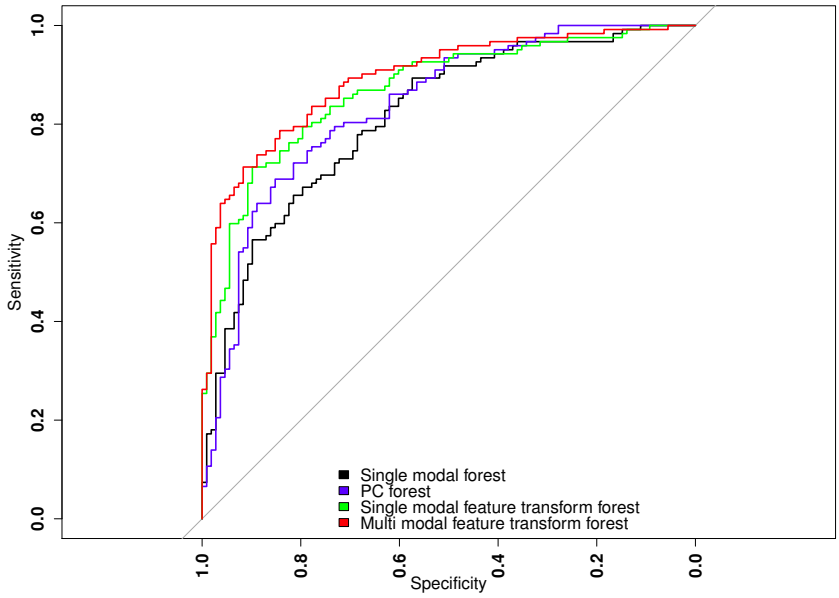


Figure 4.8: ROC curve for stable MCI vs. AD classification, single modal classification task, ADNI dataset

respectively, and the multimodal feature transforms are more effective than single modal feature transforms (p=0.0003).

The third classification task which separates sMCI from pMCI cases is potentially the most clinically relevant model. The ROC curves and performance measures for this task can be seen in Figure 4.8 and Table 4.2.

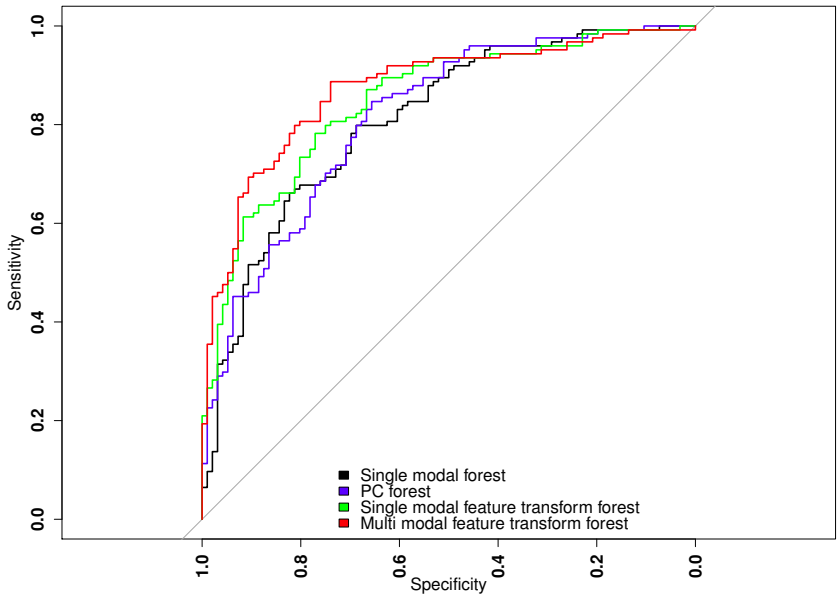


Figure 4.9: ROC curve for stable MCI vs. progressive MCI classification, single modal classification task, ADNI dataset

The trends remain the same: the tree-based feature transforms outperform a simple single modal forest with p=0.01 and and p=0.0002 for the single modal (MRI-based) and multimodal (MRI+PET) feature transforms, respectively. It can also be seen that the PC-based features fail to enhance the baseline forest to a statistically significant level (p=0.672). Similar to the previous experiments, the multimodal feature transforms yield a larger AUC than single modal feature transforms with p=0.01.

4.3.3 Comparison with other work on ADNI

The block wise missing value problem is a well-known issue of the ADNI dataset and it is addressed in many papers in literature. However, none of them has the same goal and assumptions as our study. For instance,

Table 4.3: Accuracy (Acc), Sensitivity (Sens), Specificity (Spec) and Area under ROC curve (AUC) of the proposed methods and the baseline forest for the sMCI vs. pMCI single modal classification task, ADNI dataset

	Acc	*Sens*	*Spec*	*AUC*
Single modal forest	0.743	0.713	0.744	0.810
PC forest	0.757	0.769	0.746	0.815
Single modal feature transform	0.777	0.819	0.750	0.848
Multimodal feature transform	0.815	0.831	0.803	0.872

this paper focuses on improving the performance of decision forests with the assumption that a decision forest is the classifier of choice for a given multimodal dataset. However, most of the studies on the ADNI dataset use other classifiers like multi-kernel SVM for multimodal classification. As a result, it is difficult to compare our results with the available literature as any such comparison will be mostly informed by the choice of classification paradigm.

One other issue that makes the comparison difficult is the different feature sets and sample sizes impacted by patient selection criteria. A simple example is the different assumptions on the conversion time for MCI to AD for differentiating progressive versus stable MCI. In our study, we assumed a 36 month conversion time for progressive MCI cases and used the summarized set of features extracted by adnimerge R package as our feature set. This package is accessible from the ADNI website (https://adni.loni.usc.edu).

With all these differences and limitations in mind, we have gathered a list of comparable methods with performance measures reported in the literature in Table 4.4. These are all on the sMCI vs pMCI classification task. As it can be seen, the proposed method matches or surpasses the performance of the state of the art, even in cases where multimodal data is available for all cases.

Table 4.4: Comparison of the proposed single modal method with the state of the art for sMCI vs. pMCI prediction, ADNI dataset

Method	sample size	modalities	performance			
			Acc	*Sens*	*Spec*	*AUC*
Proposed method	122	MRI	0.815	0.831	0.803	0.872
[9]	99	MRI, PET, CSF	0.801	0.853	0.733	0.852
[31]	204	MRI, PET	0.759	0.48	0.952	0.746
[6]	397	MRI, PET, CSF	0.732	0.655	0.767	0.786
[14]	388	MRI	0.754	0.705	0.776	0.82
[35]	200	MRI	0.751	-	-	0.84
[40]	143	MRI, PET, CSF, APOE	0.741	0.787	0.656	0.795
[43]	91	MRI,PET,CSF	0.739	0.686	0.736	0.797
[10]	405	MRI	0.71	0.7	0.72	-
[36]	162	MRI, CSF	0.685	0.741	0.63	0.76

4.4 Summary

In this chapter a novel application of the scandent tree model as tree-based feature transforms was introduced. These feature transforms can be used for leveraging a multimodal dataset for single modal classifier design. A method to extract these single modal tree feature transforms was explained and a method to use these feature transforms within the context of a conventional random forest classifier was provided.

Using two publicly available datasets for simulation and the ADNI dataset as a real incomplete multimodal dataset, it was shown that the proposed method can be used to enhance a single modal classifier. The experiments on the ADNI dataset show if one extracts the proposed feature transforms from a scandent forest trained on both PET and MRI features and uses them together with the original set of MRI features for classification of different stages of Alzheimer's disease, the resulting classifier can outperform a conventional single modal random forest, a random forest enhanced with PCs of the MRI features and a random forest enhanced with tree-based feature transforms solely trained on the MRI data.

We examined the proposed classifier in three different scenarios of normal vs progressive MCI, stable MCI vs progressive MCI and stable MCI vs Alzheimer's disease. It was shown that in all of these scenarios the proposed method outperforms the conventional single modal forest and the other enhanced single modal forests mentioned. The stable MCI vs progressive MCI classification task was found to be the most clinically valuable classification task. Comparison of the proposed method and state of the art in this scenario show that the proposed method matches or outperforms the state of

the art even in case of larger training sets or feature sets.

Chapter 5

Conclusion

5.1 Summary

In this thesis we addressed the problem of incomplete multimodal datasets in random forest learning algorithms in a scenario where many of the samples are non-randomly missing a large portion of the most discriminative features. This missing value problem in multimodal datasets is different from the common scenarios in single modal data analysis. In our problem of interest, features of one specific modality might be missing altogether in training, or testing. This causes an issue known as block-wise missing data. We showed that this issue can not be handled by conventional imputation techniques if the number of samples that include all of the features is small. This is a common scenario in biomedical data analysis applications. In summary, this thesis has two major contributions:

- We developed the novel concept of scandent trees for enriching a multimodal classifier with a large training dataset from only a subset of modalities. The results show that the proposed method for multimodal classification outperforms the embedded missing value imputation method of decision forests introduced in [4] and other state of the art imputation methods, particularly in smaller samples sizes and when a large portion of features are missing. We showed that the proposed method enables the integration of a small genomic plus imaging dataset, with a relatively large imaging dataset. We also showed that this method is in general less sensitive to the number of missing features and to the multimodal sample size by simulation of different missing value scenarios on three publicly available benchmark datasets.

- We also proposed a novel learning method for training on multiple modalities and testing on one modality. To this end, we introduced the concept of tree-based feature transforms. We showed that using this approach, we can efficiently transfer the discriminative power of PET imaging into the training phase of building a model that would

only use the MRI data at the testing phase. By simulation on two publicly available benchmark datasets, we showed that the single modal features generated by the multimodal model can significantly improve a single modal forest especially when the single modal dataset is small or it is missing most of the discriminative features. We also showed that the model achieved through multimodal data analysis can be used to form an enhanced random forest that only needs a single modality for classification.

5.2 Discussions and limitations

In this thesis we have addressed two classification scenarios regarding the problem of missing values in multimodal datasets: leveraging a single modal dataset for multimodal classification, and leveraging a multimodal dataset for single modal classification.

5.2.1 Limitations of the implementated method

There are some limitations to our current implementation of the proposed method:

As one limitation, it should be mentioned that similar to the other tree-based imputation methods, we are proposing this algorithm with the assumption that random forest is the classifier of choice. In other words, when a different classifier outperforms the baseline random forest, the proposed method might not be the best option.

Another limitation of the current proposed method is that it is based on optimising each leaf of each tree in a forest separately which is computationally expensive in large forests. Unlike the conventional imputation methods, the computational time of the proposed method scales with the sample size of the small multimodal dataset not the large sample size of the single modal dataset. However, as the multimodal dataset grows the computation time becomes an issue. The computational complexity of the proposed method not only depends on the model parameters, but also depends on the dataset and feature-sets used in the analysis. For instance, if the missing modalities are significantly more discriminative than the available modalities, it is expected that the link nodes are limited to the root and leaves of the support forest. As a result, the computational cost of training a scandent tree forest would be in the same order as training two support forests. A similar result is also expected if the missing feature set is significantly less

discriminative. However, the testing phase of the scandent forest is exactly similar to a conventional random forest.

5.2.2 Discussions and limitations of the multimodal study

We tested the scandent tree method on a prostate cancer dataset as a real world example of our target scenario. This dataset is an example of the worst case scenario of missing data: a large non-random portion of the data is missing the potentially more powerful genomic features resulting in a very small multimodal dataset. At the same time, the number of features on the single modal (imaging) side is small. It is, therefore, revealing that even in this situation, the use of scandent tree methodology provides a clear advantage against the traditional approaches to deal missing values in a situation like this, such as simply ignoring one or the other set, or imputation approaches.

There are a few limitations to our work with prostate cancer data:

- As our experiments show, the missing modality (gene expression) is far more discriminative than the shared modality (MRI). This makes it extremely difficult for the proposed method to model the relationships between the modalities and effectively merge the two datasets.
- The small number of features in the shared modality (MRI) makes the feature-bagging in the support tree unbalanced between the modalities. As a result, many of the support trees are completely grown based on the missing modality (gene expression) and scandent trees have to follow the structure of a whole support tree. This together with small sample size of the multimodal dataset can cause over-fitting.
- Another limitation is that the small sample size of the multimodal dataset. It not only makes it hard to train the support forest needed for the scandent tree method, it also makes it hard to show statistically significant results in comparison between different methods.

Given these limitations, a more revealing test of the performance of the solution proposed for the multimodal scenario was achieved by study of the benchmark datasets. One such study was presented in chapter 3 where we examined the performance of the scandent tree method for different mutlimodal sample sizes and different feature sets using benchmark datasets publicly available from the University of California Irvine (UCI) database [23]. In comparison with the state of the art imputation method for decision forests (rfImpute), we observed that in larger multimodal sample sizes

or when only a small number of features were missing from the single modal dataset, both of the methods perform very well in handling the missing values for multimodal classification. However, in smaller multimodal datasets or when a large portion of features are missing from the single modal samples, the scandent tree method showed significantly better performance in comparison with the rfImpute method. Another observation was that for a fixed sample size, the scandent tree method is less sensitive to the number of missing features, especially in smaller multimodal sample sizes. This advantage was also evident from the results on the prostate cancer dataset.

5.2.3 Discussions and limitations of the single modal study

We proposed the tree-based feature transform method in order to leverage a multimodal dataset for single modal classification. We showed than this method is very effective in common real-world scenarios when a large number of samples are missing many of the potentially most discriminative features. We also succeeded to leverage information from PET scan in ADNI dataset for enhancement of classification of different stages of Alzheimer's disease when only MRI data is available.

Unlike the prostate cancer study we did not have the problem of sample size in this study but there is one limitation in our results on the ADNI dataset: The different stages of Alzheimer's disease are not clearly defined and vary from one study to another. For instance, the conversion time between MCI and Alzheimer's disease which determines whether an MCI case is stable or progressive is different between different studies. Also the fact that the ADNI dataset is continuously being updated, makes it hard to compare results on this dataset.

It should also be mentioned that the current implementation of the proposed method is not computationally efficient. This becomes an issue if the relationship between the available modalities and the missing modalities is very complex or the feature set size is very large. This scenario will require a large scandent forest in order to generate the tree-based feature transforms. Considering the fact that most of the scandent trees consist of at least 2 or 3 local trees, the number of the tree-based feature transforms becomes very large. The computation costs of growing a scandent forest for large datasets may make it computationally unpractical to use in many studies. However, our experiments show that only around 5% of the feature transform trees are actually useful in practice. This means that 95% of the computation time used for training of the scandent forest is not necessary if it is only grown for the purpose of generating the tree-based feature transforms.

Moreover, note that a large set of tree-based feature transforms are only locally discriminant and their number may exceed the original single modal feature set size by a factor of 5. Therefore, using an off-the-shelf random forest that does not separately rank and bag the two feature sets may not result in an improved classifier and even may result in a forest weaker than the original single modal forest. This is because a large number of artificial features may flood the set of original features. This becomes an issue in the forest growth algorithms that use a randomized feature selection approach at each division of each tree in a forest. The randomized methods may result in decreased growth time and independence between trees. However, they may also yield in a large number of trees that are grown solely based on the transformed features. These trees are very likely to over-fit.

A similar problem is when the scandent trees are not randomised efficiently. In this scenario, the transformed features are very likely to be statistically dependent because they are formed using the same feature-set without bagging. This makes it hard to form independent decision trees and as a result limits the performance of the resulting decision forest.

5.3 Future work

We can envision four future improvements to the implementation of the scandent tree method:

- The current implementation of the proposed inference method is based on the optimization of the scandent trees in a leaf by leaf manner. This process is computationally expensive and becomes an issue in case of larger multimodal datasets. An area of improvement will be redesigning the inference method so that instead of leaf-by-leaf merging of the scandent and support trees, it can merge the two trees directly at tree level.

- The second area for continued work is improving the baseline trees. Because we needed full control over each division of each tree in the forest, we could not use the off-the-shelf decision forest packages available in R. Therefore, the support forest, which is the base of the scandent tree method, is our in-house implementation. We can improve the base classifier by using the boosting methods and advanced randomizing schemes embedded in off-the-shelf random forests.

- The third area that needs more work is the design of the local predictors in each scandent tree. The current implementation uses a very

simple classification tree while any other machine learning algorithm that can mimic a local multimodal tree can be used. The current implementation has the advantage that it automatically results in a scandent tree, but methods resulting in any single modal rule-set can also be used.

- The fourth area is adapting the scandent tree model for online learning. As it was mentioned, one of the main applications of the proposed method is in data analysis tasks where only a limited set of the samples are multimodal because of the time constraints. Therefore a natural step can be to design a mechanism to insert information from new multimodal samples into the scandent tree model. This way the model can be iteratively trained as the number of multimodal samples grows.

- Another potential applications of the scandent tree concept that is worth investigating is relationship finding between modalities. The scandent tree model provides a unique tool to model the relationship between two set of features potentially from two modalities. The local trees that form each scandent tree are in fact decision trees that model the relationship between one available feature and a set of features in the missing modality. In many applications, for instance in gene expression studies, a decision tree is a perfect tool to model the relationships between features in an interpretable fashion. This can be a very simple but effective way to find relationships between an imaging modality and a set of genes which may result in finding new biomarkers for diagnosis of diseases like cancer by finding links between features.

Bibliography

[1] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[2] Francisco Arteaga and Alberto Ferrer. Framework for regression-based missing data imputation methods in on-line mspc. *Journal of chemometrics*, 19(8):439–447, 2005.

[3] Hussam Aldeen Ashab, Piotr Kozlowski, S Larry Goldenberg, and Mehdi Moradi. Solutions for missing parameters in computer-aided diagnosis with multiparametric imaging data. In *Machine Learning in Medical Imaging*, pages 289–296. Springer, 2014.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[6] Sergio Campos, Luis Pizarro, Carlos Valle, Katherine R Gray, Daniel Rueckert, and Héctor Allende. Evaluating imputation techniques for missing data in adni: A patient classification study. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 3–10. Springer, 2015.

[7] Yu Cao, Hongzhi Wang, Mehdi Moradi, Prasanth Prasanna, and Tanveer F Syeda-Mahmood. Fracture detection in x-ray images through stacked random forests feature fusion. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 801–805. IEEE, 2015.

[8] Qixuan Chen and Sijian Wang. Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in medicine*, 32(21):3646–3659, 2013.

[9] Bo Cheng, Mingxia Liu, Heung-Il Suk, Dinggang Shen, and Daoqiang Zhang. Multimodal manifold-regularized transfer learning for MCI conversion prediction. *Brain imaging and behavior*, pages 1–14, 2015.

[10] Pierrick Coupé, Simon F Eskildsen, José V Manjón, Vladimir S Fonov, Jens C Pruessner, Michèle Allard, and D Louis Collins. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. *NeuroImage: clinical*, 1(1):141–152, 2012.

[11] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

[12] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.

[13] B. Drew, E. C. Jones, S. Reinsberg, and et al. Device for sectioning prostatectomy specimens to facilitate comparison between histology and in vivo MRI. *Journal of magnetic resonance imaging : JMRI*, 32:992–996, 2010.

[14] Simon F Eskildsen, Pierrick Coupé, Daniel García-Lorenzo, Vladimir Fonov, Jens C Pruessner, and D Louis Collins. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *Neuroimage*, 65:511–521, 2013.

[15] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[16] Nandinee Fariah Haq, Piotr Kozlowski, Edward C Jones, Silvia D Chang, S Larry Goldenberg, and Mehdi Moradi. A data-driven approach to prostate cancer detection from dynamic contrast enhanced MRI. *Computerized Medical Imaging and Graphics*, 41:37–45, 2015.

[17] Yan He. *Missing data imputation for tree-based models*. PhD thesis, University OF California Los Angeles, 2006.

[18] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[19] Tin Kam Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.

[20] Biao Jie, Daoqiang Zhang, Bo Cheng, and Dinggang Shen. Manifold regularized multitask feature learning for multimodality disease classification. *Human Brain Mapping*, 36(2):489–507, 2015.

[21] Phimmarin Keerin, Werasak Kurutach, and Tossapon Boongoen. Cluster-based knn missing value imputation for dna microarray data. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pages 445–450. IEEE, 2012.

[22] EM Kleinberg et al. An overtraining-resistant stochastic modeling method for pattern recognition. *The annals of statistics*, 24(6):2319–2349, 1996.

[23] M. Lichman. UCI machine learning repository, 2013.

[24] Mehdi Moradi, Firdaus Janoos, Andriy Fedorov, Petter Risholm, Tina Kapur, Luciant D Wolfsberger, Paul L Nguyen, Clare M Tempany, and William M Wells. Two solutions for registration of ultrasound to mri for image-guided prostate interventions. In *Engineering in Medicine and Biology Society (EMBC), 2012 annual international conference of the IEEE*, pages 1129–1132. IEEE, 2012.

[25] Mehdi Moradi, Septimiu E Salcudean, Silvia D Chang, Edward C Jones, Nicholas Buchan, Rowan G Casey, S Larry Goldenberg, and Piotr Kozlowski. Multiparametric MRI maps for detection and grading of dominant prostate tumors. *Journal of Magnetic Resonance Imaging*, 35(6):1403–1413, 2012.

[26] National Institutes of Health. National cancer institute: PDQ genetics of prostate cancer, Date last modified 02/20/2015.

[27] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[28] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[29] Dan Steinberg and Phillip Colla. Cart: classification and regression trees. *The top ten algorithms in data mining*, 9:179, 2009.

[30] Carolin Strobl, Anne-Laure Boulesteix, and Thomas Augustin. Unbiased split selection for classification trees based on the gini index. *Computational Statistics & Data Analysis*, 52(1):483–501, 2007.

[31] Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, et al. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, 101:569–582, 2014.

[32] Terry M Therneau, Beth Atkinson, and Brian Ripley. rpart: Recursive partitioning. R package version 3.1-46. *Ported to R by Brian Ripley.*, 3, 2010.

[33] Terry M Therneau, Beth Atkinson, Brian Ripley, et al. rpart: Recursive partitioning. *R package version*, 3:1–46, 2010.

[34] Xian Wang, Ao Li, Zhaohui Jiang, and Huanqing Feng. Missing value estimation for dna microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC bioinformatics*, 7(1):1, 2006.

[35] Chong-Yaw Wee, Pew-Thian Yap, and Dinggang Shen. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. *Human brain mapping*, 34(12):3411–3425, 2013.

[36] Eric Westman, J-Sebastian Muehlboeck, and Andrew Simmons. Combining mri and csf measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. *Neuroimage*, 62(1):229–238, 2012.

[37] William H Wolberg, W Nick Street, Dennis M Heisey, and Olvi L Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, 26(7):792–796, 1995.

[38] Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, and Jieping Ye. Multi-source learning with block-wise missing data for Alzheimer's disease prediction. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 185–193. ACM, 2013.

[39] Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, and Jieping Ye. Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, 102:192–206, 2014.

[40] Jonathan Young, Marc Modat, Manuel J Cardoso, Alex Mendelson, Dave Cash, and Sebastien Ourselin. Accurate multimodal probabilistic prediction of conversion to alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2:735–745, 2013.

[41] Guan Yu, Yufeng Liu, Kim-Han Thung, and Dinggang Shen. Multi-task linear programming discriminant analysis for the identification of progressive MCI individuals. *PLOS One*, 9:e96458, 2014.

[42] Lei Yuan, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, and Jieping Ye. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *Neuroimage*, 61(3):622–632, 2012.

[43] Daoqiang Zhang and Dinggang Shen. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage*, 59(2):895–907, 2012.