A Review On Random Forest In Biological And Medical Data

Michael Herzog
Department of Computer Science
North Carolina State University
Raleigh, North Carolina
mherzog@ncsu.edu

Sandeep Reddy Katypally
Department of Computer Science
North Carolina State University
Raleigh, North Carolina
skatypa@ncsu.edu

ABSTRACT

This paper reviews Random Forest. Therefore we start with a paper of 2008 that uses Random Forest and review papers for each year up to 2016. New approaches and comparison are described and critically scrutinized. Extensions that has been added to the Random Forest framework are also described and compared to each other. All paper's content is the analysis of medial or biological data. Those date has certain differences to other datasets. The differences and their effects are described and the usage of Random Forest on those datasets is evaluated.

CCS Concepts

•**Computing methodologies** → **Classification and regression trees;** •**Applied computing** → *Computational genomics; Bioinformatics;* •**Random Forest** → **Machine learning algorithms;**

Keywords

Random Forest, Curse of dimensionality, Medical Data, Biological Data

1. INTRODUCTION

The course of dimensionality is a well-known problem in the field of medical and biological data science. Due to the large amount of features, it is very complicated to extract the features or combination of features with the most influence on the class labels. Medical or biological data normally consists of many rows and few columns. This paper reviews the development and the variants of a machine learning algorithm called Random Forest. Starting from the 2012 paper by Touw et al. "Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?" [19] we take a look back and forwards with focus on the applied methods. We show how Random Forest and its variants perform in the biological or medical context.

Random Forest [1] is a machine learning method for classification and regression. The motto of random forests is, "If one tree is good, why not build a whole bunch?" [14]. An ensemble of trees is used to detect the most popular class.

Schwarz et al. [16] propose Random Jungle, a software package, which is based on Random Forest. It was said to be up to 159 times faster than alternative approaches. One reason for that is the implemented so called "variable backward" elimination. The variable backward elimination drops features as long as they aren't enough significant. The significance level is set manually prior. Random Jungle also allows a parallel computation of trees, which is very suitable for datasets with a large number of features. So many trees can be grown in parallel. As mentioned by Touw et al. in their paper it was the fastest Random Forest implementation.

Willows [20] is a software package that uses Random Forest as well as a classification tree and a deterministic forest. Its purpose is to maximize the computer memory and to allow an analysis of massive genotype data, which is a particular type of biological data.

Hapfelmeier and Ulm [5] developed a new way for variable selection and tested it against eight other approaches for variable selection. Random Forest has been shown to perform well for high prediction accuracy and distinguishing between important and unimportant variables. Their new way, they say, distinguishes even better than the other approaches between important and unimportant variables while it is still as fast as the others. Their baseline idea is to use permutation to produce unbiased importance measures.

In Maken et al. [9] Random Forest is compared to kNN and the multiple instance learning method (MIL). Their subject is a biomedical topic. They use the methods for the classification of Breast Cancer using Magnetic Resonance Imaging. They use generic (tile-biased) as well as domain specific features (region-of-interest). The subject of detecting breast cancer is a more graphical use of classification, in which Random Forest also performs well. "Data mining and linked open data - New perspectives for data analysis in environmental research" [8] provides an overview on techniques that can be used. Their idea behind is to give an overview of existing data mining techniques due to the increase of the amount of data that is collected in the 22th century. They also confirm the use of Random Forest in the topics Medicine and Biology, which are of type spatial data mining, image mining or text mining. All in all the paper shows mainly the different tools which can be used for the analysis of big data in different fields of interest.

The 2016 paper "On the use of Harrel's C for clinical risk prediction via random survival forests" [15] uses the Harrel's C as a split criterion for the so called Random Survival Forest. "Random Survival Forest" [7] extends the Random Forest towards survival data. Normally Random Survival Forest use the log-rank split criterion. The new approach, Harrel's C, is compared to the common criterion in the 2016 paper. They come to the conclusion to use Harrel's C in smaller scale clinical studies and the log-rank split criterion for use in large-scale 'omics' studies.

The following sections first motivates the use of Random Forest for the medical and biological data. A short overview is given which other techniques were evaluated as an alternative to Random Forest.

After that we point out our hypothesis that Random Forest performs best for the classification of medical and biological data compared to other classifiers. Until we work our way through the series of papers about Random Forest we first give an introduction to the specific attributes of biological and medical data as well as a description of how the Random Forest works.

We then list positive pattern as a checklist for further research with Random Forest as well as a summary of bad or negative results and patterns.

After all we conclude and give a perspective on future work on medical and biological data.

2. MOTIVATION

Biological and medical data are of specific kind. They usually have a lot of features and and often few samples, compared to the number of features. This kind of data also have features that interact strongly with each other. The improving technology lets researchers collect much data that has to be analyzed. The stepping up in sequencing DNA for example reveals a lot of genetic code. The complexity of genetic diseases can be simplified by using classification models. Or the images that are obtained by imaging method like Magnetic Resonance Imaging have to be classified. The large amount of data is also to be considered for the use of computational memory and calculation power.

Random Forest is a framework that is said to perform well on this specific kind of data. It is especially known for its high precision accuracy. As said Random Forest is a framework. The motivation of this paper is to show how Random Forest is used and implemented for the specific kind of data. Different approaches and extensions are shown in this paper.

3. RELATED WORK

Genome-wide association studies have been performed on using single marker analysis. But the goal to identify the reasons for diseases caused by heritability, hasn't been reached. Biological Systems are more complex and so further mechanisms like gene-gene interaction were taken into account. But the results were still not satisfying enough. They came to the conclusion that it is a challenge to analyze and interpret the biological subject once results are received. The usage of prior knowledge about biological systems can help by this process [10]. The methods they compared, though weren't satisfying enough. Random Forest in contrast performs well as we often evaluate in this paper.

In [20] Random Forest is compared to two other deci-
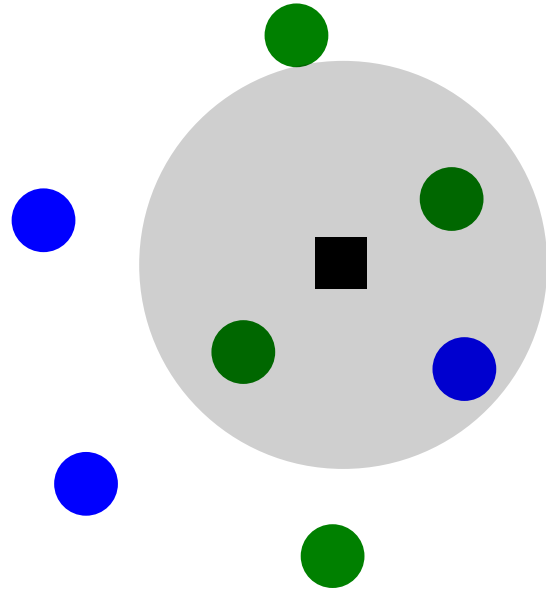
Figure 1: Illustration of the class label prediction by kNN

sion tree learners: Deterministic Forest and Classification Tree. Deterministic Forest works as follows. For building a deterministic forest, first consider all competitive splits, and construct the forest with them. The trees are limited on size and have a deterministic and reproducible manner. Compared to random forest it is more efficient to calculate. Classification Tree uses a recursive partitioning method. Its splitting criteria is the node impurity. Unfortunately they do not describe their implementation for their Classification Tree in detail, but however the results show that Random Forest performs best.

Maken et al. compare [9] Random Forest to Multiple Instance Learning and the k-Nearest Neighbors (kNN) algorithm. Multiple Instance Learning (MIL) is a variation of supervised learning. The learner receives not single instances but bags. Those are labeled positive if at least one instance of its subset is positive and negative if all instances of its subset are negative. The kNN algorothm [11] can be used for classification and regression. For a classification of a sample, kNN considers the distance of the new sample to the k nearest samples in the feature space. Figure 1 shows the qualification of the new sample, drawn as black square. In this case we consider the distance function as the Euclidian distance and the variable k is three. That means the three closest samples of the training data, drawn as circles, are taken into account. In this figure there are two green samples and one blue. So the new black square samples is classified as a sample of type green, due to the majority of green labeled samples of the three nearest.

Schmid et al [15] evaluate a new splitting criterion for Random survival forests (RSF). It is the standard approach to analyze a survival outcome using Random Forest. A RSF is a forest grown by Random Survival Trees. For those right-censored survival data is used. Like Random Trees, here a predetermined survival criterion is used as the split criterion to obtain daughter nodes. The splitting is recursively processed on each created node. The split maximizes the

survival difference of the daughter nodes.

The development of extension for the Random Forest framework still goes on. This is also shown by the 2015 paper "CloudForest: A Scalable and Efficient Random Forest Implementation for Biological Data" [2]. They still want to handle the complexity of biological data. Their extension includes methods such as dealing with unbalanced classes and missing values as shown by their table 2. The classification performance of Cloud Forest is equivalent to R and scikit-learn on the clinical dataset, but the is faster than the others. Note that Cloud Forest could handle the missing imputed to the datasets for the other implementations.

4. HYPOTHESIS

The hypothesis in the paper we studied were all about a faster implementation or a implementation that classifies their specific kind of data better than other approaches. A faster implementation is said to be done by a better use of the computer memory or a better parallelization. All in all the hypothesis is that Random Forest performs well on biological and medical data. It is said to have a high precision accuracy and that it can extract interactions between features well.

5. BACKGROUND

Data that comes from medical and biological resarches is special, because it often consists of very many features compared to their sample size. Random Forest is said to perform well on these kind of data. We first give an introduction to the data with its specialties and then fundamental overview on how Random Forest works.

5.1 Medical and Biological Data

Research in Biological topics, especially in gene-gene, protein-protein or Genome-wide interactions as well as Biological Sequence analysis or medical topics like cancer detection lead to high-dimensional data. Figure 3 for example shows the pairwise interaction of SNPs. Single nucleotide polymorphisms (SNPs) Is the most common type of genetic variation among people. A SNP represents a difference in a single DNA building block (nucleotid). If SNPs occur within a gene or in a regulatory region near a gene, it may play a more direct role in disease by affecting the gene's function [13]. On the figure, the intensity of interaction is modeled via grey scale. The darker the more intense. This is just an example of how strong interactions of features have influence on the prediction. The improving methods to collect data in those topics result in a data explosion, which makes the use of machine-learning techniques popular. Extracting meaningful information of high-dimensional data is a challenging task.

That also affects the computation time, and the convergence towards a solution. The decreasing ability of algorithms to converge by the increasing size of dimensions is called "The curse of dimensionality". The term describes the fast increasing of a volume while adding more dimensions. Data sets in life science normally have a lot more variables than samples. Figure 4 illustrated the curse of dimensionality. On the left side you see a two dimensional table. The blue square indicates the volume by using 2 dimensions and both only the range from zero to one. The volume already increases 16 times when using a range of zero to four. The
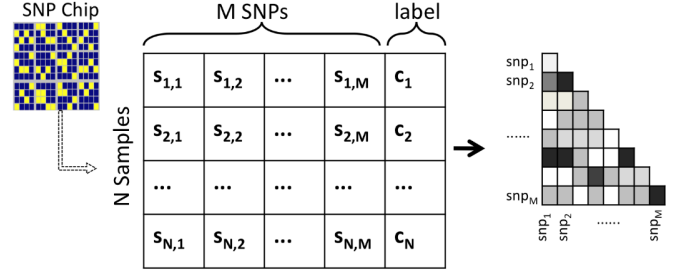
Figure 3: Illustration of feature interaction, here SNP-SNP, taken from [12]
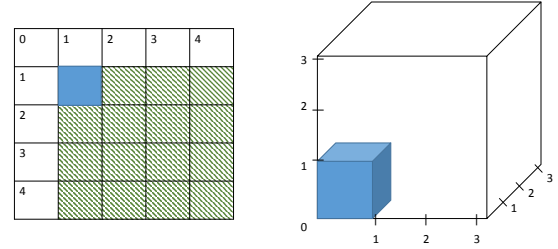
Figure 4: Illustration of the curse of dimensionality

volume even more increases by adding additional dimensions as shown by the cube in the right side. Using the range zero to three for all three dimensions increases the volume of the little blue cube 27 times.

Often in biological and medical science the sampling size is small. This in combination with many dimensions makes a computation difficult, because using so many features would require a large amount of training data. The Hughes phenomenon addresses that problem: The predictive power reduces as the dimensionality increases, using always the same number of training samples [6]. But those two are not the only issues that one has to deal with. Another phenomenon is called "confound of multimodality" [3]. This term refers to the complexity of biological systems. In biological data we often have to handle multiple interrelated biological processes, which blurs the relationship between two genes for example. Whenever needed, that problem is described specifically in this paper. At last there is also a strong correlation between the features. Interactions between variables can cause an effect, respectively variables can sometimes become important when they interact with other variable.

5.2 Random Forest

The Random Forest is very suitable to use in medical or biological data. It has not only a good prediction accuracy but also a very good interpretability compared to other machine learning algorithms [12]. Random Forest is also very fast, since it can be used concurrently.

The trees are built like follows until every feature was used.

- Randomly select d samples with replacement

Figure 2: Comparison of Cloud Forest main features in [2]

- $\sqrt{n}$ out of n features are selected for one tree.

- The features are evaluated for their ability to split the data, the one with the best ability is selected to split the set into subdatasets.

- Recursivley repeat the feature selection on the subdatasets until every leave is prune or the leaves contain a certain number of samples.

Random Forest is thus an ensemble method, the name comes from the randomly selected samples and features. Notice that the trees are not pruned afterwards. After all trees are built, the training phase is completed.

In the test phase a test sample is predicted by every tree in the forest. The final decision for the class label of the test sample goes by the majority of predicted classes.

Note that only one tree without the other of the forest is not a good decision tree, because it only handles a subset of features. But when all trees in the forest are used, you get a strong qualifier [19]. In addition there is no need for a cross-validation, because every tree already contains training and test data.

Figure 5 illustrates the building of one tree of the forest. On top F2, F4 and F5 are the randomly selected features and S3-S9 are the randomly selected samples. The randomly selected area is painted blue. F4 splits the data best, so there is a leave on the right branch containing only samples that are of class A. The left side of the branch is the rest, it is still mixed up, so we again select randomly features, here F1-F3. F2 does the best split, which results in two leaves, the one on the left side containing only samples that are of class B and on the right branch a sample of class A. Now that all leaves are not mixed up at all, the algorithm determines. Now other trees have to be built in the same way until we obtain a forest.

6. CHECKLISTS

All datasets are roughly similar structured. They all use data from biological or medical research. Some are open available and used on certain workshops like the Genetic Analysis Workshop 15. Some paper, however don't mention the origin of their dataset. In the paper about Multiple Instance Learning for Breast Cancer they split their data into only mass-like and both mass-like and non-mass-like lesions. Willows describes in his paper that he uses a binary representation of his samples, they only have four different outcomes, so they can be represented by two bits. Other datasets aren't specifically prepared for the analysis.
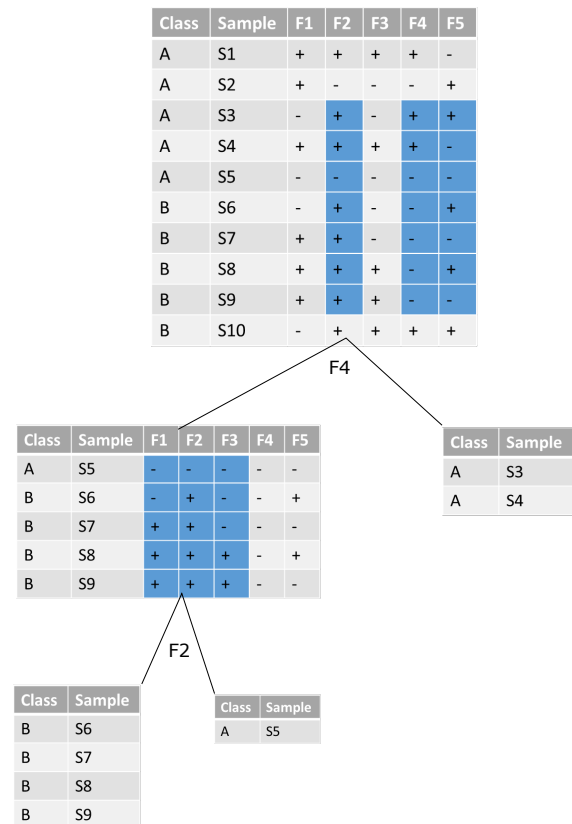
Figure 5: Building of one tree of the forest by Random Forest

Checklists are to prevent mistakes because of forgetting things. So for data science with the medical and biological context we should always know which the used algorithm is for and if it is not exactly appropriate, we should adapt it. We should keep in mind that it is complex data and therefore we need enough time to interpret the results. Due that the amount of data is large we should make safe that the environment we run our analysis on is computationally able to handle the data.

7. STUDY INSTRUMENTS

In all the papers the don't use study instruments like surveys. The data collected in medicine [18], though, normally contains a lot of data about the patient which can only be gained by asking him or letting him fill out a sheet. Features in medical data are for example

- Family history
- Patient's medical history
- Medication
- Current complaint

as well as data gained from measurements or recorded signals (EKG, EEG,...). Medical, as well as biological data, contain data like genome data which cannot be obtained by surveys. They are measured by a medical or biological procedure.

8. STATISTICAL TESTS

A statistical test provides a mechanism for making quantitative decisions to determine whether there is enough evidence to "reject" a null hypothesis. Most of the papers were about a faster performance, so it suggests itself to perform a comparison of performance. So for example in the paper "On safari to Random Jungle", they plot their results in addition which makes the result even noticeable. Figure 6 shows their bar chart.

Zhang, Wang and Chen illustrate their performance results in a table. Figure 7 shows their table used to compare run time.

In "On the use of Harrell's C for clinical risk prediction via random survival forest" Schmid, Wright and Ziegler use Boxplots of performance differences between to methods in log scale as can be seen in Figure 8.

In [16] they also introduce the proximity between samples. Those are measurements of the similarity of samples. So, outliers can also be predicted by the proximity measurements. The identification of those or of mislabeled samples can be seen as a feedback and can let the researcher take a deeper look on them.

Touw et al. compared a lot of sudies and they got to say that 66% of the studies use the "variable importance". Consider a large dataset with a lot of features. Variables that are very important for a subset can be lost when only looking at the global result. So the local variable importance can be though very important for the classification of a sample.

Zhang, Wang and Chen [20] compare values that are received across several repetition of their performed permutation to reflect the empirical distribution of importance measures and to obtain $p-values$. In addition they also use the so called "Test-Wise Error Rate (TWER)" which can be seen as the probability that the null-hypothesis is rejected, but should not be subjected and the "Family-Wise Error Rate (FWER)" which is used in multiple hypothesis testing and is the probability of getting at least one false positive test.

Bradley et al [9] use the Mean area under receiver operating characteristics curve (AUC) as the performance measurement. They also compare two AUCs arising from different algorithms using a t-test. A t-test tests the significance of a difference between two sets.

9. BASELINE RESULTS

The oldest paper we examined is "FACIL: fast and accurate genetic code inference and logo" [4]. FACIL was implemented due to a new sequencing method for DNA, which resulted in new and large datasets. Random Trees with standard parameter was used to assess the correctness of their homology-based predictions. That increased the precision and sensitivity scores. Unfortunately they do not discuss why they chose the randomForest package R.

As said, in the oldes paper they used the standard parameter. In the next two papers [16] and [20] the Random Forest is tuned to perform better.
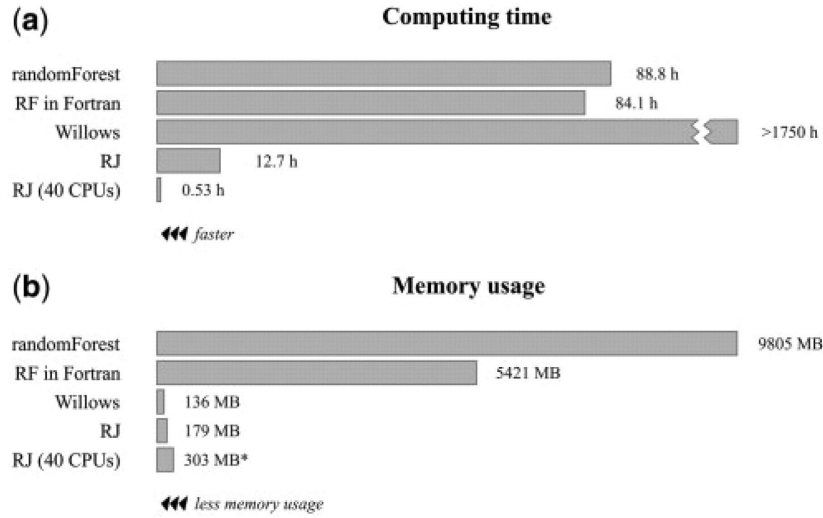
Random Jungle is said to perform up to 159 times faster than alternative implementations while still produce valid results. Figure 6 illustrate their results. The upper image shows that Random Jungle both running on one processor or 40 performs way faster than the other implementations. Random Jungle in an extension to Random Forest which can be run on multiple processors. Random Jungle can perform Random Forests on multiple processors. For that it uses multithreading and Message Passing Interface parallelization. Random Jungle also uses the "variable backward elimination". Their results for the importance scores shows equal performance compared to Random Forest. The Random Jungle package can be downloaded at http://www.randomjungle.org, which is very good, so their result are refutable.

Regarding the lower part of Figure 6 shows the memory usage. Here Willows performs best, the parallelization by Random Jungle costs a little more memory due to the parallelization demands. Willows also was implemented to run on a normal desktop computer, the authors also used a 2.33 GHz processor and 2 GB physical memory running on Microsoft Windows XP Professional Version. Willows use the recursive partitioning technique: A classification rule predicts membership and consists of a predictor and its corresponding threshold. After the splitting rule is applied on the first root, the daughter roots are split and afterwards their daughter rules - recursively. Like Random Jungle, Willlows is also available under http://c2s2.yale.edu/software/Willows, so there results can also be verified or falsified. Unfortunately their way and their algorithms are only briefly described in their paper, so it needs more than just reading the paper to understand and evaluate their results.

Figure 7 shows the performance of Random Forest in comparison to Classification tree and Deterministic Forest. Having 10 000 SNPs, Random Forest performs 10 times faster than Classification tree and 68-78 times faster than Deterministic Forest. With 100 000 SNPs the performance of Random Forest was also better than Classification tree (2.6–5 times) and approximately 7 times faster than Deterministic Forest. The main output is tree structures as shown in figure 9

The next paper "A new variable selection approach us-

Comparison of computing time and memory usage of several RF implementations.

(a) **Computing time**

randomForest — 88.8 h
RF in Fortran — 84.1 h
Willows — >1750 h
RJ — 12.7 h
RJ (40 CPUs) — 0.53 h

◀◀◀ *faster*

(b) **Memory usage**

randomForest — 9805 MB
RF in Fortran — 5421 MB
Willows — 136 MB
RJ — 179 MB
RJ (40 CPUs) — 303 MB*

◀◀◀ *less memory usage*

Daniel F. Schwarz et al. Bioinformatics 2010;26:1752-1758

Bioinformatics

Figure 6: Comparison of computing time and memory usage, in [16]

Table 1: Run time (in seconds) of the operations.

	SNPs	Classification tree	Random forest	Deterministic forest
Compressed	10, 000	70	7	480
Uncompressed	10, 000	68	6	470
Compressed	100, 000	257	96	673
Uncompressed	100, 000	249	94	640

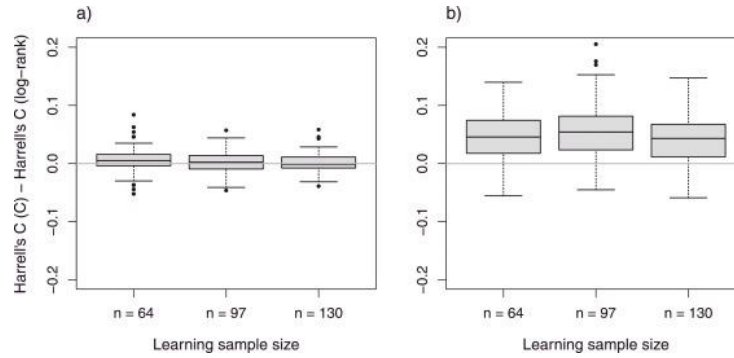Figure 7: Comparison of run time [20]

Fig. 7. Boxplots of the performance differences between C-based and log-rank-based splitting in RSF for the breast cancer data (Desmedt et al., 2007). For all learning sample sizes, predictive performance was measured by calculating Harrell's C from test da...

Matthias Schmid, Marvin N. Wright, Andreas Ziegler

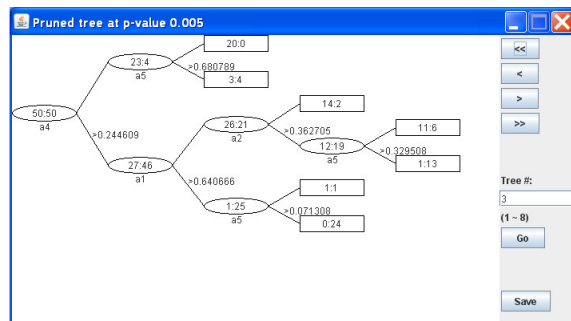**Figure 8: Comparison performance using boxplots in [15]**

Figure 1
Tree structure.

Figure 9: Sample resulting tree, [20]

ing Random Forest" [5] is beat-off in this series of different Random Forest implementations and extensions. This paper is about a new approach to select variables. Selection variables on a higher point of view is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. They did an extensive review on variable selection methods and came to the conclusion that they wanted to develop a new one. Their new approach is based on the theoretical framework of permutation tests and meets important statistical properties, and can be applied to regression and classification problems. They took four famous datasets so that they are accessible to everyone who wants to reproduce the results in the future. They selected four data sets, two regression datasets and two classification datasets. They datasets were chosen such that they had differing number of variables and observations. At the end

they came to the conclusion that to prove generalizability, the benefit of its application, concerning prediction accuracy and especially the selection of relevant variables, has to be further investigated. Further investigations need to be done to explore factors like correlation strength, block size, interactions, type of importance measure, definition of 'relevance' and kind of alpha adjustment which were identified to affect variable selection. They describe their methods very detailed so that their measurements can be comprehended. This is an improvement to the older papers.

The authors of the 2014 paper "Multiple instance learning for breast cancer magnetic resonance imaging" [9] compare Random Forest against citation-kNN and tile-kNN, and investigate how Multiple Instance Learning performs in comparison. For that they use utilize both tile-based and region-of-interest (ROI) based features.

For the tile-based features a subimage is seen taken like one bag of feature vector, which are extracted from one individual tile. Due to the size, classification performance is not affected by the accuracy of the segmented regions. A region-of-interest (ROI) based feature is a part of samples that have been specially selected due to a specific purpose. In this paper, based on the data each lesion becomes a labelled instance in the dataset. The features extracted from each lesion are then individually labelled as either benign (negative) or malignant (positive) [9].

The results show that citation-kNN works better than tile-based kNN. The performance of ckNN is the same as ROI-based SIL. They say ckNN is a suitable choice, in addition they figured out that it could be used as screening tool under certain circumstances. But apparently, the best mean AUC which is there measurement is obtained with Random Forest. The better classification performance of Random Forest
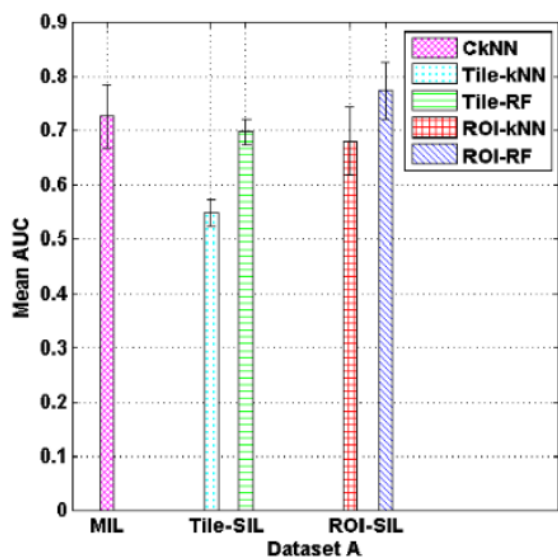
Figure 10: Comparison of results, [9]

comes from the robustness of Random Forest and its variable selection method. Figure 10 shows that a difference between citation-kNN and tile-based kNN is significant, while the difference between citation-kNN and tile-based single instance RF is not statistically significant. All in all Random Forest performs best, but not significantly.

In this paper the used software is referenced, and the parameters they used are mentioned for every study. This makes the results refutable. The dataset is taken from the MRI database, so the whole paper can be researched.

The youngest paper in our series is "On the use of Harrell's C for clinical risk prediction via random survival forests" [15]. Here Harrell's C is proposed as the split criterion. They perform their measurements on Random Survival Forest. All in all they recognize that Harrell's C performs well as a split criterion for smaller studies, while the log-rank splitting better performs in lager-scale data. They don't deliver the software package, the algorithm is described in pseudocode, though. They appended however the raw data. In addition a detailed theoretical description on all used methods is given.

10. PATTERNS / NEGATIVE PATTERNS / NEGATIVE RESULTS

In the FACIL paper the second step of the algorithm sometimes performs very badly. The homology-based prediction caused a very low precision (24%) for stop codons. But unfortunately they don't even mention why they use the randomForest R package.

A negative result of Random Jungle occurs when running the Random Jungle software to perform random forests on a multi-core processor, it consumes more memory than when run on a single-core processor. This is because helping data structures have to be provided for every core of the processor. They lack in addition in the explanation how Random Jungle technically works.

The use of Willows reduces the amount of needed memory but this decreases its speed. So an improvement of the memory usage brings a negative effect on the performance. But

in this paper there is less theoretical background explained, a implementation of their algorithm only out of the paper wouldn't work.

The paper about the a new variable selection approach lacks in where this method will not work should be explained. Furthermore why it will not work in certain situations could also be explained.

Lausch et al's paper also shows the following antipattern: The discussion is very big. It could be more precise and they did not explain why they chose the data sets the chose.

Only two paper compare their results to other methods than Random Forest.

11. NEW RESULTS

We recognized a lot of new results as well as the verification that Random Forest really works better in most cases, but always as least as good as other classifiers on medical and biological data. No other classifier could perform better than Random Forest. New result were the highly increased speed of Random Jungle, which is an extension to Random Forest and the more efficient usage of computer memory shown by Willows. A different approach for a variable selection could also not perform better than Random Forest. Another extension of Random Forest, CloudForest enables Random Forest to deal with missing values. CloudForest also speeds up the performance of Random Forest.

12. FUTURE WORK

The biological and medical data is very complex. Almost all methods need to be verified more that they were in the paper mentioned. The particular problem of those data is that the features correlate highly with each other. That means that an approach that splitting on one feature is sufficient is not accurate. A proper analysis of medical and biological systems also needs a lot of samples, since there are so many features and it is affected by the course of dimensionality. In addition biological and medical creatures always have certain unique attributes. So that is also to be considered. The 2005 paper [17] still claims those problems. There is still need for techniques to analyze the multiple layer of biological data at once. The top goal, to understand a biological or medical process as a whole is still not reached. Therefore it would need a technique to ensemble different types of medical data like images and measurements to explore the data and combine them for a prediction.

13. REFERENCES

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

[2] R. Bressler, R. B. Kreisberg, B. Bernard, J. E. Niederhuber, J. G. Vockley, I. Shmulevich, and T. A. Knijnenburg. Cloudforest: A scalable and efficient random forest implementation for biological data. *PLOS ONE*, 10:1–7, 12 2015.

[3] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37–49, 2008.

[4] B. E. Dutilh, R. Jurgelenaite, R. Szklarczyk, S. A. van Hijum, H. R. Harhangi, M. Schmid, B. de Wild, H. G. Stunnenberg, M. Strous, M. S. Jetten, et al. Facil: fast

and accurate genetic code inference and logo. *Bioinformatics*, 27(14):1929–1933, 2011.

[5] A. Hapfelmeier and K. Ulm. A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, 60:50–69, 2013.

[6] G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, Jan 1968.

[7] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

[8] A. Lausch, A. Schmidt, and L. Tischendorf. Data mining and linked open data–new perspectives for data analysis in environmental research. *Ecological Modelling*, 295:5–17, 2015.

[9] F. A. Maken, Y. Gal, D. McClymont, and A. P. Bradley. Multiple instance learning for breast cancer magnetic resonance imaging. In *Digital lmage Computing: Techniques and Applications (DlCTA), 2014 International Conference on*, pages 1–8. IEEE, 2014.

[10] J. H. Moore, F. W. Asselbergs, and S. M. Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, 2010.

[11] L. E. Peterson. K-nearest neighbor. 4(2):1883, 2009.

[12] Y. Qi. Random forest for bioinformatics. In *Ensemble machine learning*, pages 307–323. Springer, 2012.

[13] G. H. Reference. What are single nucleotide polymorphisms, 2016.

[14] M. Robillard, W. Maalej, R. Walker, and T. Zimmermann. *Recommendation Systems in Software Engineering*. Springer, New York, 2014.

[15] M. Schmid, M. N. Wright, and A. Ziegler. On the use of harrellâĂŹs c for clinical risk prediction via random survival forests. *Expert Systems with Applications*, 63:450 – 459, 2016.

[16] D. F. Schwarz, I. R. König, and A. Ziegler. On safari to random jungle: A fast implementation of random forests for high dimensional data. *Bioinformatics*, page btq257, 2010.

[17] W. Sturm, T. Schreck, A. Holzinger, and T. Ullrich. Discovering medical knowledge using visual analytics: A survey on methods for systems biology and *-omics data. In *Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine*, VCBM '15, pages 71–81, Aire-la-Ville, Switzerland, Switzerland, 2015. Eurographics Association.

[18] P. Szolovits. Nature of medical data, 2004.

[19] W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. van Hijum. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Briefings in bioinformatics*, page bbs034, 2012.

[20] H. Zhang, M. Wang, and X. Chen. Willows: a memory efficient tree and forest construction package. *BMC bioinformatics*, 10(1):1, 2009.