

# **Predicting Rules for Cancer Subtype Classification using Grammar-Based Genetic Programming on various Genomic Data Types**

**Thesis**

Submitted for a Doctoral Degree in Natural Sciences  
(Dr. rer. nat)

Faculty of Mathematics and Natural Sciences  
Rheinische Friedrich-Wilhelms-Universität Bonn

Submitted by

**Mario Deng**  
from Herne, Germany

Bonn 2017

Prepared with the consent of the Faculty of Mathematics and Natural Sciences at the Rheinische Friedrich-Wilhelms-Universität Bonn.

1. Reviewer: Prof. Dr. Sven Perner
  2. Reviewer: Prof. Dr. Joachim L. Schultze
- Date of examination: 29<sup>th</sup> January, 2018  
Year of Publication: 2018

## **Declaration**

I solemnly declare that the work submitted here is the result of my own investigation, except where otherwise stated. This work has not been submitted to any other University or Institute towards the partial fulfillment of any degree.

---

Mario Deng

## Danksagung

Ich bedanke mich bei allen, die mich in meinem akademischen Werdegang unterstützt haben. Ins besondere bei meinem Doktorvater Prof. Dr. Sven Perner, für die großartige Zusammenarbeit, Förderung und Forderung meiner persönlichen Leistungen, sowie das mir entgegengebrachte Vertrauen in wissenschaftlichen und privaten Belangen. Prof. Dr. Joachim L. Schultze danke ich für das Interesse an meiner Forschung, dieser Arbeit und die Unterstützung in fachlichen Fragen, sowie für die Übernahme des Zweitgutachtens. Auch bedanken möchte ich mich bei den Herrn Prof. Dr. Jürgen Bajorath und Prof. Dr. Alf Lamprecht für die Übernahme der weiteren Gutachten und ihrer Teilnahme an der Promotionskommission.

Ein ganz besonderer Dank geht an alle meine Arbeitskollegen des Universitätsklinikums Schleswig-Holstein und meine Arbeitskollegen des Universitätsklinikums Bonn. Ohne die enge Zusammenarbeit, die mir entgegengebrachte Geduld und die vielen Diskussion wäre die Anfertigung dieser Arbeit nicht möglich gewesen.

Ganz besonders bedanke ich mich bei meinen Eltern, Roswitha und Thomas für die Unterstützung während meines gesamten bisherigen Lebens und die Motivation, immer über den Tellerrand hinaus zu blicken. Bei meinen Schwestern, Vivian und Pauline, für die langen Diskussionen und das offene Ohr in den anstrengenden Phasen meines Lebens. Bei Erwin, der mir mein Studium und diese Arbeit mit ermöglicht hat, sowie meiner gesamten Familie.

Weiterhin möchte ich mich bei meinen guten Freunden bedanken, die mir während meines Studiums zur Seite standen und auch heute noch zur Seite stehen. Meinen lieben Bbr. die mir das akademische Leben und dessen Vorzüge nähergebracht haben.

Ich danke meiner Freundin, Claudia. Danke für die aufbauenden Worte, die ständige Unterstützung in Rat und Tat, sowie das Vertrauen und die Kraft auch anstrengende Abschnitte in unserem Leben zu überstehen.

# Contents

<b>1</b>	<b>Summary</b>	<b>3</b>
<b>2</b>	<b>Introduction</b>	<b>5</b>
2.1	Breast Cancer & Subtypes . . . . .	7
2.2	Prostate Cancer . . . . .	9
2.3	Modelling & Understanding Omics Data . . . . .	10
2.4	Aims of the Study . . . . .	12
2.5	Pre-Published Results . . . . .	12
<b>3</b>	<b>Materials &amp; Methods</b>	<b>13</b>
3.1	Computing Environment . . . . .	13
3.2	Test Data Sets . . . . .	14
3.3	Genomic Events . . . . .	15
3.3.1	Single Nucleotide Polymorphisms, Insertions & Deletions	15
3.3.2	Copy Number Variation & Gene Fusions . . . . .	16
3.3.3	Gene Expression . . . . .	17
3.4	Databases . . . . .	17
3.4.1	The Cancer Genome Atlas . . . . .	18
3.4.2	Firehose Pipeline . . . . .	18
3.4.3	cBioPortal . . . . .	18
3.5	Data Integration . . . . .	20
3.5.1	Background . . . . .	21
3.5.2	Implementation . . . . .	22
3.5.3	Workflow & Usage . . . . .	23
3.5.4	Data Normalization . . . . .	23
3.6	Learning from Data . . . . .	25
3.6.1	Supervised Learning . . . . .	26
3.6.2	Learning Functions . . . . .	26
3.6.3	Bias–Variance Trade-off . . . . .	27
3.6.4	Model Interpretability . . . . .	28
3.6.5	Established Models . . . . .	30

3.7	Classifier Design . . . . .	34
3.7.1	Interpretability versus Accuracy . . . . .	35
3.7.2	The Evolutionary Decision List . . . . .	36
3.7.3	Precision & Confidence Intervals . . . . .	37
3.7.4	Binary Data Representation . . . . .	39
3.8	Performance Assessment . . . . .	40
3.8.1	Classifier Performance . . . . .	40
3.8.2	K-Fold Cross Validation . . . . .	41
<b>4</b>	<b>Results</b>	<b>45</b>
4.1	Test Data . . . . .	45
4.1.1	The Tic Tac Toe data . . . . .	46
4.1.2	The Titanic data . . . . .	46
4.1.3	The Mushrooms data . . . . .	48
4.1.4	The Cars database . . . . .	50
4.1.5	Summary . . . . .	50
4.2	FirebrowseR + Web-TCGA = Data Foundation . . . . .	53
4.2.1	Mutational Data . . . . .	53
4.2.2	Expression Data . . . . .	53
4.2.3	Copy Number Variation Data . . . . .	55
4.3	Breast Cancer Subtyping . . . . .	55
4.3.1	Assembling a Cohort . . . . .	55
4.3.2	Classifying Breast Cancer Subtypes . . . . .	58
4.4	Prostate Cancer Subtyping . . . . .	61
4.4.1	Designing a Cohort . . . . .	62
4.4.2	Classifying Primary & mCRPC Samples . . . . .	64
<b>5</b>	<b>Discussion</b>	<b>69</b>
5.1	Data Aggregation & Normalization . . . . .	69
5.2	Evolutionary Decision List . . . . .	71
5.3	Breast Cancer Findings . . . . .	72
5.4	Prostate Cancer Findings . . . . .	73
<b>6</b>	<b>Conclusion</b>	<b>77</b>
	<b>Bibliography</b>	<b>79</b>
	<b>Abbreviations</b>	<b>95</b>
	<b>Glossary</b>	<b>101</b>

<b>A Appendix</b>	<b>107</b>
A.1 Test Data Decision List . . . . .	107
A.1.1 The Tic Tac Toe Decision List . . . . .	107
A.1.2 The Titanic Decision List . . . . .	107
A.1.3 The Mushrooms Decision List . . . . .	108
A.1.4 The Cars Database decision lists . . . . .	108
A.2 Common Altered Genes in Prostate Cancer . . . . .	108
<b>B Curriculum Vitae</b>	<b>113</b>



# List of Figures

2.1	Breast Cancer Hyperplane . . . . .	6
2.2	PAM50 Survival . . . . .	8
2.3	US Cancer Statistic . . . . .	9
3.1	Overview of structural variants . . . . .	16
3.2	Firehose/cBioPortal information flow . . . . .	19
3.3	cBioPortal data integration overview . . . . .	20
3.4	Firebrowses API root . . . . .	22
3.5	FirebrowseR Workflow . . . . .	24
3.6	Bias-Variance Trad-eoff . . . . .	28
3.7	Decision Tree Example . . . . .	32
3.8	SVM Hyperplane . . . . .	34
3.9	Evolutionary Decision List . . . . .	38
3.10	K-Fold Cross-Validataion . . . . .	43
4.1	Tic Tac Toe classification performance . . . . .	47
4.2	Titanic classification performance . . . . .	48
4.3	Mushrooms classification performance . . . . .	49
4.4	Cars Database classification performance . . . . .	51
4.5	Web-TCGA: Global mutation profile . . . . .	54
4.6	Web-TCGA: Global expression profile . . . . .	56
4.7	Web-TCGA: CNV profile . . . . .	57
4.8	Breast cancer classification performance . . . . .	59
4.9	Breast cancer decision list & graph . . . . .	61
4.10	Prostate sancer subtype inconsistency . . . . .	63
4.11	Prostate cancer classification performance . . . . .	65
4.12	Prostate cancer decision list & graph . . . . .	67



# List of Tables

2.1	Breast Cancer Subtype Patterns . . . . .	7
3.1	Decomposition of categorical predictor variables . . . . .	39
3.2	A blank confusion matrix . . . . .	40
3.3	An example for cat and dogs classification . . . . .	41
4.1	Classifier performance on Tic Tac Toe data . . . . .	47
4.2	Classifier performance on Titanic data . . . . .	49
4.3	Classifier performance on Mushrooms data . . . . .	50
4.4	Classifier performance on Cars Database . . . . .	50
4.5	Classifier performance on the breast cancer data set . . . . .	59
4.6	Classifier performance on prostate cancer . . . . .	64



# Chapter 1

## Summary

With the advent of high-throughput methods more genomic data than ever has been generated during the past decade. As these technologies remain cost intensive and not worthwhile for every research group, databases, such as the The Cancer Genome Atlas (TCGA) and Firebrowse, emerged. While these database enable the fast and free access to massive amounts of genomic data, they also embody new challenges to the research community.

This study investigates methods to obtain, normalize and process genomic data for computer aided decision making in the field of cancer subtype discovery. A new software, termed FirebrowseR is introduced, allowing the direct download of genomic data sets into the R programming environment. To pre-process the obtained data, a set of methods is introduced, enabling data type specific normalization. As a proof of principle, the Web-TCGA software is created, enabling fast data analysis.

To explore cancer subtypes a statistical model, the Evolutionary Decision List (EDL), is introduced. The newly developed method is designed to provide highly precise, yet interpretable models. The EDL is tested on well established data sets, while its performance is compared to state of the art machine learning algorithms. As a proof of principle, the EDL was run on a cohort of 1,000 breast cancer patients, where it reliably re-identified the known subtypes and automatically selected the corresponding maker genes, by which the subtypes are defined.

In addition, novel patterns of alterations in well known maker genes could be identified to distinguish primary and metastatic, castration-resistant prostate cancer (mCRPC) samples. The findings suggest that mCRPC is characterized through a unique amplification of the Androgen Receptor (*AR*), while a significant fraction of primary samples is described by a loss of heterozygosity Tumor Protein P53 (*TP53*) and Nuclear Receptor Corepressor 1 (*NCOR1*).



# Chapter 2

## Introduction

One of the first studies which combined bio-medical feature engineering, machine learning and cancer classification was published by Street, Wolberg and Mangasarian back in 1992/1995 [122, 82]. In their work Street et al digitized images of 569 Fine-needle aspirations (FNAs), 357 obtained from benign tissue, 212 from malignant tissue. Using a computer aided approach, pathologists determined ten features (shape, radius, density, etc.) for each cell. For each feature they calculated its mean, maximum and standard deviation (SD). Using these 30 features a predictive linear model for tissue detection with an accuracy of 97% could be generated. This model is shown in figure 2.1 and depicts the separating hyperplane for both tissue types, based on three manually selected features. With this simple, yet powerful approach Street et al laid the foundations for predictive modelling in the field of cancer-biology. During the past 25 years much has changed, while the foundations remain identical. Still, medical and biological data is digitized. Based on that, predictive modelling, feature selecting or outlier discovery is applied. While this workflow remains intact, new technologies emerged, established technologies became more sophisticated and affordable. Nowadays an individuals genome can be characterized on several levels, whether it is the detection of mutations to the genome or the measurement of the genes' activity. Additionally, these information are made publicly available over the internet, adding value to the scientific community. While the amount of data, whether is self generated or obtained over the network, increased drastically and investigated issues became more and more complex, new methods for data processing and analyses are needed.

This study investigates methods to share, process and analyse state of the art genomic data by the means of predictive modelling, with the aim of cancer subtype classification.

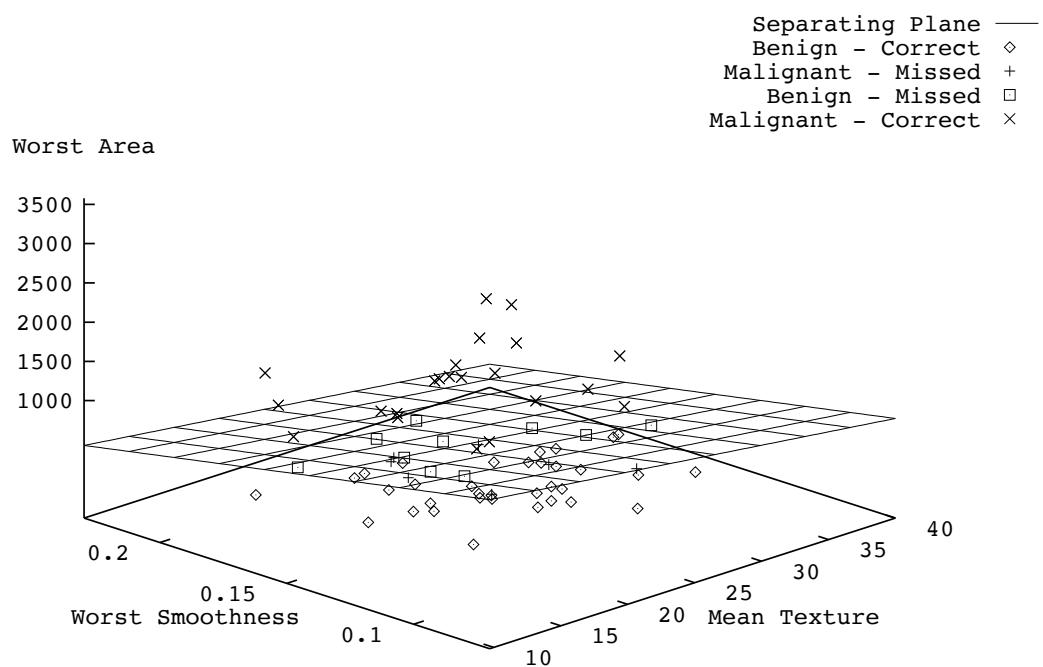


Figure 2.1: Based on the three features (area, smoothness and texture), a hyperplane separates benign from malignant tissue. Figure adapted from Street et al [122].

Table 2.1: The three marker genes which are used to define breast cancer subtypes. Table compiled from St. Gallen Criteria Catalog [51].

Subtype	Clinico-pathological definition	Therapy
Luminal A	ER <sup>+</sup> and/or PgR <sup>+</sup> , Her2 <sup>-</sup>	Endocrine therapy
Luminal B	ER <sup>+</sup> and/or PgR <sup>+</sup> , Her2 <sup>+/−</sup>	Endocrine therapy + Chemo therapy (+ Anti Her2 therapy)
Her2	ER <sup>-</sup> and/or PgR <sup>-</sup> , Her2 <sup>+</sup>	Chemo therapy + Anti Her2 therapy
Basal-like	ER <sup>-</sup> , PgR <sup>-</sup> , Her2 <sup>-</sup>	Chemo therapy
normal-like	No unique pattern	Endocrine therapy + Chemo therapy

## 2.1 Breast Cancer & Subtypes

With 246,660 new cases every year in the United States (US) alone, breast cancer is the most common cancer affecting 29% of all female cancer patients [116] (a short summary of cancer statistic is given in figure 2.3). While the term “breast cancer” is synonym with a tumor to the mammary gland, the disease characterizes through a heterogeneous profile of molecular alterations, cellular composition, and clinical outcome, allowing a classification into distinct subtypes. Despite from intuitive markers like tumor size, lymph node status, age, grade, three molecular markers are considered. Namely Estrogen Receptor 1 (*ESR1*), Progesterone Receptor (*PgR*) and Erb-B2 Receptor Tyrosine Kinase 2 (*ERBB2*) [57, 22]. The status of these markers is used to define the molecular subtypes, namely Luminal A, Luminal B, Basal-like, Her2 and normal-like. An overview which pattern of activation results in which subtype is given in table 2.1, where <sup>+/−</sup> indicate whether a gene is found positive (overexpressed) or negative (underexpressed). It is to say, that the criteria given in the table are based on observations, which are found to be statistically relevant. Hence, these patterns are observed frequently, but do not represent every single sample. These groups were initially identified by Perou et al [100]. Several years later, Parker et al [96] provided relapse-free survival estimates for each of the subtypes. As shown in figure 2.2, tumors with an enrichment of Her2 show the worst outcome, while Luminal B and Basal-like tumors show a slightly better prognosis. The best prognosis for relapse free survival has the Luminal A subtype. Additionally, Parker et al identified 50 genes, by which patterns of expression a more precise sample to subtype assignment could be achieved. This list of genes is called Prediction Analysis of Microarray (PAM) 50 and was the starting point for molecular subtyping of breast and other cancer entities. Based on these findings, several other studies emerged over the past years, identifying relations between mutations, Copy Number Variations (CNV) and the gene expression status [92, 2, 27]. Hence, the identification of *ERB2* as potential target for Her2<sup>+</sup> patients, the expression of Cyclin B1 (*CCNB1*) as marker to distinguish Lu-

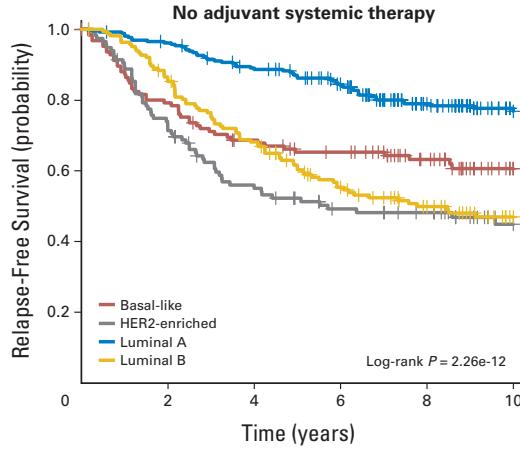


Figure 2.2: Kaplan-Meier plot for each breast cancer subtype defined by Parker et al, showing the relapse free survival probability. Figure adapted from Parker et al [96].

minal A from Luminal B samples and the Luminal A specific mutations in GATA Binding Protein 3 (*GATA3*), Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (*PIK3CA*) and Mitogen-Activated Protein Kinase Kinase Kinase 1 (*MAP3K1*) [15, 14, 77, 50].

Compared to other cancer entities, such as prostate cancer, the molecular, transcriptomic and genomic profiles of breast cancer are understood relatively well. The initially identified subtypes for breast cancer could be validated independently by other research groups [27, 93]. Additionally, more and more knowledge regarding each subtype was generated, which led to the identification of additional marker genes, influencing the genomic and transcriptomic machinery and yielding the development of a specific subtype. Also, in the light of disease treatment, the identification of potential therapeutic targets became a major advantage, as the treatment can happen personalized. These findings led to the current state of the art therapy forms, showing less side effects and afflictions.

In this study, breast cancer and its corresponding subtypes are used for model evaluation. The newly developed EDL model will be tested, despite from other data sets, on breast cancer data, with the aim to re-identify its subtypes and marker genes. Therefore, breast cancer data is used to provide a proof of principle, before the model is evaluated on prostate cancer.

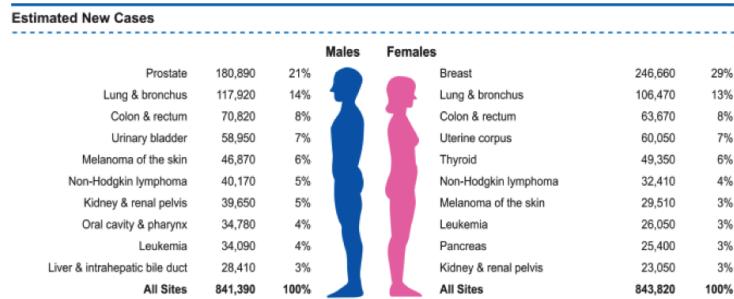


Figure 2.3: An overview of detected cancer cases in the US in the year 2013. Most prominent are breast and prostate cancer, for female or male patients. Figure adapted from Siegel et al [115].

## 2.2 Prostate Cancer

With one quarter, 180,890, of all detected cancer cases, prostate cancer is the most common cancer in the male population [115] (US alone). While breast cancer is the most common cancer in woman, which is rarely detected in men, prostate cancer is specific to men (for a ratio overview see figure 2.3). As the primary disease is asymptomatic in its early stage and unlike to cause complaints, prostate cancer is often detected in an advanced stage only. Often, in this advanced stage, metastasis have already formed in the lymph nodes and/or bone, lowering the chances of cure drastically [125]. As prostate cancer, in its early stage, does not cause any symptoms, it remains hard to detect. Therefore medical check-ups are offered to men of age 50 and older. During these check-ups the prostate is examined by touch, ultrasound and/or prostate-specific antigen (PSA) screening. While the touch examination is likely to miss tumors located at the organs front, the ultrasound examination is only capable of detecting tumors of size 10 millimeter (mm) or bigger, while smaller tumors are only detected with a probability of 20% [71]. For PSA screening, the concentration of PSA within the blood is determined. While the level of concentration is intended to be used as an indicator for prostate cancer, there exists no coherent approach for examination, as the PSA level varies vastly from individual to individual [58]. As a consequence, a high PSA level could be an indicator of prostate cancer or just be an artefact, brought about urinary retention or infection [83, 21]. While the medical check-up is just capable of detecting a primary tumor to the prostate, not yielding any information about its state, it remains unclear how the tumor will develop. While a primary tumor, detected at an old age, may be just observed without any therapy (watchful waiting), other tumors tend to be very aggressive. These aggressive tumors yield a lethal disease progress and

are known as mCRPC, as they do not show any reaction to the reduction of male sex hormones through castration or drug usage [55]. While the treatment of an early stage tumor is simple and promising, treatment of mCRPC remains an ongoing challenge [44, 43, 30].

Therefore, the identification of genomic, transcriptomic or molecular markers for the early detection and classification of prostate cancer is a crucial task, as current risk stratification systems do not provide sufficient results [24, 28, 67]. Recently, just as for breast cancer, several studies revealed correlations between genomic alterations, copy number changes and the expected disease outcome [133, 128, 125, 101, 3]. The most frequent alteration to the genome, found in 40-50% of all samples, is a gene fusion of Transmembrane Protease, Serine 2 (*TMRSS2*) and ERG, ETS Transcription Factor (*ERG*) [97, 129]. Nevertheless, the occurrence of this alteration does not seem to influence metastatic formation [93]. On the other hand side, there exists a wide variety of structural alterations such CNV, single nucleotide variant (SNP) and other copy number changes, which are observed in mCRPC, but not in primary prostate cancer [93, 107]. Highlighting the need for early stage development markers.

## 2.3 Modelling & Understanding Omics Data

The term omics functions as a proxy for proteomics, genomics and transcriptomics and describes, in the context of this study, the pooled data types. The term has been around since the breakthrough of high-throughput technologies. These high-throughput technologies transformed biology and medicine from a relative data poor discipline, into a field where massive amounts of data are generated on a daily basis. This led to several problems, ranging from the initial batch effect corrections, over the primary analyses, up to storage and distribution of these large files [64]. Not only, that this data needs to be stored, it also has to undergo a pipeline of processing steps to generate useful information. This complete processing pipeline is based in assumptions, approximations and models. Starting with the sequencing machine, which digitizes deoxyribonucleic acid (DNA) sequences based on colored cells by photography, over the mapping algorithm, which aligns reads to the most promising position in a reference genome (which is also just an approximation [110]) or the statistical model, trying to infer which alterations lead which disease type.

It is obvious that each step of this processing embodies an area of studies on its own. Therefore, at a certain point, the given information have to be accepted as gold standard, on which basis additional studies can be built

upon. With TCGA the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) booted a project to improve the ability to prevent cancer through a better understanding of the genetic basis of this disease, based on high-throughput genome analysis techniques. The TCGA collects data sets from several research projects of their partner institutions and provides the aggregated cohorts to the research community. It is ensured that the data is accessible to any researcher world wide, offering a whole new resource to scientists. While over the past decades the access to large cohorts required the seizure of pathological archives, high-throughput facilities and an interdisciplinary team, these cohorts can be accessed simply over the network.

Not only that this led to a better understanding of the diseases, it also yielded new and more sophisticated methods, which would not have been developed without such projects [87, 114, 140, 27]. A major role in this area is taken by machine learning or statistical modelling, where, under certain framework conditions, a model is fitted to data. The model is then used to infer and uncover additional information about the data, which are not obvious at the first glance. This attempt can be used to address a wide variety of issues. For example, models might be used to make predictions, to identify pattern in data or to unveil groups of samples. Each time a model is fitted, its eventual purpose should be known and declared upfront. Hence, it seems obvious that complex models<sup>1</sup> are capable of fitting complex relations in data, while more simple models can only fit aspects to a certain degree. Examples for both ends might be a neural network and a linear regression model. While the neural network is capable to even decipher complex structures within the data (hand writing recognition [72], for example), the linear regression model simply estimates two parameters, estimate and slope, not allowing such a detailed classification. While complex models seem superior at the first glance, they have an essential drawback, which is often overseen beforehand. Interpretability. As these complex models might be able to fit the data almost perfectly, they remain nearly impossible to infer and interpret. Vice versa, simple models might not gain competitive, yet acceptable performance, while unveiling true relations within the data to the user. Therefore, a trade-off between complexity, precision and interpretability should be found, allowing a precise classification while remaining interpretable. This way true coherences within groups of data can be identified.

---

<sup>1</sup>For simplicity, models with more parameters and degrees of freedom.

## 2.4 Aims of the Study

Certain cancer entities can be subdivided into several subtypes, where each of those subtypes shows a different life expectancy and therapy response. While such classifications used to be based on symptoms, phenotypes and progression, they are now investigated by alterations on the molecular level. For some cancer entities patterns of alterations could be identified, enabling an assignment for a single sample to one of the known subtypes. One of those cancer entities is breast cancer, where, depending on the identified pattern, a therapy is chosen. For prostate cancer, the course of disease shows two extremes. First, patients do not require any therapy, living a complaint free live. Second, the tumor forms metastasis, afflicting the bones and lymph node system. To identify novel driving alterations, leading to one of such extremes is the goal of this work.

This goal can be partitioned into three tasks,

1. the aggregation of omics data sets, enabling investigations on the genomic, transcriptomic and molecular level,
2. the normalization and representation of these data sets, such that they are can be inspected by a statistical model and
3. the development of a human interpretable statistical model, allowing a precise subtype assignment.

## 2.5 Pre-Published Results

Parts of this thesis have already been published in peer-reviewed international scientific journals. All paragraphs, graphics, tables, etc., where this is the case, are cited as appropriate. In addition an overview is given at this place.

- Section 3.5, including figures 3.4 and 3.5, see [31]. *Database - The Journal of Biological Databases and Curation. IF: 2.627*
- Section 4.2, including figures 4.5, 4.6 and 4.7, see [32]. *BMC Bioinformatics. IF: 2.435*
- Section 4.2, including figures 4.5, 4.6 and 4.7, has been presented at the *useR2016*, the official conference for the R programming environment, at the Stanford University.

# Chapter 3

## Materials & Methods

In this chapter the computing environment and used software tools are introduced. As the newly developed model, EDL, has to undergo performance tests, four test data sets are introduced for benchmarking in 3.2. As the eventual analyses of cancer subtypes is based on alterations of the human genome, all inspected types of alterations are introduced in 3.3. Databases, from which the cohorts for analysis are obtained, are introduced in 3.4. Required pre-processing steps the data has to undergo before analyses are then discussed in 3.5. Afterwards basic concepts machine learning are introduced and established machine learning methods are presented 3.6, which are used for performance comparison. Finally, the newly designed classifier is introduced in 3.7 along with the corresponding metrics, allowing a reliable comparison between the newly developed classifier and established models 3.8.

### 3.1 Computing Environment

All calculations produced in the context of this study are carried out by the R Programming Environment, version 3.3.2 - *Sincere Pumpkin Patch*. R is an open source programming language and software environment, provided by the R Foundation for Statistical Computing [105], offering a wide variety of extensions for statistical modelling, plotting and high performance computing. Besides the functions provided within the R core distribution, additional packages have been used. For plotting, Hadley Wickham's ggplot2 [136] library (version 2.2.1) has been utilized and C++ (version 11) extensions have been coupled to R by using Dirk Eddelbuettel's and Romain Francois' Rcpp extension [38] version (0.12.8), while Apple LLVM (version 800.0.42.1) has been used for compilation. Transferring calculations from the

R to the C++ environment drastically enhances the computing performance, as the source code is translated to byte code first. While the execution of R source code is done one a higher level, leaving type declarations undefined until computation. For more details, the reader is referred to Eddelbuettels Rcpp integration guide [37]. Other packages used for statistical modelling etc. are cited at the appropriate position.

## 3.2 Test Data Sets

As one task of this work is the development of a statistical classifier, test data sets are required to determine the models performance in comparison to already established models. For this task, four well known data sets have been chosen. The choice of those data sets has been made with regard to the final tasks of feature selection and decision making for cancer subtypes. Therefore, all test sets have identical quantities of instances and predictor variables, compared to the cancer data sets, allowing an approximation of the inter-rater reliability for cancer models. Also, non of the data sets chosen is *trivial*, that is that the label can be determined by a single predictor.

**The Tic Tac Toe data** encodes all possible board configurations (958) of the Tic Tac Toe game. Each configuration is represented by a combination of the nine fields, where each field can take the values  $x$ ,  $o$  and  $b$ , indicating if the filed is taken by a player or blank ( $b$ ). The label to predict is *TRUE* or *FALSE*, encoding if player  $x$  has won or not.

**The Titanic data** compiled by the British Government [53] provides information on the fate of passengers who traveled on the first and only voyage of the Titanic ocean liner. Recorded parameters are class, sex and age, where the label to predict is the survival. Overall 2,201 records exist, which are part of the R core package.

**The Mushrooms data** represents 8,124 different mushrooms by 22 attributes. The label to predict is if a mushroom is whether edible or poisonous. Unknown or not recommended edibility has been as encoded as poisonous as well. The data is extracted from National Audubon Society Field Guide to North American Mushrooms [120].

**The Cars database** was generated by Vladislav and Bohanec [8] in 1988 and represents a decision model, that predicts the acceptability of a car

by the customers. The label to predict can take the four values of unaccepted (“unacc”), accepted (“acc”), good acceptance (“good”) and very good acceptance (“vgood”), denoting the cars market acceptability. Each car is described by six attributes: buying prices (“buying”), maintenance effort (“maint”), number of doors (“doors”), numbers of seats (“persons”), storage space (“lug-boot”) and “safety”.

The introduced data sets serve as an ideal foundation to test a newly developed classifier for cancer subtype discovery. That is, all data sets have at least as many records as the assembled cohorts, also they come with an identical amount of attributes. Further, the Cars and Mushrooms data sets harbour a mixture of continuous, categorical and binary attributes, which is also the case for the cancer data sets.

### 3.3 Genomic Events

The foundations for carcinogenesis and cancer progression are alterations to the cascade of transcription and translation. Small changes to the DNA, mostly caused by environmental factors, effect the organisms and disrupt the cell cycle. While some alterations result in the dysfunctionality of tumor suppressor genes, others promote the hyperfunction of oncogenes, which has the potential to cause cancer. These malfunctions are caused by a wide variety of alterations to the genome. In the following an introduction of the investigated alterations in this study is given.

#### 3.3.1 Single Nucleotide Polymorphisms, Insertions & Deletions

Next Generation Sequencing (NGS) has enabled the study of the complete human genome, exome and transcriptome, unlike earlier methods, which only allowed the study of selected areas of an organisms’ genotype. Next generation sequencing led to an exponential growth of sequencing productivity, resulting in fast and cheap ways to analyze DNA sequences. Regardless of the underlying sequencing technology, the basic workflow for analysing NGS data remains identical, as each sequencing facility provides reads as output. i) An alignment is performed, where short reads are arranged to the most identical part of a reference genome [52]. ii) Mutations, such as SNPs (see figure 3.1 A) and insertions/deletions (INDELS), between the aligned reads and the reference genome are identified. Mutations may affect the translation, causing a malformed or dysfunctional protein. Affected tumor suppressor genes, can not fulfil their initial function anymore, often resulting

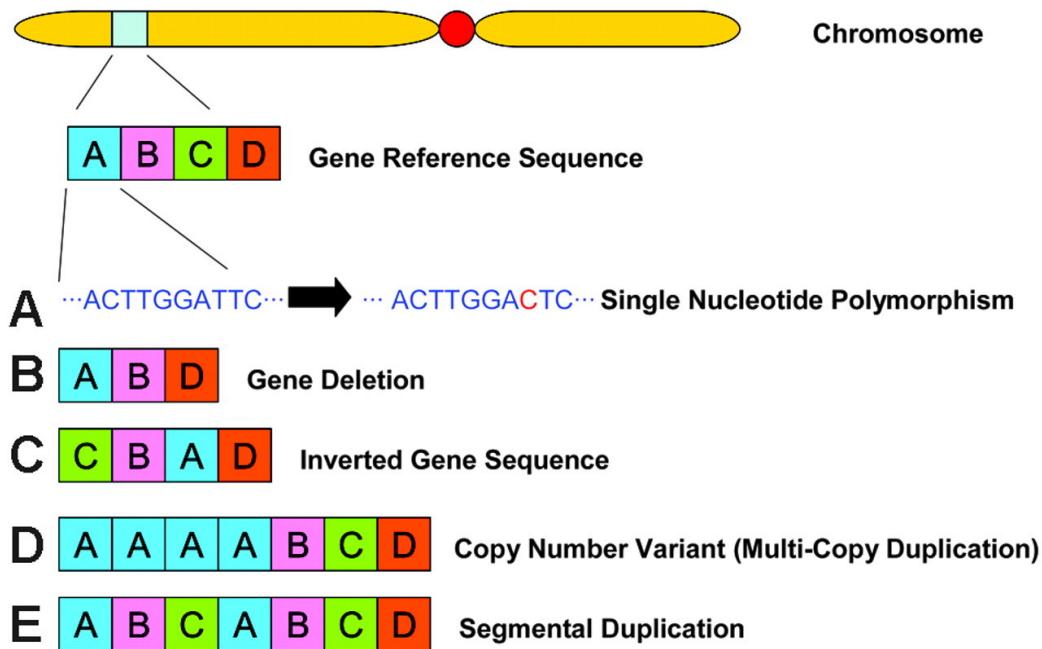


Figure 3.1: Overview of structural variants occurring A) single nucleotide polymorphism B) deletion of a complete gene C) partial inversion of gene sequence D) copy number variation / multiple copies are present E) duplication of a segment of multiple genes. Figure adapted from Mullally & Ritz [90]

in an individuals cancer disease. Further, patterns of mutations and mutated genes form footprints, which are specific to certain cancer types. For example, mutations effecting *GATA3*, *PIK3CA* and *MAP3K1* are unique to the Lumal A breast cancer subtype [92]. Also, mutations and the genes they are effecting, can serve as potential therapeutical target, as the knockdown of an oncogene may recover the initial cell cycle.

### 3.3.2 Copy Number Variation & Gene Fusions

A CNV, along with SNPs and INDELs, is another known structural variant. A structural variant is classified as CNV, if its affecting more than 50 base pairs (bp) or covers complete genes (definition adapted from Zarrei et al [143]). Regularly each gene occurs exactly two times within the individuals genome. That is one copy per chromosome set. A CNV has taken place if i) one (heterozygous) or both (homozygous) of its copies are deleted (figure 3.1 B) or ii) a single gene (gain, figure 3.1 D) or a sequence of genes occurs multiple times (high level amplification, figure 3.1 E). As with SNPs

and INDELS, CNVs can have phenotypic effects on the organism. While an increase of copie numbers can result in an increased amount of protein, heterozygous and homozygous deletions can result in a decreased or total absence of protein.

Another event is the gene fusion, which leads to a new gene out of two previously separated ones. This event can take place if i) two chromosomes are translocated, ii) a segment of the whole chromosome (and not only the gene) is deleted and iii) the chromosomes inversion. Generally, this takes place if two parts of two genes are arranged next to each other and the resulting amino acid sequences lays in between a promoter region and stop codon. Often the resulting product can produce a more active abnormal protein, causing tumor formation [39]. In the case of prostate cancer, > 50% of patients showing a overexpression of the oncogenes ETS Variant 1 (*ETV1*) and *ERG*, a gene fusion between one of the genes and *TMPRSS2* can be found [129].

### 3.3.3 Gene Expression

While NGS, CNV and fusion analysis provide an organisms footprint on the lowest level, the genotype, the gene expression analyses determines how the genotype is expressed into its final form, the phenotype. While older methods, such as gene expression chips, required a target for each gene on the chip, newly established methods make use of NGS. Here the Ribonucleic acid (RNA) is sequenced and aligned to the reference, afterwards the frequency of reads which bound to a certain transcript is determined. The final gene expression can then be determined by the genes transcripts expression. To fulfil this task, a wide variety of approaches exists. An overview is provided by Teng et al [35]. The foundation of gene expression is made on genomic level, where influences derived from CNVs, SNPs or INDELS can have an effect on how a gene is regulated and expressed into its protein. The gene expression analyses is the consequential next step after sequence analysis, as the latter describes the building blocks and the former the building blocks final product.

## 3.4 Databases

The exponential growth of sequencing data, affordable IT infrastructure and the revolution of noSQL technologies lead to a new type of databases. Not that the underlying technology would have changed, furthermore it is now possible to setup, maintain and scale large public databases, with a minimum

of cost [89].

### 3.4.1 The Cancer Genome Atlas

In the field of cancer research, TCGA has been the first of such databases. Since its launch in 2005 [135], TCGA has become the biggest portal, making large scale omics data publicly available. With the aim to improve diagnoses, treatment, and prevention of cancer through a better understanding of the disease genetics, TCGA applies high-throughput genome analysis to comparative large cohorts. At time of writing TCGA stores 15,000 cases, distributing over 29 cancer entities. While TCGA processes genomic data only to a certain level, it serves as an input for other data portals, which set up their analyses pipeline on top of TCGAs.

### 3.4.2 Firehose Pipeline

The Broad Institute’s Firehose Pipeline is one of the projects which post-processes TCGA output. It is born out of the desire to systematize analyses based on data obtained from TCGA and scale the execution of pipelines for new data to come. Thereby it processes 55 terabytes of data every month, re-running each pipeline for updated data sets. While TCGA provides rudimentary results only, Firehose integrates the output from plenty of (de facto) standard tools (figure 3.2). To distribute the generated output and to make it available to the end user, the Broad Institute provides a facility called Firebrowse. Firebrowse serves as a gateway to the analytical results. Using it, researchers can collect data in a convenient way over a web interface. Additionally, Firebrowse also holds an application programming interface (API) available. This way, users can automate their processing pipelines without the need of manual adjustment. As partial results of this thesis, an R client to the Firebrowse API is presented in section 3.5.

### 3.4.3 cBioPortal

As depicted in figure 3.2 another post-processing tool, named cBioPortal, obtains its input from TCGA and Firebrowse. The cBioPortal has been published in 2012 by Cerami et al [16], describing it as portal for “visualization, analysis and download of large-scale cancer genomics data sets”. Compared to the TCGA and the Firehose Pipeline, cBioPortal offers interactive tools, which do not only allow the download of genomic data, but also the direct analyses. Using cBioPortal one can design and directly investigate the aggregated cohorts, in terms of mutations, copy number variations and

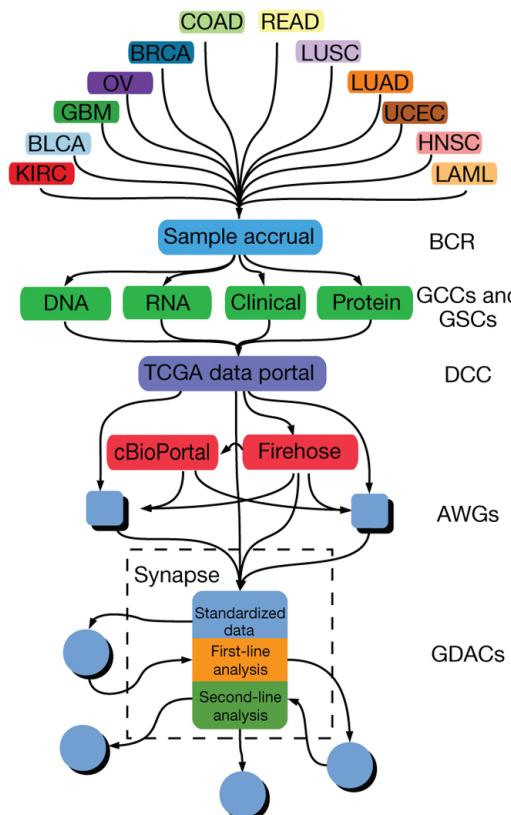


Figure 3.2: The information flow for TCGA, Firehose Pipeline and cBioPortal. It should be noted that Firehose is fed by the TCGA only, while cBioPortal also obtains data from other resources, which is not depicted in this illustration. Figure adapted from The Cancer Genome Atlas Research Network et al [134].

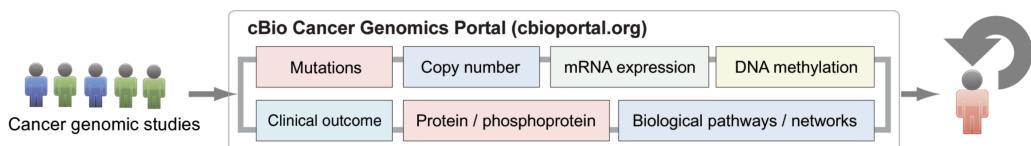


Figure 3.3: Schematic workflow of the cBioPortal: Data is collected from different studies or cohorts and the shown data types are harmonized and analyzed to be accessed by the user. Figure adapted from Cerami et al [16].

many more. Figure 3.3 depicts the available data types. Another unique characteristic is the integration of data which is not derived from TCGA / Firehose Pipeline only. cBioPortal also includes data sets and published findings from larger studies, offering an easy way of reproduction, serving as building block for own investigations. Data from both portals has been used in this thesis. For information on how the cohorts have been assembled, the reader is referred to 4.4.1, while the process of data integration is outlined in 3.5.

## 3.5 Data Integration

To integrate data into the R programming environment an R client to the Firebrowse Representational State Transfer (REST) API has been implemented. This client provides several benefits over manual downloads, as updates to the data can be obtained easily, changes to the database do not effect the data processing pipeline and let the developer focus on his task. As, during time of its development, the API was prone to changes and updates<sup>1</sup>, a workflow to automatically update, test and deploy changes to the API client has been developed. The workflow decouples changes on the server side from the client, as it automatically updates the client based on changes to the servers REST interface. This workflow is utilized to provide FirebrowseR, an R client to the Broad Institute's Firehose Pipeline (for more details the reader is referred to Deng et al [31]). As the source code is made publicly available<sup>2</sup> and transparent, both, the workflow and its deployed software product, FirebrowseR, are actively used in-house, but also by the research community. Finally, FirebrowseR became the Broad Institute's official R

<sup>1</sup>The first public beta was launched on 23<sup>rd</sup> April, 2015 and left its beta status on 2<sup>nd</sup> March, 2016.

<sup>2</sup>FirebrowseR's source code repository can be found under <https://github.com/mariodeng/FirebrowseR>.

client<sup>3</sup>.

Once the data is made available to the programming environment, additional steps for data normalization need to take place. As working with data sets obtained from TCGA and the Firehose Pipeline took a central part of this study, the Web-TCGA application has been created and published by Deng et al [32]. Web-TCGA is a graphical front end to Firehose Pipeline, enabling users to quickly inspect cohorts and obtain a brief summary. As the pre-processing methods required by Web-TCGA are identical to those in this study, the methods implemented by the Web-TCGA software form the foundation for data integration and normalization.

### 3.5.1 Background

To share information is a common task in the field of cancer research. The method of file transfer and chosen file type often strongly depend on the providers infrastructure. Data sets of low complexity are often organized as Comma Separated Values (CSV) files, as done with Variant Call Format (VCF) (see Danecek et al [29] for details on the format), or just stored as plain text file, as done with Sequence Alignment Map (SAM) format described by Li et al [80]. An alternative for storing information is provided by Database Management System (DBMS), where information is persisted in a structured way. It is the structure that reduces the data overhead when DBMS are used, as each entry is only persisted once and other occurrences are linked to that entry. Both of these methods mark an extreme at each end. While storing data in the CSV format means easy input and output to programming environment, it comes with a massive overhead of storage, as redundant information are persisted. Data stored using DBMS reduces this overhead, but makes data integration and modelling a bit of task. Further, it is almost impossible to receive or provide data to a DBMS if its structure is unknown. Also, the user needs to take updates to database or CSV structure into account, denoting a potential weak spot in the analyses pipeline.

One way to overcome these obstacles is the use of an RESTful API. While the API is the interface to an application, REST provides a framework for how the machine-machine interaction is realized. This machine-machine interaction is commonly realized over Hypertext Transfer Protocol (HTTP) verbs, defined by Berners-Lee and Fielding back in 1996 [7, 42]. If the data transport is encapsulated through an RESTful API, changes made to the database will not effect the communication, as the API remains stable. Also

---

<sup>3</sup>See press release: <https://confluence.broadinstitute.org/display/GDAC/FireBrowse+Release+Notes>.

```
{
  "apiVersion": "1.1.35 (2016-09-27 10:12:23 6a47e74011281b2aae7dc415)",
  "apis": [
    {
      "description": "Fine grained retrieval of sample-level data",
      "path": "/v1/Samples"
    },
    {
      "description": "Fine grained retrieval of analysis pipeline results",
      "path": "/v1/Analyses"
    },
    {
      "description": "Bulk retrieval of data or analysis pipeline results",
      "path": "/v1/Archives"
    },
    {
      "description": "Retrieve disease, sample, and datatype descriptions, sample counts, and more",
      "path": "/v1/Metadata"
    }
  ],
  "swaggerVersion": "1.2"
}
```

Figure 3.4: The root entry of the Firebrowse API, providing meta information, as well as sub-APIs, which can be traversed to unfold all functions provided by the API. Figure adapted from Deng et al [31].

the underlying structure of the database is completely decoupled from the communication. This allows the implementations of more advanced methods, which will not affect the usability. Applications, such as Firebrowse, realize their communication over Uniform Resource Locator (URL) queries and deliver results in a structured format, such as JavaScript Object Notation (JSON) or CSV. If the API itself receives an update, its definition changes and the client software can automatically adapt the new definition, as it is public available.

### 3.5.2 Implementation

The benefit of using an API over other technologies is, that the its definition is made available through the API itself. Hence, it can be reached from any computing environment over the network. This definition is structured in a hierachical fashion, starting from the entry point, the root. For Firebrowse, this root can be found by the following URL <http://firebrowse.org/api/api-docs/>. For convenience this definition is also depicted in figure 3.4. At the top level, three entries can be found i) “apiVersion“, ii) “apis“ and iii) “swaggerVersion“, where i) and iii) are meta information, by the software used to generate this definition. Traversing the “apis“ entry, all definitions of the API and its functions can be found. These entries provide the developer with information needed to communicate with the API, such as methods names, parameter data types and HTTP verbs used for interaction. Out of these definitions, almost all code required to build a client software can be

generated. Therefore, a blank template for R functions is created, which is completed with the information obtained from the APIs definition. For template creation Rs mustache<sup>4</sup> implementation whisker<sup>5</sup> is utilized. The template is designed in a way, that for each function provided by the API a corresponding R function is created. This function, again, interacts with centralized download manager. This has the benefit that no code is duplicated and the number of potential sites of fractures can be reduced to a minimum. Now, to combine definitions and the templates, the API is traversed and for each definition a template is completed.

### 3.5.3 Workflow & Usage

The complete workflow is a combination of free and publicly available web-services and depicted in figure 3.5. A cron-job<sup>6</sup> checks if a new API version is available. If so, the new versions source is build using the whisker templates and the new API definitions. Afterwards the new code is pushed to development branch on GitHub. The code is then tested by Travis-CI with upfront written unit tests. If an error occurs, the developer is notified, otherwise a new release is finalized by pushing the code to the master branch. The FirebrowseR package is publicly available on GitHub (master branch) and can be installed and used by anybody. After installation, data provided by Firebrowse can be downloaded directly into the R environment. Further the user can chose whether to use matrix or JSON objects, allowing maximum flexibility.

### 3.5.4 Data Normalization

Data obtained from the Firehose Pipeline is already processed to a certain level, reducing the workload for pre-processing. Nonetheless, some pre-processing is still required with regard to the follow-up analyses. All methods used for pre-processing and normalization are discussed on the example of Web-TCGA, an online platform for integrated analysis of molecular cancer data sets by Deng et al [32]. Web-TCGA has been developed as a side project of this thesis, highlighting the normalization, usage and depiction of data obtained from the Firehose Pipeline. While the first version of Web-TCGA required manual data download, the new version<sup>7</sup> utilizes FirebrowseR, making manual downloads redundant. That is possible, as both software packages

---

<sup>4</sup>See <https://mustache.github.io/> for details.

<sup>5</sup>Whisker is available via GitHub <https://github.com/edwindj/whisker>.

<sup>6</sup>The cron-job is hosted on <https://cron-job.org/>.

<sup>7</sup>Currently under development.

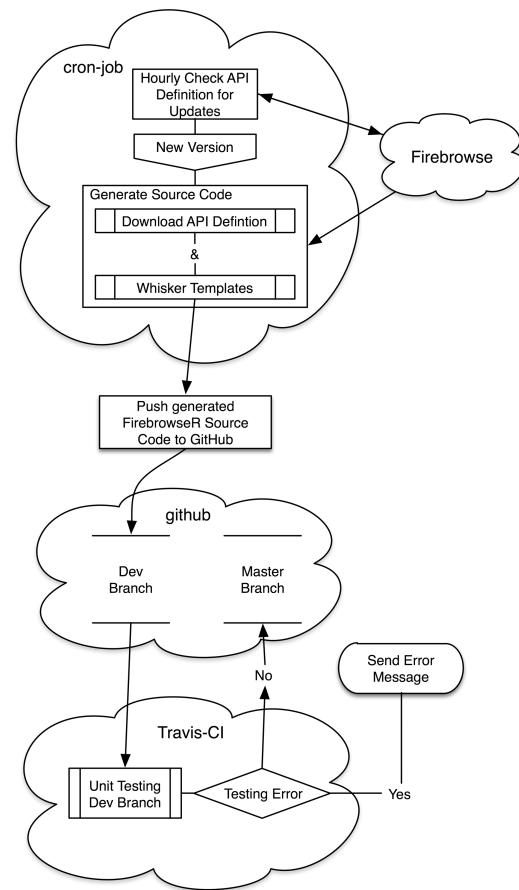


Figure 3.5: The complete workflow which is run to build a new version of FirebrowseR. The cron-job service checks for new API definitions and generates the source code for the new version, if necessary. The source code is then checked by Travis-CI and pushed to the repositories master, if not errors occur. Figure adapted from Deng et al [31].

are written in the R environment. Further, by utilizing FirebrowseR, Web-TCGAs data repository is always synchronized with the latest version from Firehose Pipeline.

Within TCGA the data are provided on different levels. Levels range from one to three, indicating an increasing state of pre-processing and data aggregation each. Raw-data only is provided on level one, the second level is characterized by canonical pre-processing or filtering (depending on the data type, see below). The third level provides data which is appropriate for analyses.

To reduce calculation time and to keep the amount of data as small as possible, data used in this study and by Web-TCGA always includes the highest data level available for each type. Somatic mutation data (level 2) and somatic CNV data (level 3, GISTIC2.0 output [86]) is directly used, as it does not require any further processing. For gene expression profiling, level 3 data is imported by FirebrowseR and processed as described below. For the expression status, two different preprocessing methods are available, namely RNA-SeqV1 (Reads Per Kilobase per Million (RPKM)) and RNA-SeqV2 (RNA-Seq by Expectation Maximization (RSEM)). Here, RNA-SeqV2 is used, which takes transcript length into account and is found to provide more accurate results [79] for downstream analysis. For RNA-SeqV2, gene expression profiles are calculated using RSEM data. Due to the lack of normal samples, the relative expression for a specific gene is calculated using its expression status in a tumor sample of a given entity, compared to its average expression status in the remaining samples of the same entity [94]. The degree of differential expression is calculated using the z-score. The z-score is defined as number of standard deviations above or below the mean of the gene's expression levels in the reference cohort.

$$Z = \frac{X - \mu}{\sigma}, \quad (3.1)$$

where  $X$  is a random variable,  $\mu$  the populations mean and  $\sigma$  its standard deviation.

Furthermore, Web-TCGA provides utilities to analyse and visualize the methylation status. For more details the reader is referred to Deng et al [32], as this data type is not used within the scope of this thesis.

## 3.6 Learning from Data

There exists a wide variety of terminology when it comes to the process generating knowledge from data. Most prominently the terms information

retrieval, data mining, machine and statistical learning should be mentioned. While information retrieval and data mining have their focus on the side of data generation and aggregation, machine and statistical learning are mainly used when it comes to building models, their interpretation and conclusions. While the main task of machine learning is to make predictions, statistical learning aims to infer conclusions from such predictions and the models used to generate them. Of course, these are fine lines and somewhat arbitrary, which will be seen when different models are discussed, but they help to frame the context of this work: The learning of information which are interpretable by a user of a certain domain.

### 3.6.1 Supervised Learning

Machine or statistical learning can be subdivided into two major disciplines, supervised and unsupervised learning. While the supervised task is to predict a measured *label* (*outcome*), based on a number of *variables* (*predictors* or *features*), unsupervised learning aims to organize data into groups without any pre-knowledge about the true label. Another distinction has to be made for supervised problems, as they can be subdivided into regression and classification problems. For regression problems, the label to predict is continuous (e.g. body height or expectancy of life), while for classification problems the label takes categorical states, such as home country or disease state. This thesis only focusses on supervised classification problems, as for all studied cases a true class label is available. A *training set*, which consists of observations and a label for various samples, is used to build a model. This model is used afterwards to make predictions of new, unseen, samples, where the true class is unknown.

### 3.6.2 Learning Functions

Any model performs a projection from the input variables  $X$  to the label  $Y$ . This projection is performed by function and it is the trainings goal to identify such a function, which minimizes the training error  $E$ . We assume  $X \in \mathbb{R}^p$  to be a real valued input vector of measurements for a single sample of  $p$  features and  $Y \in \mathbb{G}$  be the corresponding label, with the joint distribution  $Pr(X, Y)$ . If there is no error within the measurements of vector  $X$  and  $Y$  depends on  $X$ , then there exists a function  $f(\cdot)$ , such that  $f(X) = Y$ . That at hand, the goal of any supervised classification model is the approximation of a function  $\hat{f}(\cdot)$ , for which  $\hat{f}(X) = f(X) = Y$ . To evaluate the quality of such an approximation, a loss function,  $L(\cdot)$ , is required, indicating how far  $\hat{f}(\cdot)$  is away from the true mapping function  $f(\cdot)$ . A simple loss function can

be represented by a  $K \times K$  matrix  $\mathbf{L}$ , where  $K = \text{card}(\mathbb{G})$  is the cardinality of  $\mathbb{G}$  and  $L(f(X), \hat{f}(X)) = 1_{f(X) \neq \hat{f}(X)}$ . Therefore, the loss function takes a value of 0, if a prediction is made correctly, 1 elsewhere. For  $n$  samples, any algorithm minimizes

$$E = L(f(X), \hat{f}(X)) = \arg \min_X \frac{1}{n} \sum_{n=1}^n L(f(x_i), y_i), \quad (3.2)$$

where,  $E$  is the mean error made by the model with respect to the training data [47]. It is to say that the approximated function,  $\hat{f}(\cdot)$ , can be over complex (e.g. when generated through a neural network) or fairly intuitive (as for linear models). As some methods outperform others, they still might be impractical due to their complex output and intractability.

### 3.6.3 Bias–Variance Trade-off

The bias-variance tradeoff is a dilemma that occurs for every supervised learning problem. It describes the problem of the simultaneous minimization of two error terms, the bias and the variance:

- The bias describes the problem of an algorithm, not being capable of modelling the true relation between training data  $X$  and label  $Y$ . This error is based on false assumptions made by the algorithm and known as underfitting.
- The error of variance occurs if an algorithm reacts over-sensitive to the training data. This results in an overfitting of the model, as the algorithm interprets noise within the data as signal.

This dilemma takes a central role for classification tasks, as it holds true for all supervised regression and classification models [74]. Ideally the model is capable of detecting all relations between the input data and the corresponding label, simultaneously keeping its property of generalization to unseen test data. For example, a linear model may not fit the data in perfect detail, missing some observations, but provides a constant performance when evaluated on test sets. Therefore it has low variance, but a high bias. This behaviour corresponds to the bottom-left bullseye in figure 3.6. On the other hand, if a spline is added to the regression model, it may perform very well on the training data, but suffers from high variance in the test scenario. Applied onto multiple test sets it tends to perform poorly or highly accurate for one or the other set, then on the training set. Vice versa, this model has high variance and low bias, as depicted in the top-right corner of figure 3.6. Given

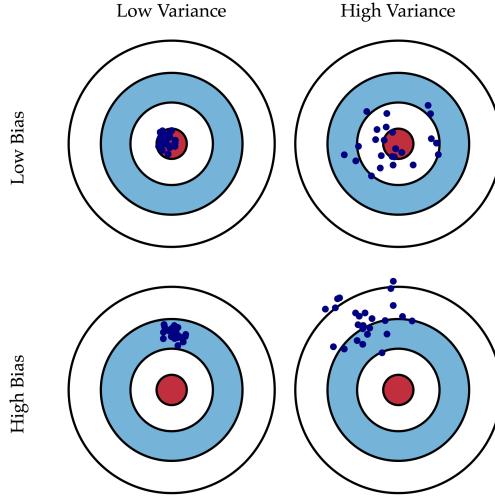


Figure 3.6: Simplified illustration of the bias-variance trade-off. Figure slightly adapted from Scott Fortmann-Roe (<http://scott.fortmann-roe.com/docs/BiasVariance.html>).

a set of training samples  $X_1, \dots, X_n$  and corresponding labels  $Y_1, \dots, Y_n$ , we aim to approximate the true relation  $f(\cdot)$ . Unlike the assumption made in 3.6.2, data  $X$  has noise to it and an error term,  $\epsilon$ , is required. Therefore an approximation of  $Y_i = f(X_i) + \epsilon$  has to be performed. Also, as the irreducible error,  $\epsilon$ , is unknown, the models fit has to be measured for the training and test set independently, seizing the models true performance and error tolerance [47].

### 3.6.4 Model Interpretability

As machine learning effects a broad spectrum of critical areas, such as medicine, criminal justice or the financial markets, there is strong a will to understand and interpret these models. As the approximated function of any model is just an imperfect assumption about a real-life process, there exists a certain interest in understanding how an approximated function came to its decisions.

A well organised review of motifs and required properties is given by Lipton [81]. In his study on model interpretability he gives four motivators why it is important to understand a models output.

**Trust:** Simplified, a model can be taken as trustworthy if its shown to perform well on the task it is trained for. But trust might also be more subjectively. Therefore, a user might feel more comfortable with a model he can totally understand. This might not be the case for speech recognition on a dial service, but it's becoming an issue, if the user invest money to a certain stock, suggested by a model.

**Causality:** The main focus of modelling is to make predictions. But in some scenarios, as in this thesis, models might be used to infer properties of the underlying problem. For example, a simple regression model could reveal the association between tobacco abuse and lung cancer. As for correlation, not any association might imply causality, but they might point their user into the right direction.

**Transferability:** Mostly, models are trained and evaluated on a data set which is split into two chunks. The first chunk is used for training and the second is used for evaluation (as discussed in 3.6.3). But how will this model perform when brought into a productive environment. In real-life, a model could be trained on gene expression data to classify patients regarding their cancer status. Likely this model will become invalid and produce false predictions if the underlying technology changes, such as the expression chip. If a model is interpretable, transferring it becomes an easier task.

**Informativeness:** Sometimes the model itself doesn't perform any automated task, it just suggests some likely options to user, as done in decision support systems. As the model reduces an formal error, the user might be interested in the real-world purpose of the suggested action.

Based in these motifs, properties regarding the model and its approximated function can be defined, helping to evaluated a learner with respect to its interpretability.

**Transparency:** A model is considered transparent, if its simulatable. Here simulatability can be understood in the way, that the user is able to reproduce the decision made, just with the input data and model parameters at hand. To reproduce a fitted regression spline is intractable to a human without any access to a computing machine, but also is the reproduction of a deep classification tree (classification trees are discussed in 3.6.5) with thousands of leaves.

**Decomposability:** It should be guaranteed that each part of the model - the input, the parameters, the calculation etc. - are assessable by the user. This might be the case for decision trees, if the input variables are defined clearly with respect to the user. But considering feature aggregation procedures, like a principal component analysis (PCA), which accumulates features to achieve a better prediction, predictors can become black boxes to the user.

**Algorithmic transparency:** Undeniable users are able to understand the procedures of recursive splitting and partitioning, made by trees. Whereas it takes more to understand and reproduce the complex training process of a neural network. Therefore, a model is simpler to understand of its underlying algorithms are intuitive.

**Post-Hoc interpretability:** Even if the model trained well and only provides a non complex mapping function which is understandable to the user, it is of interest to provide factors which simplify the decision being made. Therefore, predictor variables could be extended with further information, providing context for the user. Also it is an important factor, that the model can be visualized well. When struggling with complex scenarios, a well visualized model might allow to focus on the problem and not the model itself. Lastly, examples should be given. Examples should be chosen in a way that they are intuitive, so that the user can focus on the understanding of the model, before investigating the decisions made.

These concepts are both, important and slippery at the same time. It seems clear that a single model cannot achieve all of the above goals. Therefore it is always a trade-off between interpretability and accuracy. Complex problems might be solved by simple and intuitive models, but with the cost of feature engineering, violating one of the above criteria.

### 3.6.5 Established Models

To asses the value of the EDL, it is necessary to compare its performance to other models. Here, four well known models are introduced, namely Support Vector Machine (SVM) [25], random forest [12], multinomial regression [9] and classification trees [13]. The models have been chosen with respect to the analyzed data types, their complexity and interpretability. As a first criterion, all models have to be capable of handling continuous, categorical and binary predictor variables at the same time. This is required as a broad

variety of data types is tested. Second, the SVM and random forest models have been chosen, as they are known to perform well on complex data sets, but remain difficult to interpret. Vice versa, the regression and tree models are intuitive to interpret, but should not perform as good as the other models. Further, all models react different to irreducible errors, introduced in 3.6.3, and therefore show different behaviours regarding the bias-variance trade-off.

### Multinomial Regression

First, the multinomial logistic regression, short multinomial regression, is introduced. It generalizes the logistic regression model to handle multiclass problems, therefore more than two discrete outcomes are possible. It assumes that the label can be modeled as a weighted linear combination of the predictor variables, but is not perfectly predictable from a single variable. As with other regression models, statistical independence and collinearity of the predictor variables can be neglected [47]. Basically the model can be written as

$$\text{score}(X_i, y) = \beta_y * X_i, \quad (3.3)$$

where  $X_i$  is the vector of observations of a single sample and  $y$  its corresponding class.  $\beta_y$  is the vector of weights to be multiplied with  $X_i$ , to model the combinations. As the multinomial regression decomposes multi label classification into  $k - 1$  binary classification problems, where  $k = \text{card}(y)$ , the above problem has to be solved  $k - 1$  times. Therefore, one class has to be chosen as reference beforehand and the final classification decision is made by maximum class probability for each regression model. For each regression problem the identification of the coefficients from equation 3.3, is then solved by Maximum a posteriori estimation (MAP). Here the implementation by Venables and Ripley's R package `nnet`<sup>8</sup>, described in [130], is used.

### Classification Trees

Classification trees belong to oldest methods of classification and rely on the concept of recursive partitioning. Here the `rpart`<sup>9</sup> package for R has been used, which implements the classification tree described by Breiman et al [13]. In this implementation, the algorithm recursively splits the input data  $X$  by testing each predictor variable and each of its values as threshold, to

---

<sup>8</sup><https://cran.r-project.org/package=nnet>.

<sup>9</sup><https://cran.r-project.org/package=rpart>.

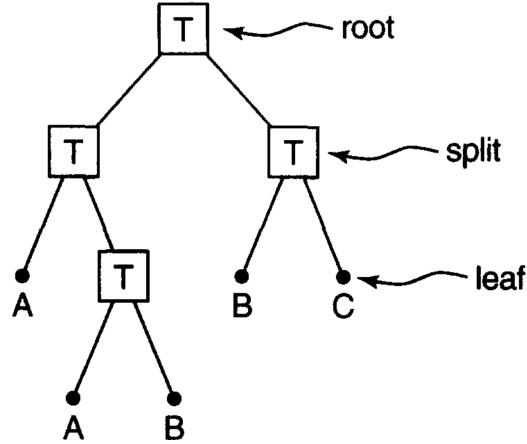


Figure 3.7: An outline of a decision tree. The splits within the tree (labeled with T) are called nodes, and the outcomes (labeled with A,B and C) are called leaves. The top node is called the root. Figure adapted from Breiman et all [13].

minimize the gini index. With

$$GI(S) = \sum_{i=1}^k f_i(1 - f_i), \quad (3.4)$$

where  $k$  is the cardinality of  $y$  and  $f_i$  that fraction of items labeled with  $i$ . This gini index can be understood as a measure for set impurity. For example, the gini index of a set  $a, a, a, a, b, b, b, c, c, c$  would be 0.66, as  $(0.4 * (1 - 0.4)) + (0.3 * (1 - 0.3)) + (0.3 * (1 - 0.3)) = 0.66$ . A more pure set would be  $a, a, a, a, b, b$ , with  $(0.67 * (1 - 0.67)) + (0.33 * (1 - 0.33)) = 0.44$ . The predictor and value combination yielding the smallest gini index is then used to split the data. For the remaining data, this growing procedure is repeated until all samples are represented by the tree. If it is not possible to perform a pure split, the subset with the smallest gini index is chosen. This impure subset is then searched again with the remaining predictors for a pure split. This way the characteristic tree structure (see figure 3.7) is created. For this algorithm the choice of the next split only relies on the current state. This is referred to as greedy algorithm and introduces weaknesses, as the algorithm performs a local optimization to find the current best split. This local search does not aim to optimize any global criterion and tends to create overfitted models [46]. To avoid overfitting, tree pruning is introduced. The pruning procedure removes *unimportant* splits within in sequence, not exceeding a purity threshold given by the user.

### Random Forest

A random forest is an ensemble method for classification and relies on multiple bagged trees. To overcome the weaknesses associated with classification trees, Breiman introduced the random forest model back in 2001 [12]. To create a random forest, a fixed number of  $m$  bagged trees are grown and the final classification decision is carried out by voting. Here, a bagged tree is a tree grown by the mean of Bootstrap Aggregation (Bagging) (for details on bagging the reader is referred to Breiman et al [11]). With bagging,  $n'$  samples are drawn with replacement from the original  $n$  samples, forming a new training set  $D_i$ , the classification is then performed on the remaining samples, which have not been chosen for training. The procedure is repeated  $m$  times, while each tree has one vote for the classification of each sample. By default  $n' = n$ , yielding an expected ratio of  $(1 - 1/e) \approx 0.632$  of training to test cases. This procedure can be understood as meta-learning algorithm as it is applicable to any classification or regression model. While random forests are known to heavily increase the performance of trees, this model becomes almost uninterpretable as often several hundreds or thousands of trees are grown. In this work, the random forest implementation by Wright et al is used [137].

### Support Vector Machines

Other than the methods introduced before, SVMs can only be used for binary classification and they require the input data to be linear separable. That is, there exists a vector, called hyperplane,

$$0 = w * X + b, \quad (3.5)$$

where  $w$  is a normal vector of  $X$  and  $b$  a simple scalar. For a better understanding figure 3.8 left shows linear separable data with a hyperplane and its margins. All data points are labeled regarding the side of the hyperplane they are located on. If their distance from a data point to the hyperplane is high, there probability of belonging to the class is high and vice versa. Thus,  $w$  and  $b$  have to be chosen, so that equation 3.5 is fulfilled. If the input is not linear separable, as shown in figure 3.8 right, exceptions can be made allowing a fraction of samples to be miss-labeled while training. This is called soft margin, while the first version is called hard margin. If the data is not linear separable and soft margins are applied, then, for any hyperplane 3.5, there exist  $x_i \in X$ , such that

$$y_i[w * x_i + b] \not\geq 1, \quad (3.6)$$

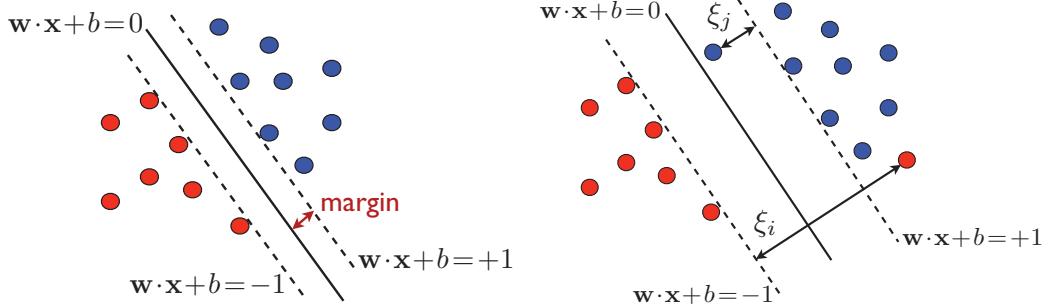


Figure 3.8: Left: Linear separable data in a two dimensional space with one possible maximum margin (hard margin). Right: Not linear separable data where an exception is made (soft margin). Figure adapted by Mohri et al [88].

which is not in agreement with 3.5. But with the introduction of the error term (also known as slack term )  $\xi_i$ , we can reformulate 3.5, such that

$$y_i[w \cdot x_i + b] \geq 1 - \xi_i. \quad (3.7)$$

Here  $\xi_i$  is a measure of distance, describing the gap between  $x_i$  and the hyperplane. This is illustrated in figure 3.8, right.  $x_i$  is classified incorrectly and  $x_j$  violates the hyperplane property. Hence, the error  $\xi_i$  is incorporated proportionally to the distance for both samples.

As most data sets are, even with soft margins, not linear separable, the SVM uses a transformation. Using a kernel function the data is transformed into a higher dimensional space and tested for linear separability again. This process is repeated until a hyperplane is found. The implementation used for all classifications in this work relies on Changs `libsvm` [17] and realizes multi class classification by solving  $k$  times *one-versus-all* classifications. Here the distances to hyperplane are used as score and the model with the maximum distance for a samples wins.

## 3.7 Classifier Design

The goal is to predict models which are accurate, yet interpretable according the criteria introduced in 3.6.4. Hence, the resulting model should be readable and semantically intuitive to any user with knowledge of the problems domain. Also some measure of evidence should be provided, indicating the *importance* or *relevance* of a decision being made. One model fulfilling the criteria given, is the decision list introduced by Rivest [106]. A decision list

consists out of consecutive *IF... THEN...* rules and can be understood as single path of a classification tree. Not only that this model is intuitive, it also compares well to medical scoring systems such as the CHADS<sub>2</sub> score for stroke prediction [48] or *ERB2* scoring systems, which are utilized for breast and gastric cancer scoring models [117, 60]. While Rivest only provided the theoretical framework for decision lists, they were implemented first by Quinlan and Quinlan [103, 104] afterwards, as a simplification of decision trees. In the following, problems with established decision list methods are discussed and a generative model to overcome weak spots is given. The introduced algorithm solves the classification problem through global optimization of a single decision list.

### 3.7.1 Interpretability versus Accuracy

The two most well known methods for the generation of a decision list are the C4.5Rules algorithm by Quinlan [103] and the PART procedure, described by Witten et al [45]. Both algorithms build decision lists by the means of a classification tree model described in 3.6.5. Basically, both methods collapse each path through the tree into a single rule, which is then pruned. The resulting set of rules is assembled into a decision list using optimization procedures. This has the benefit that the yet simple tree model becomes even more intuitive. But, as these procedures rely on decision trees only, resulting models are pretty unlike to exceed the classification accuracy of their predecessor trees [74]. As there might be a small increase in accuracy, due to rule pruning, the main drawback is, as discussed in 3.6.5, the greedy procedure to grow the trees. Common approaches to overcome problems introduced by greedy algorithms are boosting or bagging of the classification models, as done with random forests (3.6.5). Utilizing such methods results in complex and hard to understand models, as these techniques build their decisions on the basis of several hundreds of models. Another way to improve classification accuracy is the global optimization of the classification problem. To optimize a decision tree, its initial structure has to be given upfront. Therefore, the number of splits, the number of leaves, the leave labeling and tree depth become additional model parameters, making it uncomfortable to work with. Further, the search space becomes needlessly large (Bennett [5] provides a summary on tree optimization). Most of those model parameters can be avoided, if the problem is formulated as a decision list, as splits are decided automatically by the lists structure and only the list length has to be given beforehand.

### 3.7.2 The Evolutionary Decision List

A decision list consists of consecutive *IF...THEN...* rules. If one rule does not apply, the next one is tested. If none of the rules apply, a default rule is chosen. A single rule can be described by the context-free grammar  $G = \{N, T, F, S\}$  (slightly adapted from Espejo et al [40]), with

$$\begin{aligned} F &= \{ \text{rule} := \text{IF } \text{antecedent} \text{ THEN consequent} \mid \\ &\quad \text{ELSE label;} \\ &\quad \text{antecedent} := \text{test} \mid \text{antecedent} \& \text{ test;} \\ &\quad \text{test} := \text{name operator value;} \\ &\quad \text{operator} := ==; \\ &\quad \text{consequent} := \text{label} \} \\ T &= \{ \text{IF}, \text{ THEN}, \&, \text{name}, \text{ value}, ==, \text{label} \} \\ N &= \{ \text{rule}, \text{ antecedent}, \text{ test}, \text{ operator}, \text{ consequent} \} \\ S &= \text{rule} \end{aligned}$$

where  $F$  is the set of state transition functions,  $T$  the set of terminal symbols,  $N$  the set of non-terminal symbols and  $S$  the set of start symbols. It is easy to see that this grammar can only create languages, capable of testing positive predictor occurrences. Hence, boolean false expressions, numerical and categorical variables are decomposed upfront (discussion provided in section 3.7.4). Using such a simple language reduces the search space drastically, as the majority of decision problems are decided during feature engineering. Vice versa, expanding the grammars operator set would create languages that are capable of testing more complex expressions. As given by the grammar, an antecedent can recursively be expanded to an arbitrary long series of conditions. This number is called the rules cardinality. The grammar also generates the default rule, which is necessary to terminate the decision list, if no rule is applicable. To generate rules by this grammar, any arbitrary algorithm for frequent item set mining can be utilized. Here the FP-Growth algorithm by Borgelt [10] is used. This allows also to take the rule support into account, requiring the rule to cover a certain amount of samples.

Due to the grammar which generates the rules, each rule can be interpreted as a nondeterministic pushdown automaton (PDA). This way, each rule acts as a small program on its own, solving a sub-problem of the classification task. To form a global classifier, a *population* (figure 3.9 B) is initialized from all available rules. The population initially consists out of  $p_s$  randomly assembled decision lists, where each decision list is of the predefined length  $l$ . Based on this population a simple genetic algorithm iteratively forms the global classifier. First, each lists performance is assessed using the

hamming loss (see figure 3.9 C, "Measure Performance"),

$$H(x_i, y_i) = \frac{1}{n} \sum_{i=1}^n \frac{xor(x_i, y_i)}{|Y|}, \quad (3.8)$$

where  $n$  is the number of samples,  $|Y|$  the labels cardinality,  $y_i$  the samples ground truth and  $x_i$  its predicted label. That is, the fraction of false predictions made by a decision list to the sum of all labels. The best decision lists are determined by tournament selection, where two lists are chosen randomly and the one with smaller hamming loss is tagged for breeding (see figure 3.9 C, "Selection"). This approach is known to put less selective pressure onto the population, yielding a slower convergence, but also providing a higher chance to find the optimum [70]. Tagged decision lists are then considered for breeding, figure 3.9 D, with equal chances for crossover and mutation. The crossover operator randomly samples two lists from the tagged once, defines a random splitting point and swaps their trailing rules, right behind the splitting point. For a mutation, a decision list is chosen randomly and a random rule is replaced random by a rule sampled from the set of all available ones. This process is repeated for a pre-defined number of generations,  $G$ . During all generations, the best list found is kept aside and returned after the genetic algorithm terminates. This list is the final decision list. The idea of genetic programming was introduced by Koza [70] back in 1992. During the past years this paradigm has helped to solve several optimization problems, due to its flexibility. Compared to other numerical optimization methods, such gradient decent, it is more likely to find a global optimum, but can therefore be more time consuming. Other recently introduced decision list construction techniques by Letham et al [78], Yang et al [139] and Wang et al [132] rely on the Bayesian statistics and provide an alternative way of construction. Initially introduced by Letham et al, a single decision list assembled at random and then modified according to the posterior distribution. This reduces the memory footprint and calculation time, but introduces additional burden on implementation. Also the model is only considered for binary class prediction, which is unusable in the most cases. Executing such a model with a meta algorithm, where the winner is determined by voting or probability (as implemented by the SVM 3.6.5), would indeed carry out a multi class prediction, but also obfuscate the interpretable decision list.

### 3.7.3 Precision & Confidence Intervals

To provide some further insights into the data and the fitted model, the precision and confidence intervals (CI) for each rule are given. These coefficients are intended to provide a measure of *correctness* for each rule. Hence,

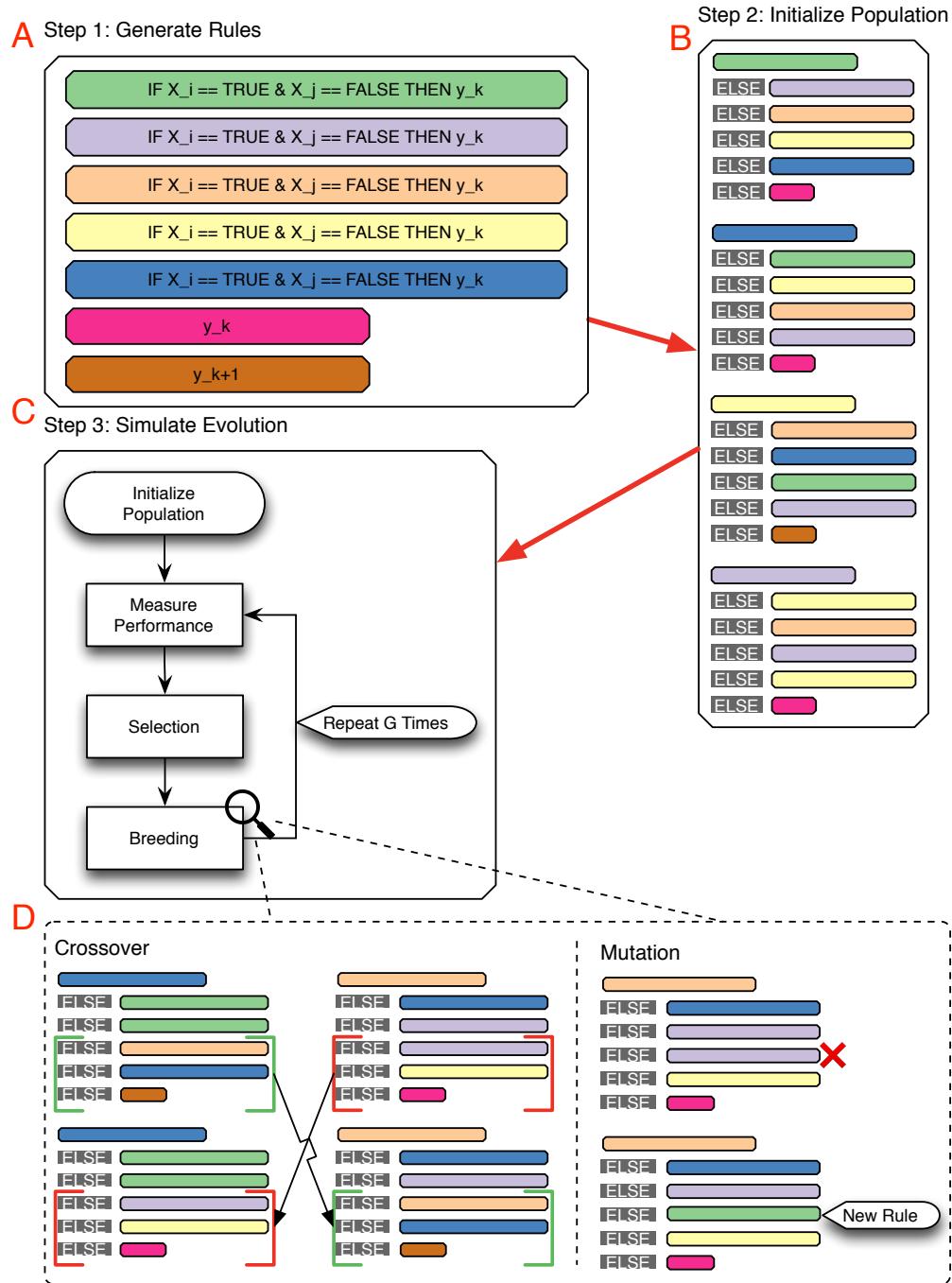


Figure 3.9: The workflow of the EDL. In the first step, A, all available rules are generated from a data matrix. Afterwards, in B, a set of decision lists, the population, is initialized at random. C depicts the workflow of the genetic algorithm, where the performance of each rule determined and the best rules are selected for breeding. This process is repeated  $G$  times. For breeding, D, rules are chosen at random from tournament and undergo a crossover or a mutation.

Table 3.1: Decomposition of categorical predictor variables

V_all	V_cat	V_dog	V_chicken
cat	T	F	F
cat	T	F	F
dog	F	T	F
chicken	F	F	T
chicken	F	F	T
chicken	F	F	T

the number of true positives (TP) classified samples per rule, divided by the number of samples classified by the rule in total,

$$P(R) = \frac{tp}{(tp + fp)}. \quad (3.9)$$

The series of decision rules can be treated as independent sequences of binomial experiments, each with a cohort of size  $n$  and a success rate of  $tp$ . With these parameters at hand, the confidence intervals for the observed outcome can be calculated by the Clopper-Pearson interval. As it is sufficient to know, that each experiment can be treated as binomial experiment, and therefore the Clopper-Pearson interval applies, the reader is referred to Clopper and Pearson [19] for more details on this particular method. In this work the confidence intervals haven been calculated using the `binom.test` function from the R core library.

### 3.7.4 Binary Data Representation

For the process of rule mining it is required that the data is in binary format. That is, only binary states of a samples predictor can be taken into account. As most of the test data sets (3.2) and aggregated cancer data sets (4.2) are not available in this binary format they need to be converted. Categorical predictor variables are decomposed by their cardinality  $k$ . Thus, each state a categorical predictor can take is represented as its own binary predictor. As shown in table 3.1, for one predictor with  $k = 3$  states, this will result in three separate predictors to represent all states.

For continuous predictors, the z-score (equation 3.1) is utilized first (if not already applied), to normalize the data. This score indicates the number of standard deviations above/below the mean. Therefore, the cuts can be chosen naturally as a grain of detail. For  $n$  cuts chosen, this results in  $n + 1$  predictor variables.

Table 3.2: A blank confusion matrix

	F	T
F	TN	FP
T	FN	TP

## 3.8 Performance Assessment

This work assesses and compares the performance of multiple models with respect to their ability of predicting unseen samples. To put different models in contrast, a unit of measure is required. Further, resampling techniques are required to ensure that a model does not only perform well on its training data, but also when applied to a new set of samples. This section introduces methods for performances assessment and resampling methods for gaining statistical confidence from observed measurements. A confusion matrix serves as basis for further calculations. This matrix provides an overview of how the samples have been labeled, compared to their true class. A blank confusion matrix is given in table 3.2. In this confusion matrix, TP equals to the number of samples being of *true* and been labeled as such, true negatives (TN) the number of samples being *negative* and labeled as such, false positives (FP) the number *negative* samples being predicted as positive and false negatives (FN) the number positive samples, which have been labeled as negative. This concept can easily be extended to multi label classification, where the rows of the matrix are labeled with true class labels and the columns with the predicted ones.

### 3.8.1 Classifier Performance

Here, the performance of a classifier or model does not describe the time it takes to execute on a data set. Rather it describes its capabilities on how well it can be suited to a given set of samples. First, it should mentioned that not all measures are suitable for all kinds of problems. For the most cases analyzed in this work, a multi label classification problem is given. Therefore, only measures to assess their performance are taken into account, leaving out binary measures of performance. One of the most well known and frequently used measures might be the classification accuracy

$$ACC = \frac{(TP/TN)}{(P + N)} = \frac{(TP/TN)}{(TP + TN + FN + FP)}. \quad (3.10)$$

According to the confusion matrix from table 3.2, the accuracy is the fraction of correctly classified samples to all samples. One drawback of this measure

Table 3.3: An example for cat and dogs classification

	dog	cat
dog	25	75
cat	0	900

is its weak spot for vastly unbalanced data. For example, considering 1,000 samples, from which 100 are of class *dog* and the remaining samples of class *cat*. A classifier which predicts the most samples as *cat* (as shown in table 3.3) performs well, according to the measure of accuracy ( $ACC = 925/1000 = 0.925$ ). Actually this classifier seems to work pretty well. Considering that the detection of dogs is an important task, there is a high chance that a model will seem to perform well (given its accuracy), but will actually miss most of the dogs.

This error can be corrected by taking the expected accuracy into account. The expected accuracy is the accuracy any random classifier is expected to achieve on the given data. It is defined by

$$EA = \frac{(TN + FP) * (TN + FN) + (FN + TP) * (FP + TP)}{(TP + TN + FN + FP)^2}. \quad (3.11)$$

This expected accuracy at hand, the inter rater agreement can be calculated. That is, the achieved accuracy under control of a random classifier and is known as kappa statistic or Cohens kappa [20],

$$\kappa = (ACC - EA)/(1 - EA). \quad (3.12)$$

Recap the cat and dog example, where most of the samples are classified as *cat*. The kappa statistic would yield a result of  $(0.925 - 0.88)/(1 - 0.88) = 0.375$ , which is far more pessimistic than the plain accuracy.

As none of the considered problems is such far of balance as the example, it is not expected that kappa and accuracy statistic differ that much. Still, the kappa statistic is the measure considered at first sight. For convenience the accuracy is given as a seconds measure.

### 3.8.2 K-Fold Cross Validation

The methods described in section 3.8.1 serve as a utility to measure how well a model is adjusted to a data set, but it does not reflect how the model will perform for predicting new cases. It might be possible that the model achieves high accuracy on its training data, but performs poorly on unseen cases. This phenomenon is known overfitting and takes place, if the models

starts to follow the *irreducible error* (3.6.3) or *noise* within the set of training samples. To ensure that the model has not been overfitted it is necessary to validate its performance on an independent test cohort, which has not been seen by the model during training. As the access to unseen samples might be limited, it is common to split the cohort into two halves, a training and a test set. This way, a model can be trained and evaluated independently.

This technique comes with two major drawbacks. First, depending on which samples are chosen for training, the difference between training and test performance can be huge. Second, only fifty percent of the data are used for training. This might effect the model, as statistical models tend to perform better, the bigger the sample size. For a detailed discussion, the reader is referred to James et al [62]. A simple, yet effective method to overcome these flaws is called K-fold cross-validation. As depicted in figure, 3.10, K-fold cross-validation splits the available data into  $k$  chunks (here 5). Afterwards,  $k - 1$  chunks are used for training and the remaining chunk for validation. This process is repeated  $k$  times, each time leaving out the  $k$ th chunk for testing. Therefore, only a small fraction of samples is left out for training. To choose the right  $k$ , the bias-variance trade-off from 3.6.3 has to be taken into account. One might argue that choosing  $k = N$  might be a good idea, where  $N$  is the number of samples. Despite from computational burden, there are  $n$  models to be trained, the cross validation will show high degree of variance. This is because all  $n$  training sets are very similar to each other, while the single left out sample can be very distinct to the training data. On the other hand, if  $k$  is chosen to low, a high bias will be observed, as the model has not enough data to correct for noise. To summarize,  $k$  has to be chosen with care. Empirically the choice of  $k = 5$  or  $k = 10$  have proven to show test error rates that suffer neither from excessively high bias nor from very high variance [62].

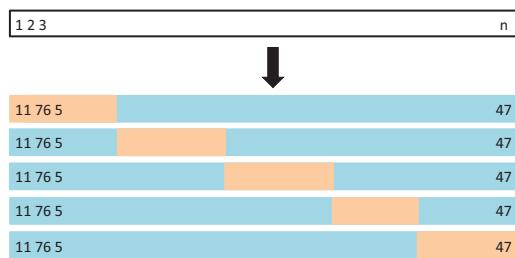


Figure 3.10: An example of K-fold cross-validation with  $k = 5$ . The data  $1, 2, 3, \dots, n$  is split into five chunks. Afterwards five sets are created (each shown in blue), each leaving out the  $k$ th chunk (shown in beige). This way, 11765 samples can be used for training, while 47 are used for validation. Figure adapted from James et al [62].



# Chapter 4

## Results

Fist, the results on the test data sets will be discussed. Afterwards, the developed EDL model is applied to a cohort of 1.000 breast cancer patients. This is done as a proof of principle, to ensure that the model is capable of both, labeling the samples according to their subtype and to re-identify known subtype driver alterations. The achieved model accuracy is compared to the accuracy achieved by the well established models. Following, a newly assembled cohort of 500 patients, suffering from primary or metastatic prostate cancer, is inspected using the EDL. Again, the accuracy is inspected in the context of the well established models. For both investigations the cohorts have been assembled using the FirebrowseR software and normalized by the methods established through Web-TCGA, as described in 4.2. Finally, the determined models are viewed under the aspects of their interpretability 3.6.4.

### 4.1 Test Data

The data sets introduced in 3.2 have been used by the machine learning community for several years up to decades. They act as benchmark to compare newly developed models to established ones. Since they are used by a broad spectrum of people not only in the machine learning community, common obstacles are known and improvements on classification accuracy can be determined easily.

To have a fair comparison between the different models, all models, except for the multinomial regression, have been tuned. For the SVM, the slack parameter  $C$  and the  $\gamma$  radial-kernel parameter have been determined using grid search with  $C \in \{2^2, 2^0, \dots, 2^6\}$  and  $\gamma \in \{2^6, 2^4, \dots, 2^2\}$ . For the classification tree, the minimum split parameter has been tuned with

$m_{split} \in \{1, 3, \dots, 19\}$ , indicating the minimum number of samples in a node to be considered for further splitting. Additionally, pruning has been utilized to simplify and potentially improve the tree. The  $m_{try}$  parameter for the random forest model, indicating the number of sampled predictors for each split, has been chosen from  $\log_2(n), \log_{10}(n), \sqrt{n}, n/2$ . As the multinomial regression does not have a tuning parameter, it has been guaranteed that the model converged. The list length parameter  $l$  for the EDL has been tuned by hand, with respect to an upfront chosen prior.

With respect to the bias-variance trade-off introduced in 3.6.3, 10-fold cross-validation has been utilized. During each run, each model has been tuned on the  $k - 1$  training samples and the final model performance has been carried out by the  $k$ th test set. Kappa and accuracy are given as mean with the corresponding SD, to detect potential outliers.

As these data sets are only used to compare EDLs performance to other models, all decision lists are postponed to the appendix A.1.

#### 4.1.1 The Tic Tac Toe data

The Tic Tac toe data consists of 9 categorical predictor variables, encoding all possible states (958) of the tic tac toe game. The aim of the classification model is to predict if player  $x$  has won.

As 3 identical symbols have to be present in either a row, a column or in the diagonal of the game matrix, only rules of cardinality 3 need to be generated. Longer rules not make any sense, as they wrap around the gaming board, while shorter rules are not capable of detecting a victory. Only rules with a rule support of at least 5% have been generated, yielding 152 classification rules in total. To create an EDL from those rules, the genetic algorithm has been run for 50.000 generations and with a population size of 152 decision lists. The length of each decision list was chosen to be 8, as there are exactly 8 possible board configurations to win the game. For the EDL model, the categorical variables have been decomposed as described in 3.7.4, resulting 27 binary predictors.

As shown in figure 4.1, it is clear to see that no other model, except to the EDL, achieved a perfect classification for all ten runs. The results can be viewed in more detail in table 4.1 and a final decision list, over all samples, can be found in the appendix section A.1.1.

#### 4.1.2 The Titanic data

The 3 categorical predictor variables form the titanic data set have been decomposed 8 binary predictors. Out of these, 58 rules of cardinality  $1, \dots, 4$

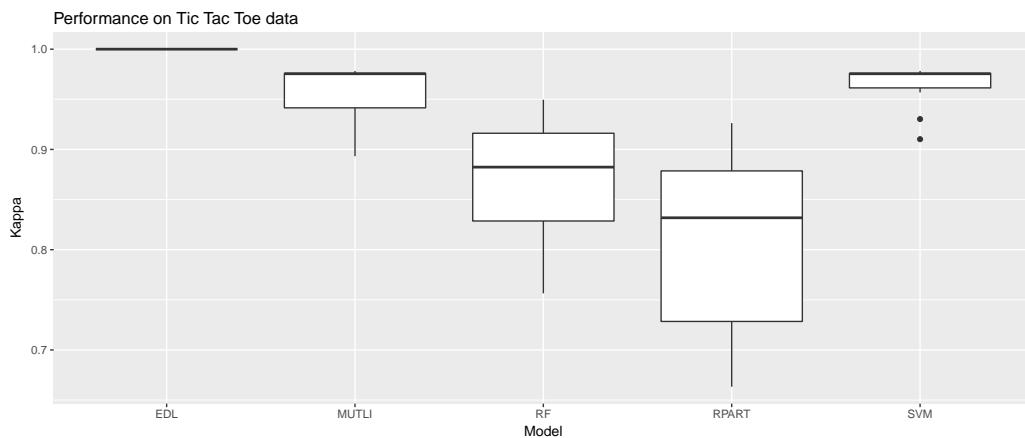


Figure 4.1: Boxplots indicate the classification performance (kappa statistic) of all five models for the Tic Tac Toe data set. Scattering indicates the 10-fold cross-validation.

Table 4.1: Tic Tac Toe classification performance. Given measures are the kappa statistic, accuracy and their SDs

	Kappa	Kappa SD	Accuarcy	Accuracy SD
EDL	1.000	0.000	1.000	0.000
Multinom Reg	0.957	0.032	0.980	0.016
Random Forest	0.869	0.060	0.943	0.028
Class Tree	0.812	0.095	0.916	0.040
SVM	0.963	0.024	0.983	0.011

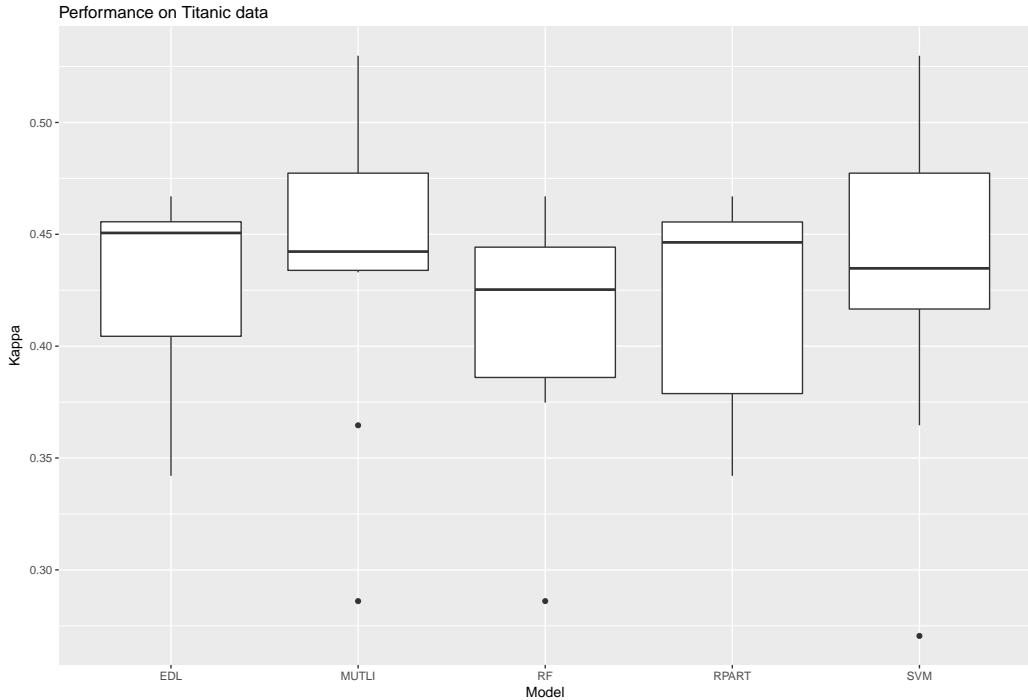


Figure 4.2: Boxplots indicate the classification performance (kappa statistic) of all five models for the Titanic data set. Scattering indicates the 10-fold cross-validation.

have been generated. The genetic algorithm has been run for 50.000 generations and the ideal decision list length was found to be 3. The results are shown in figure 4.2 and table 4.2. After cross-validation, a final decision list on all samples has been generated. This list is given in the appendix section A.1.2 with the corresponding precision and confidence intervals.

### 4.1.3 The Mushrooms data

The mushrooms data set consist of 20 categorical predictors, which have been decomposed to 111 binary predictors. Therewith 8116 rules of length  $1, \dots, 3$  and a minimum support 5% have been generated. The best performance could be achieved with a list length of 8, where 20.000 generations were run to build the model. The performance is given in figure 4.3 and table 4.3. Again, a final decision list has been generated using all available data set and is shown in A.1.3.

Table 4.2: Titanic classification performance. Given measures are the kappa statistic, accuracy and their SDs

	Kappa	Kappa SD	Accuarcy	Accuracy SD
EDL	0.429	0.044	0.791	0.022
Multinom Reg	0.439	0.070	0.778	0.026
Random Forest	0.411	0.054	0.781	0.029
Class Tree	0.421	0.046	0.788	0.025
SVM	0.432	0.073	0.776	0.027

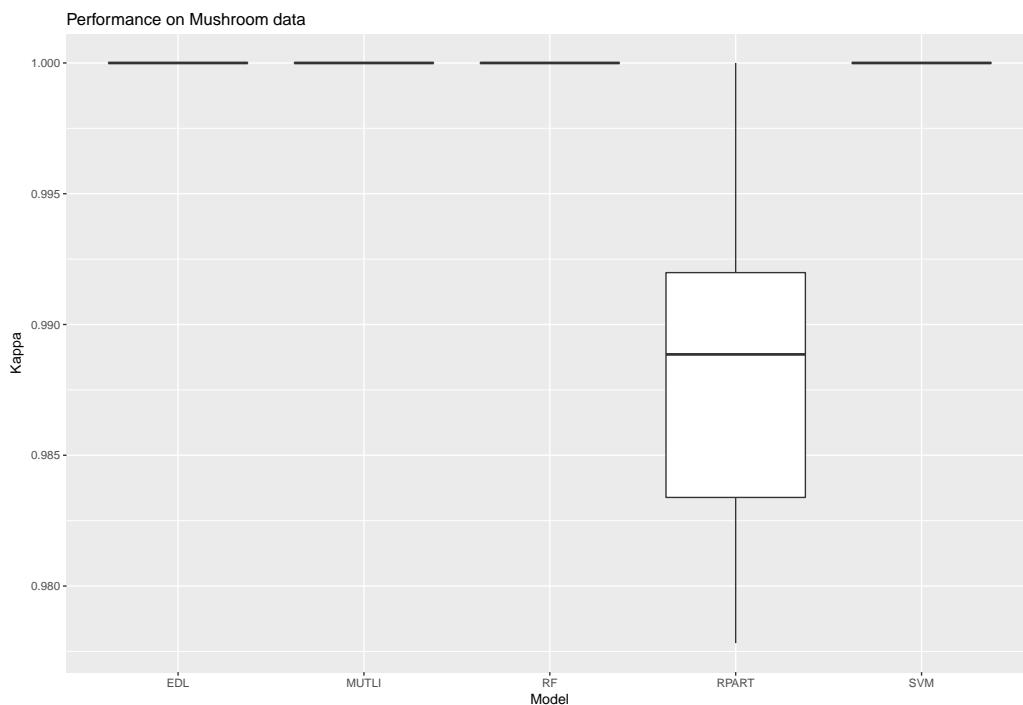


Figure 4.3: Boxplots indicate the classification performance (kappa statistic) of all five models for the Mushrooms data set. Scattering indicates the 10-fold cross-validation.

Table 4.3: Mushrooms classification performance. Given measures are the kappa statistic, accuracy and their SDs

	Kappa	Kappa SD	Accuarcy	Accuracy SD
EDL	1.000	0.000	1.000	0.000
Multinom Reg	1.000	0.000	1.000	0.000
Random Forest	1.000	0.000	1.000	0.000
Class Tree	0.988	0.006	0.994	0.003
SVM	1.000	0.000	1.000	0.000

Table 4.4: Cars Database classification performance. Given measures are the kappa statistic, accuracy and their SDs

	Kappa	Kappa SD	Accuarcy	Accuracy SD
EDL	0.879	0.041	0.936	0.029
Multinom Reg	0.853	0.037	0.932	0.018
Random Forest	0.932	0.043	0.968	0.022
Class Tree	0.877	0.020	0.942	0.013
SVM	0.853	0.030	0.932	0.015

#### 4.1.4 The Cars database

The cars data consists of 1728 samples, described by 6 categorical predictors. These have been decomposed to 21 binary variables. 416 rules of length  $1, \dots, 2$  and a support of 1% have been generated. The genetic algorithm ran for 20.000 generations, providing the best decision list with a length 25. The decision list trained on all samples is depicted in appendix section A.1.4. The classification performance to other models is depicted in figure 4.4 and summarized in table 4.4.

#### 4.1.5 Summary

As it can be seen, the EDL performed comparably well on all test data sets. For the Tic Tac Toe problem, no other method performed as well as the EDL. This is due the fact, that a lot of prior knowledge can be incorporated into the model. It is clear that there are only 8 states to win the game (for player  $x$ ), each consisting of exactly 3 predictor variables. This information can be taken into account by the models parameters, reducing the search space drastically and enabling EDL to find the global optimum. That important observation confirms that the EDL is capable of finding the global optimum if it exists. The reason the classification tree struggles at this problem, is be-

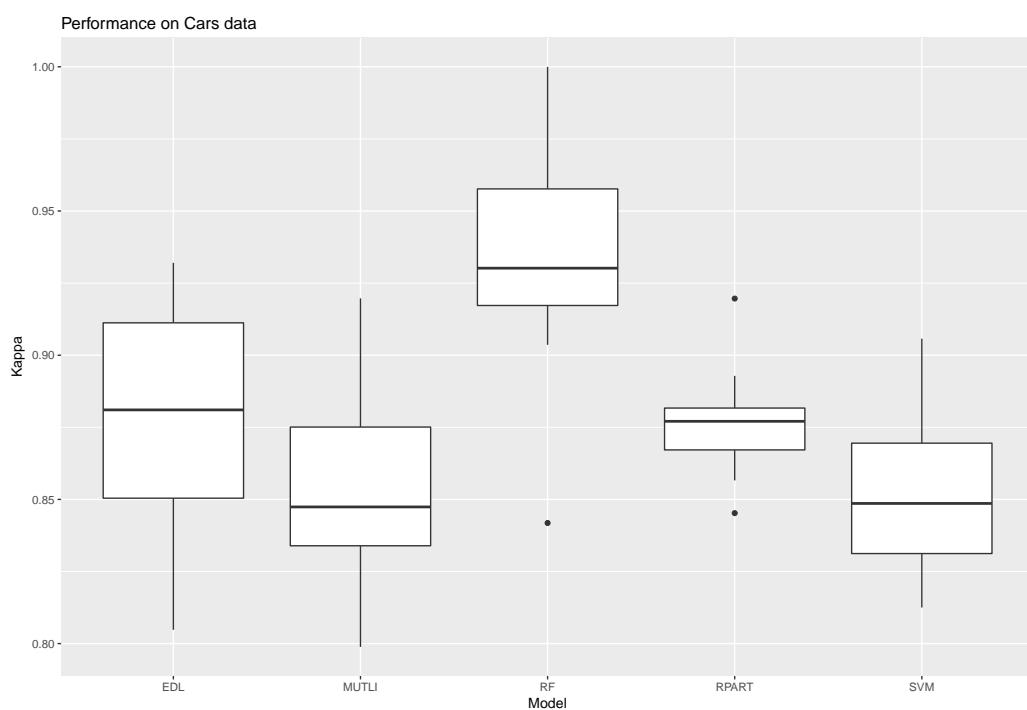


Figure 4.4: Boxplots indicate the classification performance (kappa statistic) of all five models for the Cars Database. Scattering indicates the 10-fold cross-validation.

cause of its greedy nature. The tree chooses the best split currently available. Hence, at the root of the tree, negative examples are chosen first, because more negative samples exist in the data set, which can be classified by a single split. The random forest model is unlike to find the global optimum, because of its voting system. Even if there exists several trees in the forest indicating the perfect solution, this solution can not be carried out, as it gets blurred by other trees during the process of voting. When inspecting the generated decision list A.1.1, it can be seen that the EDL model found the right combination of states, avoiding detours through combinations of non optimal states.

For the Titanic data set, the generated decision list A.1.2 recovered the policy by which the passengers have been rescued. Hence, children and female passengers have been saved first, while passengers traveling third class had a poor outcome of survival. As the discovered list gives an overview of the odds of survival and death, slight modifications to the rule set might reveal additional insights. Hence, the EDL could just be run with positive or negative (*survived/death*) rules, yielding single label probabilities only. Overall it should be mentioned that all of the methods performed well on that problem. Except for the multinomial regression, showing strong outliers regarding the kappa statistic.

The mushrooms data is another example, where the greedy strategy of the classification tree results in a decrease of performance. Even if its just a marginal effect, the tree is not capable to achieve a perfect classification. All other methods, which optimize for a global maximum, can overcome this issue. As a result, the EDL provide a decision list which can be used by anybody to detect poisonous mushrooms A.1.3.

For the Cars Database the EDL performance is comparable to the other models, but is clearly outperformed by the random forest. Further it shows a high variance for the cross-validation, indication that noise has been learned. As the overall performance seems acceptable, in special with a mean kappa of 0.879, some flaws come along with the decision list model. The overall decision list A.1.4 includes 25 rules with an additional default rule. This might be to long to still call this model interpretable. Here it might be worth to accept a less precise model, which in turn would yield a shorter and more transparent set of rules.

## 4.2 FirebrowseR + Web-TCGA = Data Foundation

The breast 4.3 and prostate cancer 4.4 subtyping analyses are based on the data sets obtained from the Firehose Pipeline. To obtain and normalize these datasets, the FirebrowseR package and methods adapted from Web-TCGA are adapted. The validity and applicability of the methods described in 3.5 are elaborated in this section. To demonstrate the capabilities of FirebrowseR, Web-TCGA and their combination, examples for each data type are provided. As a proof of principle, analysis of well known mutations, expression profiles and CNVs are provided, highlighting the tools strengths when working with genomic data obtained from TCGA or the Firehose Pipeline, respectively.

### 4.2.1 Mutational Data

For mutational data, a global profile is created. This profile aggregates the occurrences of somatic mutations of the genes *TP53*, Teashirt Zinc Finger Homeobox 3 (*TSHZ3*) and Von Hippel-Lindau Tumor Suppressor (*VHL*) within the cancer entities of breast invasive carcinoma (992 samples) and kidney renal clear cell carcinoma (437 samples). Within this cancer entities, *TP53* is known to be highly mutated in breast cancer, while *VHL* is known for its hight mutation rate in the kidney clear cell carcinoma (for details see Kandoth et al [65]). *TSHZ3* is added as a negative control and should not occur highly mutated in any of those entities. As shown in figure 4.5, *TP53* is highly mutated for breast cancer (32.5%) but is barely for the kidney entity (1.8%). Vice versa, *VHL* is highly mutated in the entity of kidney renal clear cell carcinoma (48.5%) but only on 1.4% of all breast cancer samples. *TSHZ3*, as negative control, shows mutation rates <1% in both entities.

### 4.2.2 Expression Data

For lung adenocarcinoma it is known that KRAS Proto-Oncogene, GTPase (*KRAS*), Epidermal Growth Factor Receptor (*EGFR*) and Transcription Termination Factor 1 (*TTF1*) show higher levels of over expression, than under expression [145, 95]. To re-identify those findings, the expression data sets related to these genes and the cancer entity are download by Web-TCGA over Firehose Pipeline. Then the z-score, equation 3.1, is used to calculate the SD from the populations mean. Using these methods Web-TCGA in combination with FirebrowseR was able to re-identify the described patterns

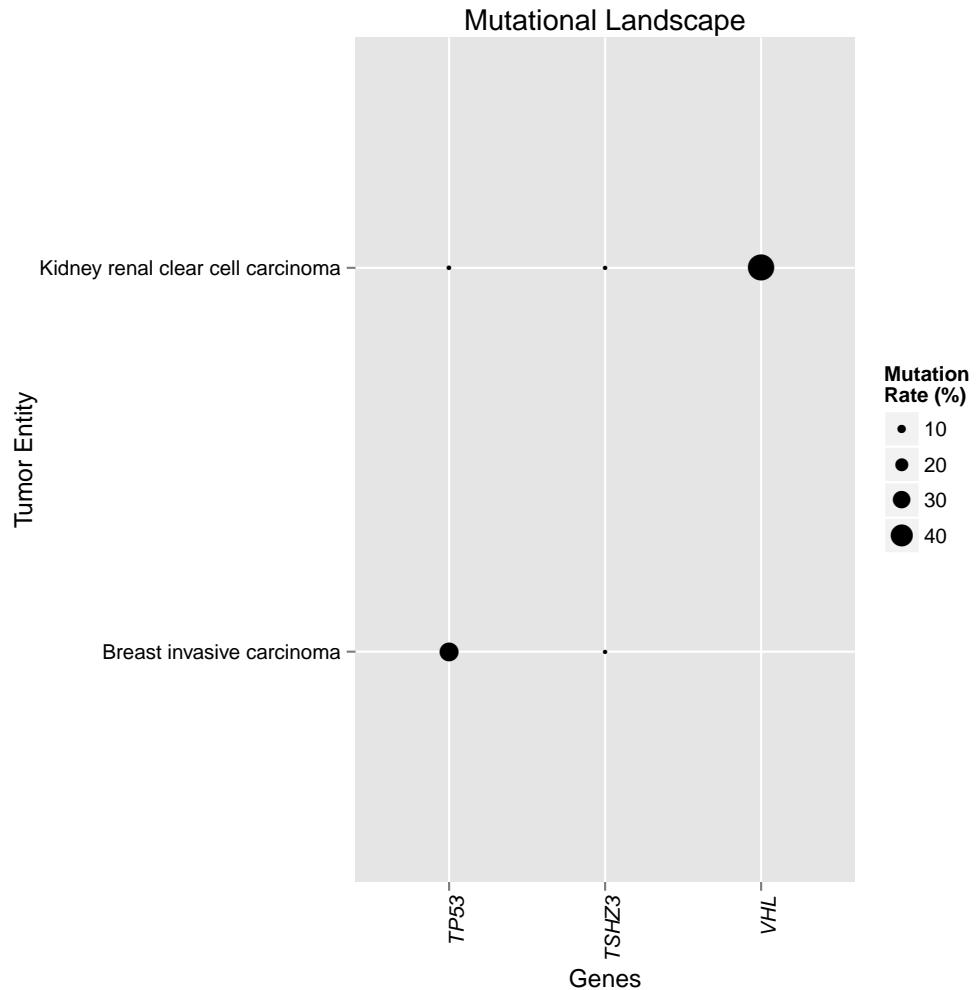


Figure 4.5: The global mutation profile for *TP53*, *VHL* and *TSHZ3* within the Breast and kidney cancer entities. While *TP53* is highly mutated in breast cancer, *VHL* shows almost no mutations. The reverse pattern is observed for kidney renal clear cell carcinoma, where *VHL* is highly mutated. In both entities, the negative control *TSHZ3*, occurs <1% mutated. Figure adapted from Deng et al [32].

as depicted in 4.6. It is clear to see, that more samples suffer from an over, than from an under expression.

### 4.2.3 Copy Number Variation Data

To handle CNV data, GISTIC2.0 is used by the Firehose Pipeline. Therefore, each gene is characterized by a copy number level, either -2, -1, 0, 1, 2, encoding a homozygous deletion, a heterozygous loss, no change (diploid), a gain and a high level amplification. To demonstrate the usability of this data type for cancer classification, the CNV status for Fibroblast Growth Factor Receptor 1 (*FGFR1*) and *PIK3CA* in lung squamous and lung adenocarcinomas is compared. Findings depicted in figure 4.7 reflect the results provided by Ciriello et al [18], where *PIK3CA* is highly amplified in more than 50% of all lung squamous cell cancer entities, but only in 4% of all lung adenocarcinomas. Furthermore, a high level amplification of *FGFR1* is detected in more than 27% of all lung squamous cell cancer samples, but only in a small subset of adenocarcinomas.

## 4.3 Breast Cancer Subtyping

As discussed in section 2.1, breast cancer is a heterogeneous disease, which can be classified into several subtypes. While these subtypes are characterized by certain pattern, a crisp classification still seems impossible. Outgoing from an initial clustering of expression profiles from breast cancer patients [96], other more advanced methods and analyses have been introduced. These incorporate additional data types [87, 92, 2] and were therefore able to provide a broader assessment of the disease, which let to a better understanding.

In this study, breast cancer and its known subtypes are utilized to evaluate the combination of EDL, the data provided over FirebrowseR and the normalization procedures described in the context of Web-TCGA. Therefore, using these techniques should yield identical results to state of the art models, recently published by Mo et al [87], Mer et al [85] and Curtis et al [27]. The combination of the introduced methods should be capable of detecting the molecular subtypes, as well as the identification well known driver genes.

### 4.3.1 Assembling a Cohort

Breast cancer data has been obtained from Firehose Pipeline using the FirebrowseR R client. Overall there where 1.097 samples available. There where 1.089 samples with copy number data, 977 samples with mutational data

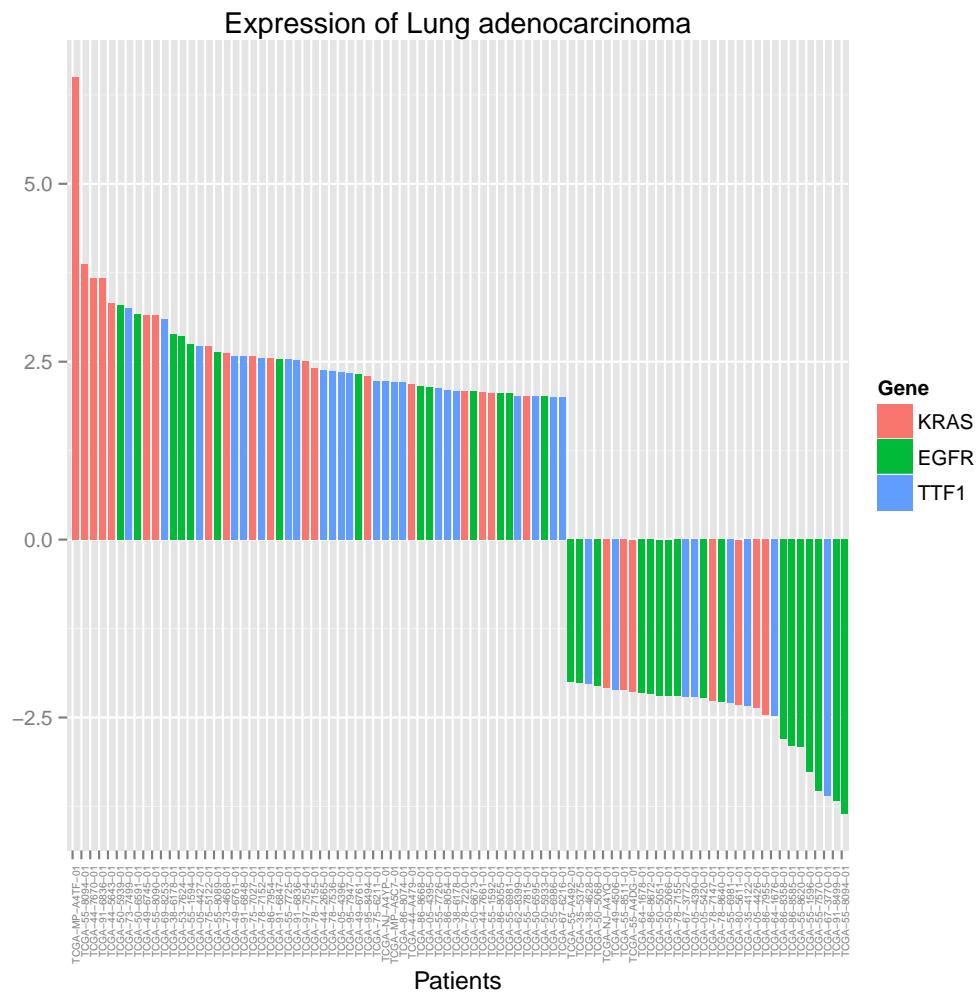


Figure 4.6: Expression profile for *KRAS*, *EGFR* and *TTF1* on the lung adenocarcinoma cohort obtained from FirebrowseR and analyzed Web-TCGA. As expected, the well known oncogenes show an increased level of over expression. Figure adapted from Deng et al [32].

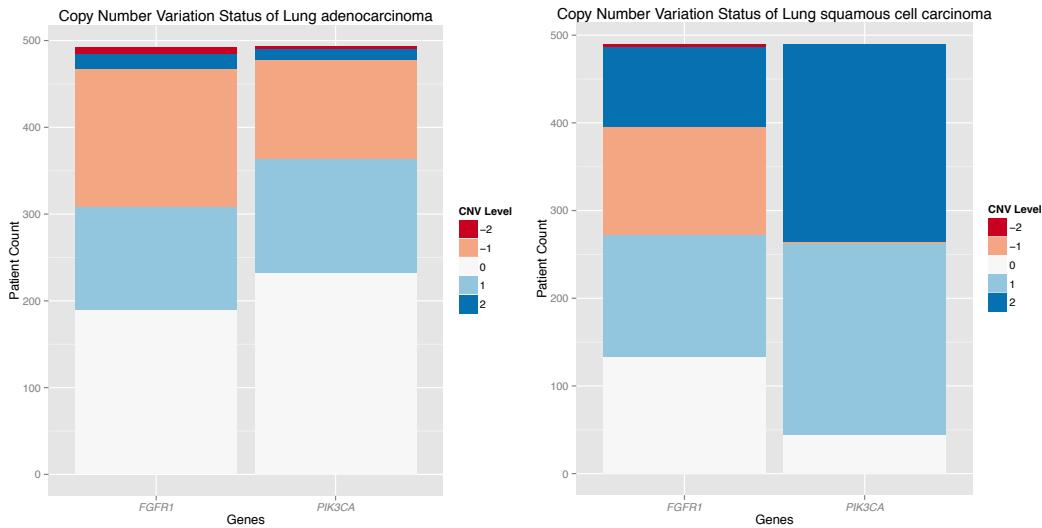


Figure 4.7: CNV profile for *FGFR1* and *PIK3CA* on lung adenocarcinoma and lung squamous cell carcinoma cohorts. Pattern described by Ciriello et al [18] could be re-identified using FirebrowseR and Web-TCGA. Figure adapted from Deng et al [32].

and 1.093 with samples with expression data available. As neither TCGA nor the Firehose Pipeline provide full PAM50 information, the annotations have been adapted from Keenan et al [68]. After intersection of all TCGA barcodes, 959 samples where identified, having mutational, copy number and expression data available.

All applied analysis provide the mutational, expression and CNV status for each gene. This results in several thousand predictor variables for each analyses type, making the problem untraceable due to the combinatorial explosion (for more details on the combinatorial explosion, the reader is referred to Venables et al [130]). Because of that, breaking down the number of predictors is a crucial task. For mutational data, only significant mutated genes obtained from MutSig analyses [76], implemented in the Firehose Pipeline, have been considered for analysis. For the CNV status, according to Kuhn [74], predictor variables with zero or near zero variance<sup>1</sup> and strongly correlated predictors<sup>2</sup> have been dropped. For expression data, only PAM50 genes haven been taken into account. This let to 88 predictor variables for the CNV status, 125 for the mutation status and 50 for the expression sta-

<sup>1</sup>Cutoff ratio for the most common value to the second most common value: 95/5.

<sup>2</sup>A cutoff of 90% pearson correlation has been applied.

tus. As described in 3.7.4, all variables have been converted to binary form. Each predictor for the expression data was converted to 5 binary predictors, indicating high underexpression, low underexpression, no differential expression, low overexpression and high overexpression. These classes were defined by calculating each genes z-score (equation 3.1), with the thresholds of two and one standard deviations away from the populations mean, for positive as for negative values. As GISTIC2.0 output is provided by the Firehose Pipeline (3.5.4), each predictor can consist out of five copy number states. These categorical predictors are decomposed to five binary predictors. After decomposition 352 CNV and 250 expression predictors were determined. All combined, together with the mutational data, this resulted in 677 binary predictor variables.

### 4.3.2 Classifying Breast Cancer Subtypes

Decision rules where generated with a cardinality of  $1, \dots, 3$  and a support of at least 5%, yielding 60,853 rules. The genetic algorithm was run for 100.000.000 generations and the ideal list length was found to be ten. As usual, the population size was chosen the same number as rules exist, 60,853. As proceeded with the test data, all other models were tuned as described in 4.1. The results are shown in figure 4.8 and table 4.5. It is clear to see that the EDL does not perform as good as the SVM and the random forest ( $\kappa$ : 0.704, 0.795 and 0.803), but still outperforms the multinomial regression and classification tree model ( $\kappa$ : 0.704, 0.501 and 0.659). Also the EDL provides its results with a variance, comparable to the SVM and random forest (0.042, 0.031 and 0.037). To ensure that the engineered features are not hindering the classification and sufficient information is provided, the results were compared to the finding of Mer et al [85]. Using a random forest model, they achieved an accuracy of 0.87 for 800 breast cancer samples, which compares well the random forest model trained here (accuracy: 0.852). It has to be mentioned that Mer et al did only include PAM50 genes as predictor variables. Therefore, an optimal classification can be provided, as the breast cancer subtypes are defined by these 50 genes. Further, noise is introduced into the cohort of this study, as additional data types are incorporated, from which not every single predictor provides signal, but can potentially improve the classification performance.

After ten fold cross-validation, a final decision list with the same model parameters has been generated on all available samples. This list and a corresponding classification graph is provided in figure 4.9.

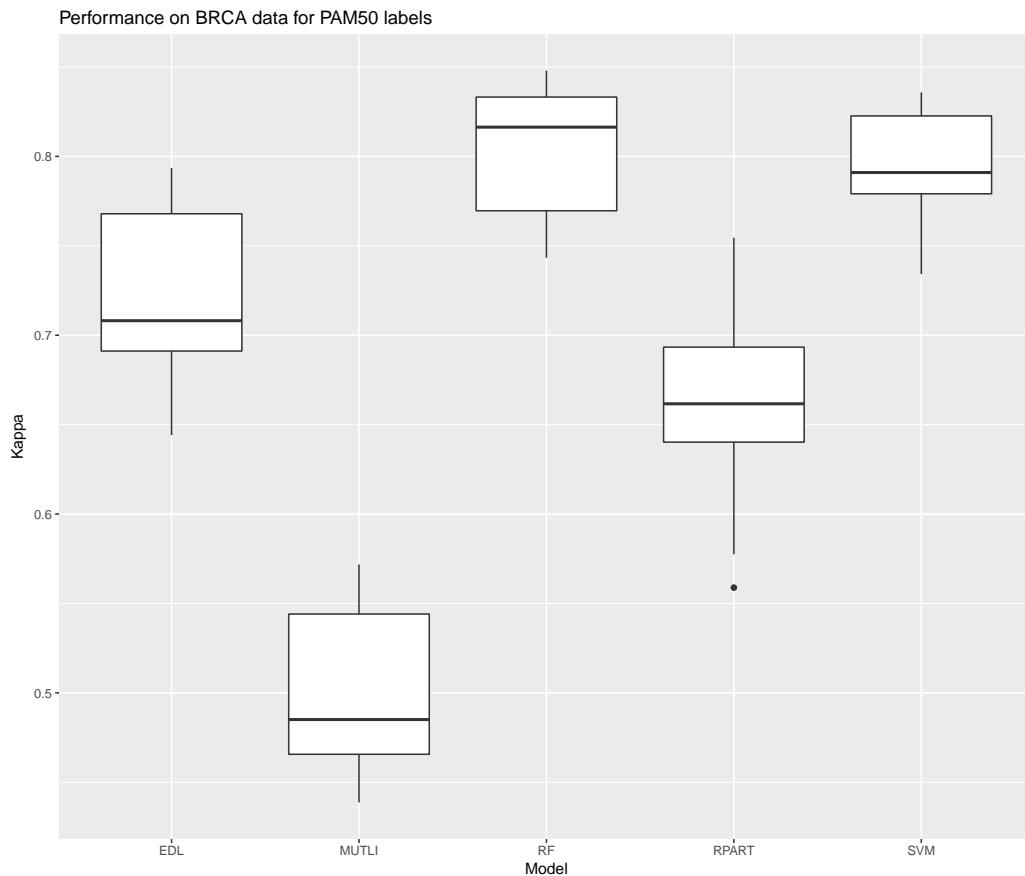


Figure 4.8: Boxplots indicate the classification performance (kappa statistic) of all five models for Breast cancer data with PAM50 label. Scattering indicates the 10-fold cross-validation.

Table 4.5: Breast cancer classification performance on PAM50 labels. Given measures are the kappa statistic, accuracy and their SDs

	Kappa	Kappa SD	Accuarcy	Accuracy SD
EDL	0.704	0.042	0.776	0.055
Multinom Reg	0.501	0.049	0.612	0.041
Random Forest	0.803	0.037	0.852	0.027
Class Tree	0.659	0.058	0.743	0.044
SVM	0.795	0.031	0.845	0.023

**Her2 Subtype** Surprisingly no *ERB2* variation has been incorporated into the decision list for the detection of the HER2 subtype, although this subtype is defined by an overexpression of this gene. Instead almost all samples of this subtype can be characterized by a combination of the expression and CNV status of six genes. For one thing, the Luminal A&B subtype specific gene *ESR1* is not known to be differentially expressed in the HER2 subtype, but is recorded with a low underexpression. A study by Denkert et al [33] lately revealed correlations between the response to trastuzumab therapy, indicating a good response for HER2 patients with an overexpression of *ESR1*, but no response was observed in *ESR1* underexpressed samples. This might be an indicator for the poorer outcome for HER2 classified patients. Another gene in the first rule is Forkhead Box C 1 (*FOXC1*), an inducer of epithelial-to-mesenchymal transition (EMT) and already linked to the HER2 subtype [124, 141]. Transmembrane Protein 45B (*TMEM45B*) and Forkhead Box A1 (*FOX1*) from the second rule have initially been found overexpressed by Parker et al [96] in one third of all HER2 samples. The gain of Phenylethanolamine N-Methyltransferase (*PNMT*) is likely to be an artefact of an *ERB2* gain, as its is located on the *ERB2* amplicon.

**Luminal B Subtype** In two out of three Luminal B rules *ESR1* occurs overexpressed. This behaviour is one essential marker for the Luminal B subtype in breast cancer. Despite from that, the important marker *CCNB1* has been detected. *CCNB1* regulates the cell proliferation and is commonly active in Luminal B tumors, making them more aggressive compared to Luminal A samples [98, 99].

**Basal Subtype** Two thirds of the Basal subtype is characterized by a single rule, for the other third of samples the default rule is applied, indicating that no sufficient information could be found. Within the single rule which characterizes most of the basal samples MYB Proto-Oncogene Like 2 (*MYBL2*) was identified with an overexpression. Just as *CCNB1*, *MYBL2* is an important player for cell proliferation [98, 99]. Along with *MYBL2*, Secreted Frizzled Related Protein 1 (*SFRP1*) it occurs within the basal rule. *SFRP1* is known to be key player in prostate cancer, effecting the Wnt signaling pathway [63, 34]. For the last gene, NDC80, Kinetochore Complex Component (*NDC80*) no association to cancer could be found.

**Luminal A Subtype** As well as for the Luminal B subtype *ESR1* and *FOXA1* are found in the Luminal A subtype with the same pattern of alteration. These samples are set apart by N-Acetyltransferase 1 (*NAT1*), *PgR*,

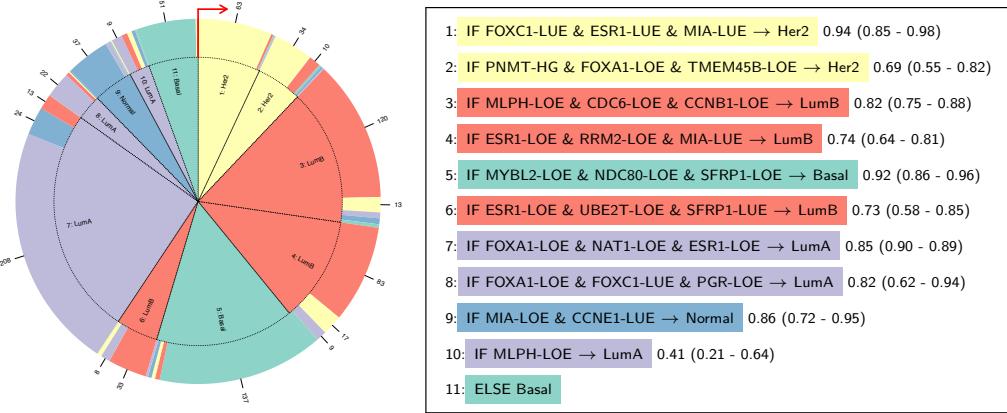


Figure 4.9: The final decision list generated for breast cancer data set. On the left hand side, the graph displays each rule applied (starting at 12 o'clock, clockwise) with prediction made (inner circle) and true labels on the outer circle. Numbers indicating the amount of TP samples (number < 5 are omitted, due to reasons of space). Coloring is done regarding the subtype and matched between both lists. The decision list on the right hand side shows the classification rules, precision and the corresponding 95% confidence intervals.

*FOXC1* and Melanophilin (*MLPH*). While *NAT1* and *MLPH* are not associated with either the Luminal A subtype nor any other, *PgR* is a key player for breast cancer disease. It is observed overexpressed Luminal A&B tumors and acts as a key marker for the choice of therapy [118, 56, 4]. Further, the EMT inducer *FOXC1* is found underexpressed in this subtype, which be another indicator for the better outcome in contrast to Luminal B and HER2.

**Normal Subtype** Within the single rule for the normal subtype only Cyclin E1 (*CCNE1*) is the only unseen gene in the decision list. This might be due to fact that this subtype is known to show no molecular signatures. Despite from that, *CCNE1* was identified to be distinct of this subtype. *CCNE1* is associated with a poor outcome for breast and ovarian cancer [91, 69], with a life expectancy of less than five years.

## 4.4 Prostate Cancer Subtyping

Prostate cancer is considered the male counterpart to breast cancer. Although its primary type might not be as aggressive, prostate cancer forms

different types of metastasis, decreasing the patients survival expectancy drastically. To classify prostate cancer according to its molecular profiles and to obtain predictions for survival rates from such profiles remains an ongoing task. While for breast cancer the initial PAM50 subtypes have been validated and proven as a reliable base, there is no such initial classification scheme for prostate cancer, that could be validated. In this section, problems and contradictions with existing prostate cancer classification schemes are discussed and an alternative model for the exploration of subtypes is proposed. This model builds on the means of the elaborated methods used in section 4.3 for breast cancer. The novel design of the cohort and the resulting decision list allow an accurate classification of primary and metastatic prostate cancer samples. Additionally, the EDL sheds new light into the underlying machinery driving prostate cancer progression and metastasis.

#### 4.4.1 Designing a Cohort

As described in section 2.1, there exist a broad variety of prostate cancer classification schemes and proposed methods yielding potential subtypes. While some methods rely on expression data only, others take advantage of additional genomic data types or meta information, such as pathways information. Within their study You et al [140] compared their identified subtypes to those propose by Tomlins et al [127] and the TCGA Network [93]. The overlap of the predicted subtypes for all three studies is shown in figure 4.10. As it is clear to see, there barely exists any overlap, obtruding the question of reliability for the determined subtypes. While the TCGA network did not provide any survival or relapse statistic, Tomlins et al could identify a small but significant trend towards prostate cancer-specific mortality free survival for one of their identified subtypes (called “triple negative”). At least, only You et al could identify a decreased probability for metastatic free survival for their PCS1 subtype, which is characterized by the expression of luminal cell associated genes.

Despite from the problems discussed in section 2.2, all three studies rely on different data types. While Tomlins et al and You et al used expression data only, the TCGA Network utilized multiple data types. As all data, except for the TCGA, is derived from different sources, batch effects could heavily bias the outcome. Also, for most of the prostate cancer data available, there is a lack of clinical and followup data. For example, the TCGA only has follow up data for 2% of their prostate cancer patients available. Due to those problems, it might be helpful not to identify potential subtypes first, but to examine functional differences between primary and mCRPC. Also, to avoid side effects, all samples should be processed in the same way and

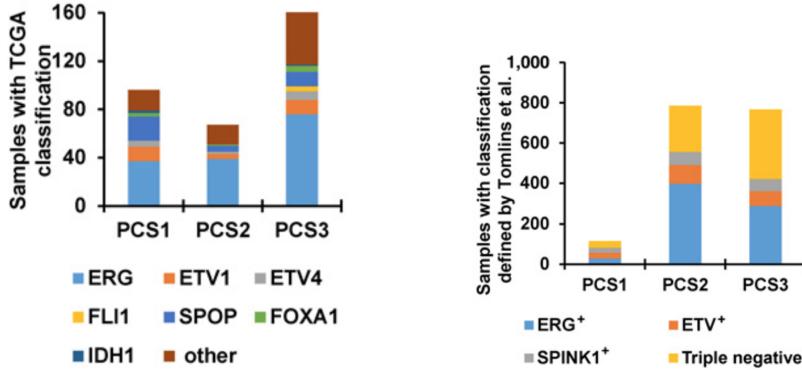


Figure 4.10: The suggested prostate cancer subtypes. In both figures the subtypes by You et al are compared to the findings by the TCGA Network (left) and Tomlins et al (right). PCS1, PCS2 and PCS3 denote the findings made by You et al, other labels are adapted from the original publication. Figure adapted from You et al [140].

meta predictors should be used as done in for the breast cancer data (4.3.1).

To overcome these obstacles, a cohort of 483 samples is analyzed by the means of the methods developed in this study. The cohort consists of the 333 primary tumor samples used by the TCGA Network [93] and additional 150 mCRPC samples, used by Robinson et al [107]. Data for the 333 samples is derived over the FirebrowseR software and the remaining 150 samples are obtained from cBioPortals raw data download section. As discussed in the Firehose and cBioPortal sections (3.4.2, 3.4.3), both portals rely on identical datasets for their analyses, allowing an easy merge of their raw data sets. Both cohorts provide CNV, mutational and fusion data. To merge them, a set of genes commonly altered within prostate cancer is created. This set serves as predictor variables in the EDL model. The list of 49 genes can be found in the appendix A.2. For each gene, all alterations have been assessed. Gene fusions have been encoded binary, denoting if a gene for a sample is fused with any other gene. As proceeded with the breast cancer samples, the CNV status has been decomposed to five states. Namely homozygous deletion, heterozygous loss, no change (diploid), gain and amplification. As mutational data can directly be used as binary predictor, no conversion is required. This resulted in 343 binary predicted variables, 49 for mutations, 49 for gene fusions and 245 for the CNV status. Samples have been labeled according to their cohort, primary or mCRPC.

Table 4.6: Prostate Cancer Classification performance on primary and mCRPC labels. Given measures are the kappa statistic, accuracy and their SDs. All results are carried out by 10-fold cross-validation

	Kappa	Kappa SD	Accuarcy	Accuracy SD
EDL	0.877	0.029	0.950	0.013
Multinom Reg	0.751	0.111	0.896	0.048
Random Forest	0.922	0.069	0.967	0.028
Class Tree	0.775	0.099	0.911	0.039
SVM	0.917	0.068	0.965	0.028

#### 4.4.2 Classifying Primary & mCRPC Samples

For 483 samples and the corresponding 343 features decision rules of cardinality 1, 2 and a minimum support of 1% have been generated, resulting in 1,817 rules. The optimization using the genetic algorithm was run for 100,000,000 generations, with a population size of 1,817 decision lists. The best performing decision list was found to be of length 14. Again, all other models were tuned as described in 4.1 and the given results are determined by 10-fold cross-validation. As shown in table 4.6, the EDL (kappa: 0.877) clearly outperforms the multinomial regression (kappa: 0.751) and the classification tree (kappa: 0.775). Compared to the more advanced models, SVM (kappa: 0.917) and random forest (kappa: 0.922), the EDL performs almost as good. Also the EDL maintains the lowest standard deviation of all cross-validation runs, yielding more reliable results. Additionally to the table, the classification performance has been visualized using boxplots depicted in figure 4.11.

Eventually a final decision list, based on all samples, has been generated. The decision list and the classification graph is depicted in figure 4.12. Considering that mCRPC is known to be more heterogeneous than the primary type, the EDL could identify four rules which are specific to samples of the non metastatic entity.

#### Primary Subtype

Interestingly the first rule of the decision list is specific to primary prostate cancer. As the mCRPC samples are considered to be more heterogeneous, this first rule, with a coverage of 9.5%, reveals outstanding characteristics for the primary cohort. The rule consists of heterozygous losses for the tumor suppressor gene *TP53* and the *AR* pathway member *NCOR1*. The second rule for primary samples consists of gains for Phosphatidylinositol-4,5-

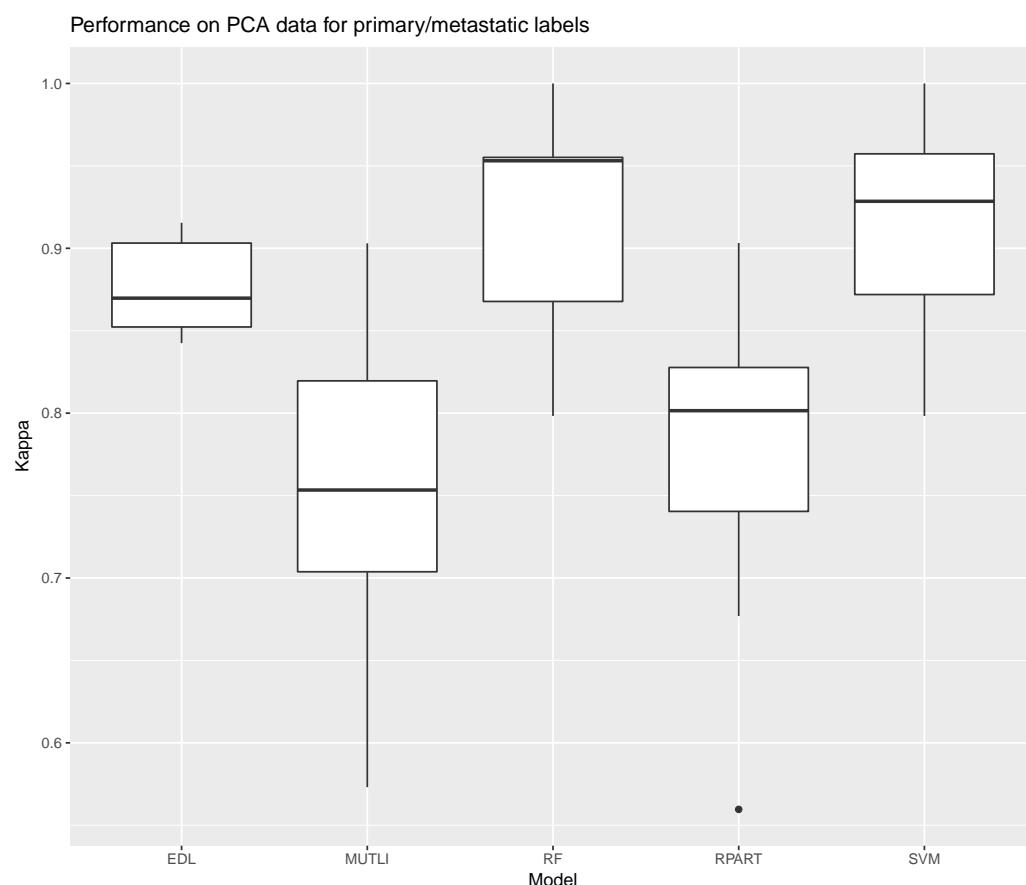


Figure 4.11: Boxplots indicate the classification performance (kappa statistic) of all five models for the Prostate cancer samples with primary and metastatic labels. Scattering indicates the 10-fold cross-validation.

Bisphosphate 3-Kinase Catalytic Subunit Beta (*PIK3CB*) and *ETV1* only. *PIK3CB*, a PI3K signaling pathway member, is likely to be upregulated due Phosphatase And Tensin Homolog (*PTEN*) loss (not covered by the list) [93]. The *ETV1* gain, an ETS transcription factor family (ETS) member, is likely to cover the more rare ETS<sup>+</sup> samples [127]. The last two rules utilize Zinc Finger Homeobox 3 (*ZFHX3*) and *ERG*. As *ZFHX3* is reported to be frequently mutated within primary prostate tumor samples [6], this heterogeneous loss might be the result of such mutations. The homozygous deletion of *ERG* covers most of the identified samples. This is not surprising, as an *ERG* fusion is observed in almost 50% of all primary tumor cases [128, 93], which could result in such a loss. Finally, for 158 primary tumor samples no decision rule could be identified and therefore they are characterized by the default rule.

### mCRPC Subtype

While a couple of metastatic samples are classified by rules only covering a small fraction of samples, the majority is identified by a single rule. This rule, the second rule in the list, utilizes the amplification of *AR* to identify 52% of all metastatic samples. The remaining half of the metastatic samples is largely identified by mutations to *TP53* and *AR* and CNV changes to Speckle Type BTB/POZ Protein (*SPOP*).

### Final Decision List

Overall, exclusive alterations to *NCOR1*, *PIK3CB* and *ERG* could be detected and utilized on the primary cohort by EDL. Additional analyses revealed that the heterozygous loss of AR regulator *NCOR1* is exclusive to 12.2% of all primary patients and does not occur in metastatic samples (<1%). A *PIK3CB* gain was found 9.5% of the samples. As *PTEN*-deleted tumors likely depend on *PIK3CB*, due to the inhibition of *PIK3CA*, a co-occurring alteration to *PIK3CB* and *PTEN* might effect the PI3K pathway output, as suggested by Schwarz et al [111]. A homozygous deletion of *ERG* was found in 10.4% of all primary samples, promoting a fusion with potential partners such as *TMPRSS2*.

In comparison, the only events which could be observed exclusive for mCRPC samples were a gain of *AR* and mutations of *AR*. Although not at the very first position, the gain of *AR* could be observed exclusively in 16.2% of all cases. Also the mutation to *AR* takes place in 5.4% of all mCRPC cases. Other alterations were not exclusive or nearly exclusive to one of the two entities. Hence, the EDL made use of its hierarchical decision model and

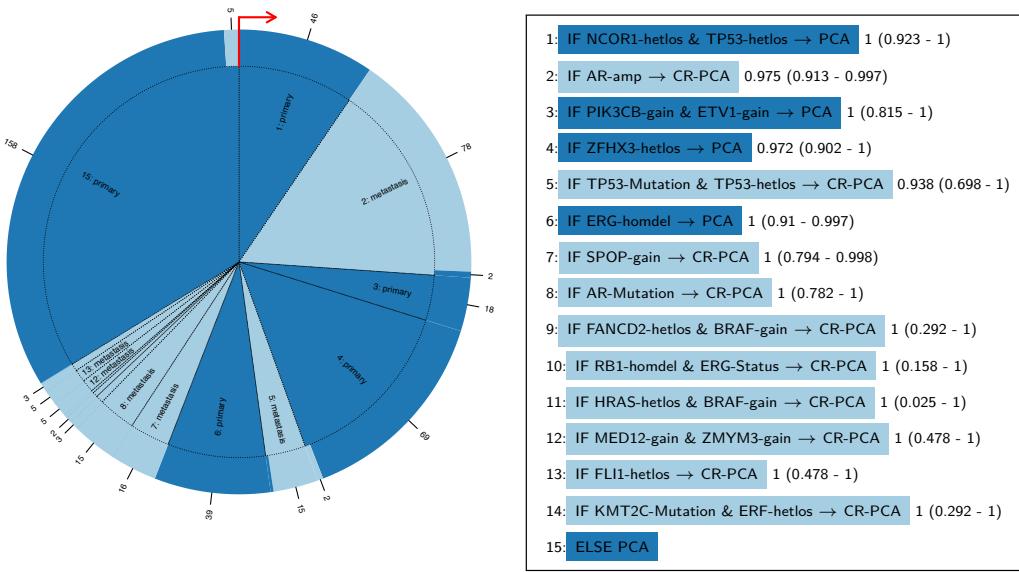


Figure 4.12: The final decision list generated for the prostate cancer data set. On the left hand side, the graph displays each rule applied (starting at 12 o'clock, clockwise) with prediction made (inner circle) and true labels on the outer circle. The numbers indicate the amount of TP samples (numbers < 1 omitted, due to reasons of space). Coloring is done regarding the subtype and matched between both lists. The decision list on the right hand side shows the classification rules, precision and the corresponding 95% confidence intervals.

excluded the FPs first, before integrating an impure rule.

# Chapter 5

## Discussion

### 5.1 Data Aggregation & Normalization

With the spread of the internet, the process of data sharing became key in the field of bio-medical research. Over the past years several publicly available platforms emerged. While, in the beginning, issue centric information to a problem has been provided as a database [121], recently emerged architectures act more as a dump for high-throughput data [66]. This led to questions, which have not been addressed by the research community before. Hence, the researcher is not only confronted the evaluation of the data, but more with a complete workflow requiring problem-specific fine tuning on several levels. To overcome such obstacles in the field of cancer research, the Firehose Pipeline has been implemented, enabling the access to almost analyses ready data sets. But still, depending on the problem, several steps are required to work with data sets derived from multiple high-throughput technology resources.

To address theses problems, several approaches have been introduced, spanning a broad space of application. Frameworks like TCGA2STAT [131], RTCGAToolbox [109] and TCGA-Assembler [146] aim to enable easy access to such data and pre-processing procedures provided by TCGA and Firehose Pipeline, respectively. Comparing those software packages to FirebrowseR, the first difference is the underlying software architecture, which keeps FirebrowseR updated. Non of the other packages provides an automated update mechanism, often leaving the software product behind the API schedule. When comparing the projects data sources, TCGA2STAT, TCGA-Assembler and FirebrowseR all rest upon Firehose data only, while RTCGAToolbox incorporates both, Firehose and TCGA data. This has the advantage, that different levels of data can be accessed, which is not possible

when relying on Firehose data only. Additionally RTCGAToolbox already implements linear models and empirical Bayesian methods provided through the limma [119] package, to determine differentially expressed genes. Based on these models, a method for survival analyses is provided as well. Just as RTCGAToolbox, TCGA-Assembler provides methods to merge different data types into a single object, for multi platform analyses. Additionally, methods to compare methylation data, derived from different platforms, are provided. Just as FirebrowseR, TCGA2STAT provides no additional functionalities. Despite from that, there are a number important differences between FirebrowseR and all other packages. As FirebrowseR is the only package, which relies on Firehose data only, while being fully compatible with the Firebrowse API. While other packages require the user to download whole data dumps organized by cohort and/or technology, FirebrowseR enables the targeted download of pre-defined samples, genes and data types. This has the advantage, that no data overhead is obtained. Vice versa, a pre-defined set of genes for a given technology must be provided, as the download of all genomic information over an API would exceed its capacities. Also, as for each data type a function on its own is provided by the Firehose Pipeline, FirebrowseR realizes the download for each data type through a function on its own. While this has the benefit of remaining compatible with the API, it could complicate the first steps, if the user is new to TCGA and/or Firebrowse.

The cBioPortal [16] provides an intuitive web-interface to analyse TCGA and Firehose data on a coarse level. Additionally, the cohorts provided can be downloaded for offline analyses, just as the results themselves. When comparing cBioPortal to the other software tools discussed, their different purpose becomes clear. As the other tools are intended to make the data available to the R programming environment, cBioPortal aims provides an exploratory data-laboratory. Additionally, a software package for the R programming environment is available [61], making cBioPortal functionalities accessible from R. This package enables the access to the data portal for additional investigations within the programming environment. While Web-TCGA [32], from which the normalization methods have been adapted, does not provide such an abundance of features, their intended use remains identical. With both utilities the user is able to investigate data on a coarse level, which can be further downstreamed within the R programming environment. In direct comparison, Web-TCGA does not provide survival and co-expression analyses. However, these analyses can easily be added from within the R environment by utilizing the FirebrowseR package.

## 5.2 Evolutionary Decision List

The introduced EDL provides a powerful, yet easy to interpret alternative to existing methods. While the model showed competitive results, when compared to state of the art machine learning methods (section 4), there still exists issues which need to be addressed in the future.

Technically the EDL is a machine learning model which consists out of two different models. While the first one solves an association rule learning problem, the second model solves an optimization problem when constructing the decision list. For the extraction of association rules from databases, a wide variety of algorithms is available. The most prominent once are the Apriori algorithm [1], the Eclat algorithm [142] and the FP-growth algorithm [10]. Although all algorithms perform an identical task, the algorithms choice still remains important, when taking the database size into account. In this study, only relatively small data sets have been investigated. While ranging from several hundred, to several thousand samples and being described through several hundred predictor variables, the FP-growth algorithm performed sufficient. The particular choice becomes important, if the data sets become larger. Therefore, the algorithm should be chosen with respect to the underlying architecture and desired rule criteria [59]. That is, some algorithms are designed for parallel or distributed computation, while other show better performance when used to generate longer rules. Despite from the aspect of scaling, different concepts of association rules could be taken into account. For example, association rules covering multiple relations [36, 138] or mining only context specific rules [112] should be investigated, as they might improve the models performance with reference to accuracy and speed. The second problem, the optimization of a decision list, is carried out by a simple genetic algorithm. In this study crossover and mutation rates have been hold fixed, also the population size has always been hold constant, with respect to the number of generated rules. As genetic algorithms have been around for more than 25 year now [70], there exists a large number of problem oriented approaches. Hence, mutation and crossover rates could be chosen with respect to the current performance or performance progression during the evolutionary procedure [102, 73]. Additionally, the proposed system operates on decision lists of a fixed length only, adding an additional tuning parameter to the model. Therefore, by allowing crossover operations without respect to the list length, the system could automatically discover lists of optimal size. While this would remove an additional tuning parameter, it introduces a problem termed bloating [41, 102]. Bloating can result in extreme long decision lists, being only marginal superior to short and intuitive list.

When comparing the introduced model to the Bayesian Rule List (BRL)

by Letham and Yang<sup>1</sup> [78, 139], the EDL is missing a measure of class probability. This is due to the simple genetic algorithm, which does not allow probability estimates, but enables an easy implementation of the multi class classification problem, which is not given for the BRL. Additional extensions, based on these foundations, could be the monotone property. This property has been introduced by Wang et al [132], where the class probabilities are monotonically decreasing with descending list depth. This could be beneficial for risk estimations of a single event in relation to additional observations. Other properties, introduced by Goessling et al [49], are the routing of the direction for the decision list. In the presented concept, one possible option is the learning with a fixed default class, where the search space would be “carved out” for informative rules. The other option only learns rules for one class, collapsing all remaining ones into a default class. This procedure would be useful, when estimating likelihoods for a single group only.

The EDL embodies a model, fulfilling the criteria introduced in 3.6.4. Hence, it provides i) transparancy, as the final model is simulatable, ii) decomposability, as its input is intuitive to the user, iii) algorithmicly transparent, as the algorithm is a simple echo of the darwinian evolution theory and iv) post-hoc interpretable as the output is human readable and can be visualized easily. Further, it can be modified easily, by changing the underlying concept of association rules, the set of class labels taken into account or by changing the underlying optimization algorithm.

### 5.3 Breast Cancer Findings

For the analysis of breast cancer subtypes, the EDL represents an appropriate tool. In terms of accuracy it outperforms the multinomial regression and the classification tree, while being a bit less accurate than the SVM and the random forest. Additionally, the EDL was able to identify almost all driving alterations, by which the PAM50 subtypes are defined. Hence, for the detection of both luminal subtypes, the decision list utilized the overexpression of *ESR1* as a marker. *ESR1* is used to define these two breast cancer subtypes, as ER<sup>+</sup> cancer cells depend on estrogen for their growth [96, 26]. Additionally, *CCNB1* has been chosen as predictor to distinguish the luminal B from luminal A samples. *CCNB1* is a proliferation regulator and is considered to be the driving force making the luminal B subtype more aggressive, than the luminal A subtype [98, 99]. Also, *PgR* hast been chosen as a marker for the detection of the luminal A subtype (PR<sup>+</sup> subtype). Together with *ESR1* this gene is utilized to determine an appropriate hormone therapy form [108, 118].

---

<sup>1</sup>The model is identical, yet the implementation differs.

While the EDL was capable to detect the genes by which the luminal subtypes are defined, it missed the detection of *ERB2*. While *ERB2* defines the Her2 subtype, several Her2 samples are reported with an overexpression of *ESR1* and no overexpression of *ERB2* itself, making the clear separation difficult [57, 108]. As *ESR1* was utilized within the first rule of the list, this might explain why here no *ERB2* marker was used, as *ESR1* in combination with Melanoma Inhibitory Activity (*MIA*) and *FOXC1* might yield a better classification performance.

While the EDL was able to achieve a competitive classification performance, it only utilized expression data (with the exception of a *PNMT* gain) within the final decision list. Although mutational and CNV data has been incorporated as well. Hence, it was expected that mutations and CNVs of the genes *GATA3*, *FOXA1*, *PIK3CA* and *MAP3K1* would have been considered for classification as well, as they have been reported to be exclusively altered in the luminal subtypes [92, 2]. For triple negative breast cancer samples (basal like), a *MYBL2* overexpression was identified as splitting criterion. *MYBL2* is proliferation marker which has already been observed overexpressed in breast cancer [126]. Due to ambiguous patterns for the Her2 subtype, a combination of *ESR1*, *FOXC1* and *MIA* was chosen to identify most of Her2 samples.

In sum, the EDL correctly identified the driver genes for luminal subtypes and the true marker gene to distinguish the luminal A from the luminal B subtype. In addition known marker genes for the basal subtype could be re-identified correctly, while the significant marker for the Her2 subtype has been missed. Further, a novel signature of genes (associated with a poor disease outcome in general) for the normal subtype was identified, reliably separating those samples from the others.

## 5.4 Prostate Cancer Findings

After it has been shown that the EDL is an appropriate classifier in general and capable of classifying cancer subtypes, with the additional ability to unveil the important, subtype specific, predictors, a final evaluation on the aggregated data set of prostate cancer samples has been performed. During the 10-fold cross validation runs the EDL achieved results which compared well to the state of the art models, SVM and random forest. It clearly outperformed both models which are considered interpretable, while holding the properties introduced in 3.6.4. Further, there was no other model showing such a low degree of SD during cross-validation, highlighting the stability for classification.

The finally aggregated decision list revealed known and novel finding between primary and metastatic prostate cancer samples.

It is known that primary prostate cancer suffers less frequent from mutations and CNVs compared to mCRPC [107, 93, 54]. However, the EDL identified a decision rule to distinguish primary from metastatic cases by utilizing several genes affected through mutations and CNVs. In particular it identified a subgroup of samples by two specific gains, which are only observed in the primary cohort. The genes utilized by that first rule, *TP53* and *NCOR1*, have previously been observed in both cohorts, but not attracted attention as a unique characteristic [107, 93] for any subtype. While in combination, their heterozygous loss is specific to 10% of all primary samples. This is contrary to observations made in breast cancer, where *NCOR1* has been reported mutated and differentially expressed in lymph node metastasis [144]. Another important finding is the identification of the homozygous *ERG* deletion. While a gene fusion between *TMPRSS2* and *ERG* is observed in 50% of all primary prostate cancer samples [6, 129], the partial deletion of *ERG* can be considered as a precursor for this event. Another novel finding is that the *AR* was found to be gained in more than 50% of all mCRPC samples, but not in the previously identified set of primary samples, harbouring a *NCOR1/TP53* variant. It is the *AR* which is therapeutically drugged, but to which mCRPC patients develop a resistance. As the *AR* is normally observed gained in both states of the disease [123, 113], it has not been under consideration as a distinguishing marker. On the other hand, the *AR* regulator *NCOR1* was found to be exclusive for primary samples, which might determine a preliminary stage for the gain of *AR*.

In comparison to other studies which proposed prostate cancer subtypes [127, 93, 140], this study differs as it investigates potential subtypes by assuming primary and metastatic prostate cancer as given class labels, which are then investigated by the EDL model. Compared to the other clustering approaches, this procedure seems more target-aimed, as all the other studies totally disagreed on the their identified subtypes (figure 4.10). Further, it remains unclear if mCRPC samples had been included. While there have been subtypes identified which differ in survival [140, 127, 84], such an analysis could not be provided here, as the data is not available for the investigated cohorts. Additionally, there is no data available whether the patient has already undergone a therapy, which could bias the outcome. Also, prior studies mostly defined prostate cancer subtypes by the *TMPRSS2:ETS* fusion status. This allowed for correlations with survival probability, but led to conflicts with respect to grade and the probability of forming metastasis and [75, 54, 125].

When inspecting the decision list itself (figure 4.12), the biggest group

consists out of 158 primary samples. For these samples no specific alteration could be found. Here additional investigations should be performed, including a broader range of potential genes and genomic data types. Further, due to its heterogeneity [107, 54], several rules applying for only a small portion of mCRPC samples, have been utilized. These rules could potentially be collapsed by identifying common alterations for the mCRPC samples, or by expanding the rule cardinality. It is to say that all findings are based in computational analyses and require a wet laboratory evaluation, based on an independent cohort. Nevertheless, the findings revealed by the EDL shed novel light into the yet sparsely understood process of mCRPC development. Hence, the combination of *TP53* and *NCOR1* as novel distinguishing marker for primary cases, as well as the *AR* gain for metastatic cases deserve and require additional investigations, but provide a promising starting point for follow-up studies.



# Chapter 6

## Conclusion

In conclusion this study showed the usability of multi omics data types for cancer subtype classification. To achieve this goal, it has been shown that data obtained using the newly developed FirebrowseR software and normalization methods adapted from Web-TCGA provide an ideal foundation for such an analyses.

A newly developed classifier, the evolutionary decision list, has been proven as reliable model for cancer subtype classification, achieving competitive results to state of the art machine learning models. Through its structure, the model automatically provides a build in mechanism for feature selection and model interpretation which is naïvely amenable to any user with knowledge of the problem domain. Benchmarks run on well known example data sets underlined the models performance in competition with established highly accurate and easy-to-interpret models.

The combination of FirebrowseR, data normalization methods and the EDL was able to re-identify the known breast cancer subtypes and highlighted the important marker alterations of *ERB2*, *ESR1*, *PgR* and *CCNB1*, by which these subtypes are defined. For the novel classification of primary and metastatic prostate cancer samples, the method utilized well known genes, which have not been considered as unique characteristic to one of the two cohorts. Hence, a combined gain of *TP53* and *NCOR1* is specific to primary prostate cancer, while a gain of the *AR* describes more than 50% of all mCRPC samples.



# Bibliography

- [1] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.
- [2] Shantanu Banerji, Kristian Cibulkis, Claudia Rangel-Escareno, Kristin K Brown, Scott L Carter, Abbie M Frederick, Michael S Lawrence, Andrey Y Sivachenko, Carrie Sougnez, Lihua Zou, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486(7403):405–409, 2012.
- [3] Christopher E Barbieri and Scott A Tomlins. The prostate cancer genome: perspectives and potential. In *Urologic Oncology: Seminars and Original Investigations*, volume 32, pages 53–e15. Elsevier, 2014.
- [4] Katrina R Bauer, Monica Brown, Rosemary D Cress, Carol A Parise, and Vincent Caggiano. Descriptive analysis of estrogen receptor (er)-negative, progesterone receptor (pr)-negative, and her2-negative invasive breast cancer, the so-called triple-negative phenotype. *Cancer*, 109(9):1721–1728, 2007.
- [5] Kristin P Bennett. Global tree optimization: A non-greedy decision tree algorithm. *Computing Science and Statistics*, pages 156–156, 1994.
- [6] Michael F Berger, Michael S Lawrence, Francesca Demichelis, Yotam Drier, Kristian Cibulkis, Andrey Y Sivachenko, Andrea Sboner, Raquel Esgueva, Dorothee Pflueger, Carrie Sougnez, et al. The genomic complexity of primary human prostate cancer. *Nature*, 470(7333):214–220, 2011.
- [7] Tim Berners-Lee, Roy Fielding, and Henrik Frystyk. Hypertext transfer protocol—http/1.0. Technical report, 1996.

- [8] Marko Bohanec and Vladislav Rajkovic. Knowledge acquisition and explanation for multi-attribute decision making. In *8th Intl Workshop on Expert Systems and their Applications*, pages 59–78, 1988.
- [9] Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200, 1992.
- [10] Christian Borgelt. An implementation of the fp-growth algorithm. In *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, pages 1–5. ACM, 2005.
- [11] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- [12] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [13] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.
- [14] Lisa A Carey, E Claire Dees, Lynda Sawyer, Lisa Gatti, Dominic T Moore, Frances Collichio, David W Ollila, Carolyn I Sartor, Mark L Graham, and Charles M Perou. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. *Clinical cancer research*, 13(8):2329–2334, 2007.
- [15] Lisa A Carey, Charles M Perou, Chad A Livasy, Lynn G Dressler, David Cowan, Kathleen Conway, Gamze Karaca, Melissa A Troester, Chiu Kit Tse, Sharon Edmiston, et al. Race, breast cancer subtypes, and survival in the carolina breast cancer study. *Jama*, 295(21):2492–2502, 2006.
- [16] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, 2012.
- [17] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.
- [18] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Sen-babaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133, 2013.

- [19] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.
- [20] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
- [21] Mary McNaughton Collins and Michael J Barry. Controversies in prostate cancer screening: analogies to the early lung cancer screening debate. *Jama*, 276(24):1976–1979, 1996.
- [22] Pierre-Emmanuel Colombo, Fernanda Milanezi, Britta Weigelt, and Jorge S Reis-Filho. Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. *Breast Cancer Research*, 13(3):212, 2011.
- [23] Colin S Cooper, Rosalind Eeles, David C Wedge, Peter Van Loo, Gunes Gundem, Ludmil B Alexandrov, Barbara Kremeyer, Adam Butler, Andrew G Lynch, Niedzica Camacho, et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nature genetics*, 47(4):367–372, 2015.
- [24] Matthew R Cooperberg, Jeanette M Broering, and Peter R Carroll. Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. *Journal of the National Cancer Institute*, 101(12):878–887, 2009.
- [25] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [26] Chad J Creighton et al. The molecular profile of luminal b breast cancer. *Biologics*, 6(2):289–297, 2012.
- [27] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.
- [28] Anthony V D’amico, Richard Whittington, S Bruce Malkowicz, Delray Schultz, Kenneth Blank, Gregory A Broderick, John E Tomaszewski, Andrew A Renshaw, Irving Kaplan, Clair J Beard, et al. Biochemical

- outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *Jama*, 280(11):969–974, 1998.
- [29] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.
  - [30] Johann Sebastian De Bono, Stephane Oudard, Mustafa Ozguroglu, Steinbjørn Hansen, Jean-Pascal Machiels, Ivo Kocak, Gwenaëlle Gravis, Istvan Bodrogi, Mary J Mackenzie, Liji Shen, et al. Prednisone plus cabazitaxel or mitoxantrone for metastatic castration-resistant prostate cancer progressing after docetaxel treatment: a randomised open-label trial. *The Lancet*, 376(9747):1147–1154, 2010.
  - [31] Mario Deng, Johannes Brägelmann, Ivan Kryukov, Nuno Saraiva-Agostinho, and Sven Perner. Firebrowser: an r client to the broad institute’s firehose pipeline. *Database*, 2017:baw160, 2017.
  - [32] Mario Deng, Johannes Brägelmann, Joachim L Schultze, and Sven Perner. Web-tcga: an online platform for integrated analysis of molecular cancer data sets. *BMC bioinformatics*, 17(1):72, 2016.
  - [33] Carsten Denkert, Jens Huober, Sibylle Loibl, Judith Prinzler, Ralf Kronenwett, Silvia Darb-Esfahani, Jan C Bräse, Christine Solbach, Keyur Mehta, Peter A Fasching, et al. Her2 and esr1 mrna expression levels and response to neoadjuvant trastuzumab plus chemotherapy in patients with primary breast cancer. *Breast Cancer Research*, 15(1):R11, 2013.
  - [34] Theresa A DiMeo, Kristen Anderson, Pushkar Phadke, Chang Feng, Charles M Perou, Steven Naber, and Charlotte Kuperwasser. A novel lung metastasis signature links wnt signaling with cancer cell self-renewal and epithelial-mesenchymal transition in basal-like breast cancer. *Cancer research*, 69(13):5364–5373, 2009.
  - [35] Brenton R Dobin, Sheng Li, Christopher E Mason, Sara Olson, Dmitri Pervouchine, Cricket A Sloan, Xintao Wei, Lijun Zhan, and Rafael A Irizarry. A benchmark for rna-seq quantification pipelines.
  - [36] Sašo Džeroski. Multi-relational data mining: an introduction. *ACM SIGKDD Explorations Newsletter*, 5(1):1–16, 2003.

- [37] Dirk Eddelbuettel. *Seamless R and C++ integration with Rcpp*. Springer, 2013.
- [38] Dirk Eddelbuettel and Romain Francois. Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, 40(1):1–18, 2011.
- [39] Paul AW Edwards. Fusion genes and chromosome translocations in the common epithelial cancers. *The Journal of pathology*, 220(2):244–254, 2010.
- [40] Pedro G Espejo, Cristóbal Romero, Sebastián Ventura, and César Hervás. Induction of classification rules with grammar-based genetic programming. In *Conference on Machine Intelligence*, pages 596–601, 2005.
- [41] Pedro G Espejo, Sebastián Ventura, and Francisco Herrera. A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 40(2):121–144, 2010.
- [42] Roy Fielding, Jim Gettys, Jeffrey Mogul, Henrik Frystyk, Larry Masinter, Paul Leach, and Tim Berners-Lee. Hypertext transfer protocol—http/1.1. Technical report, 1999.
- [43] Karim Fizazi, Michael Carducci, Matthew Smith, Ronaldo Damião, Janet Brown, Lawrence Karsh, Piotr Milecki, Neal Shore, Michael Rader, Huei Wang, et al. Denosumab versus zoledronic acid for treatment of bone metastases in men with castration-resistant prostate cancer: a randomised, double-blind study. *The Lancet*, 377(9768):813–822, 2011.
- [44] Karim Fizazi, Howard I Scher, Arturo Molina, Christopher J Logothetis, Kim N Chi, Robert J Jones, John N Staffurth, Scott North, Nicholas J Vogelzang, Fred Saad, et al. Abiraterone acetate for treatment of metastatic castration-resistant prostate cancer: final overall survival analysis of the cou-aa-301 randomised, double-blind, placebo-controlled phase 3 study. *The lancet oncology*, 13(10):983–992, 2012.
- [45] Eibe Frank and Ian H Witten. Generating accurate rule sets without global optimization. 1998.
- [46] Mark A Friedl and Carla E Brodley. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3):399–409, 1997.

- [47] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.
- [48] Brian F Gage, Amy D Waterman, William Shannon, Michael Boechler, Michael W Rich, and Martha J Radford. Validation of clinical classification schemes for predicting stroke: results from the national registry of atrial fibrillation. *Jama*, 285(22):2864–2870, 2001.
- [49] Marc Goessling and Shan Kang. Directional decision lists. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2762–2766. IEEE, 2015.
- [50] A 2011 Goldhirsch, WC Wood, AS Coates, RD Gelber, B Thürlimann, H-J Senn, et al. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2011. *Annals of oncology*, page mdr304, 2011.
- [51] Aron Goldhirsch, WC Wood, RD Gelber, AS Coates, B Thürlimann, H-J Senn, et al. Progress and promise: highlights of the international expert consensus on the primary therapy of early breast cancer 2007. *Annals of oncology*, 18(7):1133–1144, 2007.
- [52] Martin Gillery. Bioinformatics: Sequence and genome analysis, david w. mount. cold spring harbor, ny: Cold spring harbor laboratory press, 2004, 692 pp.. isbn 0-87969-712-1. *Clinical Chemistry*, 51(11):2219–2219, 2005.
- [53] British Government. *Report on the Loss of the S.S. Titanic*. St Martin’s Press, 1998.
- [54] Catherine S Grasso, Yi-Mi Wu, Dan R Robinson, Xuhong Cao, Saravana M Dhanasekaran, Amjad P Khan, Michael J Quist, Xiaojun Jing, Robert J Lonigro, J Chad Brenner, et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, 487(7406):239–243, 2012.
- [55] Peter Grimm, Ignace Billiet, David Bostwick, Adam P Dicker, Steven Frank, Jos Immerzeel, Mira Keyes, Patrick Kupelian, W Robert Lee, Stefan Machtens, et al. Comparative analysis of prostate-specific antigen free survival outcomes for patients with low, intermediate and high risk prostate cancer treatment by radical therapy. results from the

- prostate cancer results study group. *BJU international*, 109(s1):22–29, 2012.
- [56] M Elizabeth H Hammond, Daniel F Hayes, Mitch Dowsett, D Craig Allred, Karen L Hagerty, Sunil Badve, Patrick L Fitzgibbons, Glenn Francis, Neil S Goldstein, Malcolm Hayes, et al. American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). *Archives of pathology & laboratory medicine*, 134(7):e48–e72, 2010.
- [57] Reina Haque, Syed A Ahmed, Galina Inzhakova, Jiaxiao Shi, Chantal Avila, Jonathan Polikoff, Leslie Bernstein, Shelley M Enger, and Michael F Press. Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. *Cancer Epidemiology and Prevention Biomarkers*, 21(10):1848–1855, 2012.
- [58] Julia H Hayes and Michael J Barry. Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence. *Jama*, 311(11):1143–1149, 2014.
- [59] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter*, 2(1):58–64, 2000.
- [60] M Hofmann, O Stoss, D Shi, R Büttner, M Van De Vijver, W Kim, A Ochiai, J Rüschoff, and T Henkel. Assessment of a her2 scoring system for gastric cancer: results from a validation study. *Histopathology*, 52(7):797–805, 2008.
- [61] A Jacobsen. cgdsr: R-based api for accessing the mskcc cancer genomics data server (cgds). *R package version*, 1:30, 2013.
- [62] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.
- [63] Margaret S Joesting, Steve Perrin, Brian Elenbaas, Stephen E Fawell, Jeffrey S Rubin, Omar E Franco, Simon W Hayward, Gerald R Cunha, and Paul C Marker. Identification of sfrp1 as a candidate mediator of stromal-to-epithelial signaling in prostate cancer. *Cancer research*, 65(22):10423–10430, 2005.

- [64] Andrew R Joyce and Bernhard Ø Palsson. The model organism as a system: integrating'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210, 2006.
- [65] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 2013.
- [66] Lavanya Kannan, Marcel Ramos, Angela Re, Nehme El-Hachem, Zhaleh Safikhani, Deena MA Gendoo, Sean Davis, David Gomez-Cabrero, Robert Castelo, Kasper D Hansen, et al. Public data and open source tools for multi-assay genomic investigation of disease. *Briefings in bioinformatics*, page bbv080, 2015.
- [67] Michael W Kattan, James A Eastham, Alan MF Stapleton, Thomas M Wheeler, and Peter T Scardino. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *Journal of the National Cancer Institute*, 90(10):766–771, 1998.
- [68] Tanya Keenan, Beverly Moy, Edmund A Mroz, Kenneth Ross, Andrzej Niemierko, James W Rocco, Steven Isakoff, Leif W Ellisen, and Aditya Bardia. Comparison of the genomic landscape between primary breast cancer in african american versus white women and the association of racial differences with tumor recurrence. *Journal of Clinical Oncology*, 33(31):3621–3627, 2015.
- [69] Khandan Keyomarsi, Susan L Tucker, Thomas A Buchholz, Matthew Callister, YE Ding, Gabriel N Hortobagyi, Isabelle Bedrosian, Christopher Knickerbocker, Wendy Toyofuku, Michael Lowe, et al. Cyclin e and survival in patients with breast cancer. *New England Journal of Medicine*, 347(20):1566–1575, 2002.
- [70] John R Koza. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press, 1992.
- [71] Murray D Krahn, John E Mahoney, Mark H Eckman, John Trachtenberg, Stephen G Pauker, and Allan S Detsky. Screening for prostate cancer: a decision analytic view. *Jama*, 272(10):773–780, 1994.
- [72] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

- [73] Gabriel Kronberger, Stephan Winkler, Michael Affenzeller, Andreas Beham, and Stefan Wagner. On the success rate of crossover operators for genetic programming with offspring selection. In *International Conference on Computer Aided Systems Theory*, pages 793–800. Springer, 2009.
- [74] Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.
- [75] Jacques Lapointe, Chunde Li, John P Higgins, Matt Van De Rijn, Eric Bair, Kelli Montgomery, Michelle Ferrari, Lars Egevad, Walter Rayford, Ulf Bergerheim, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 101(3):811–816, 2004.
- [76] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.
- [77] Brian D Lehmann, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, and Jennifer A Pietenpol. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation*, 121(7):2750–2767, 2011.
- [78] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.
- [79] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.
- [80] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.
- [81] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

- [82] Olvi L Mangasarian, W Nick Street, and William H Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.
- [83] Marilyn Marcione. Prostate testing’s dark side: Men who were harmed, 2011.
- [84] Elke K Markert, Hideaki Mizuno, Alexei Vazquez, and Arnold J Levine. Molecular classification of prostate cancer using curated expression signatures. *Proceedings of the National Academy of Sciences*, 108(52):21276–21281, 2011.
- [85] Arvind Singh Mer, Daniel Klevebring, Henrik Grönberg, and Mattias Rantalainen. Study design requirements for rna sequencing-based breast cancer diagnostics. *Scientific reports*, 6, 2016.
- [86] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, 12(4):R41, 2011.
- [87] Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.
- [88] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.
- [89] ABM Moniruzzaman and Syed Akhter Hossain. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv preprint arXiv:1307.0191*, 2013.
- [90] Ann Mullally and Jerome Ritz. Beyond hla: the significance of genomic variation for allogeneic hematopoietic stem cell transplantation. *Blood*, 109(4):1355–1362, 2007.
- [91] Naomi Nakayama, Kentaro Nakayama, Yeasmin Shamima, Masako Ishikawa, Atsuko Katagiri, Kouji Iida, and Khoji Miyazaki. Gene amplification ccne1 is related to poor survival and potential therapeutic target in ovarian cancer. *Cancer*, 116(11):2621–2634, 2010.

- [92] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418):61, 2012.
- [93] Cancer Genome Atlas Research Network et al. The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025, 2015.
- [94] William D Orsi, Virginia P Edgcomb, Glenn D Christman, and Jennifer F Biddle. Gene expression in the deep biosphere. *Nature*, 499(7457):205–208, 2013.
- [95] J Guillermo Paez, Pasi A Jänne, Jeffrey C Lee, Sean Tracy, Heidi Greulich, Stacey Gabriel, Paula Herman, Frederic J Kaye, Neal Lindeman, Titus J Boggon, et al. Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 304(5676):1497–1500, 2004.
- [96] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160–1167, 2009.
- [97] Sven Perner, Francesca Demichelis, Rameen Beroukhim, Folke H Schmidt, Juan-Miguel Mosquera, Sunita Setlur, Joelle Tchinda, Scott A Tomlins, Matthias D Hofer, Kenneth G Pienta, et al. Tmprss2: Erg fusion-associated deletions provide insight into the heterogeneity of prostate cancer. *Cancer research*, 66(17):8337–8341, 2006.
- [98] Charles M Perou, Stefanie S Jeffrey, Matt Van De Rijn, Christian A Rees, Michael B Eisen, Douglas T Ross, Alexander Pergamenschikov, Cheryl F Williams, Shirley X Zhu, Jeffrey CF Lee, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences*, 96(16):9212–9217, 1999.
- [99] Charles M Perou, Stefanie S Jeffrey, Matt Van De Rijn, Christian A Rees, Michael B Eisen, Douglas T Ross, Alexander Pergamenschikov, Cheryl F Williams, Shirley X Zhu, Jeffrey CF Lee, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences*, 96(16):9212–9217, 1999.
- [100] Charles M Perou, Therese Sørlie, Michael B Eisen, Matt van de Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T

- Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.
- [101] Dorothee Pflueger, Stéphane Terry, Andrea Sboner, Lukas Habegger, Raquel Esgueva, Pei-Chun Lin, Maria A Svensson, Naoki Kitabayashi, Benjamin J Moss, Theresa Y MacDonald, et al. Discovery of non-nets gene fusions in human prostate cancer using next-generation rna sequencing. *Genome research*, 21(1):56–67, 2011.
  - [102] Riccardo Poli and William B Langdon. On the search properties of different crossover operators in genetic programming. *Genetic Programming*, pages 293–301, 1998.
  - [103] J. Ross Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.
  - [104] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.
  - [105] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.
  - [106] Ronald L Rivest. Learning decision lists. *Machine learning*, 2(3):229–246, 1987.
  - [107] Dan Robinson, Eliezer M Van Allen, Yi-Mi Wu, Nikolaus Schultz, Robert J Lonigro, Juan-Miguel Mosquera, Bruce Montgomery, Mary-Ellen Taplin, Colin C Pritchard, Gerhardt Attard, et al. Integrative clinical genomics of advanced prostate cancer. *Cell*, 161(5):1215–1228, 2015.
  - [108] Hege G Russnes, Nicholas Navin, James Hicks, and Anne-Lise Borresen-Dale. Insight into the heterogeneity of breast cancer through next-generation sequencing. *The Journal of clinical investigation*, 121(10):3810–3818, 2011.
  - [109] Mehmet Kemal Samur. Rtcgatoolbox: a new tool for exporting tcga firehose data. *PloS one*, 9(9):e106397, 2014.
  - [110] Valerie Schneider and Deanna Church. Genome reference consortium. 2013.
  - [111] Sarit Schwartz, John Wongvipat, Cath B Trigwell, Urs Hancox, Brett S Carver, Vanessa Rodrik-Outmezguine, Marie Will, Paige Yellen, Elisa

- de Stanchina, José Baselga, et al. Feedback suppression of pi3k $\alpha$  signaling in pten-mutated tumors is relieved by selective inhibition of pi3k $\beta$ . *Cancer cell*, 27(1):109–122, 2015.
- [112] Muhammad Shaheen, Muhammad Shahbaz, and Aziz Guergachi. Context based positive and negative spatio-temporal association rule mining. *Knowledge-Based Systems*, 37:261–273, 2013.
- [113] Michael M Shen and Cory Abate-Shen. Molecular genetics of prostate cancer: new prospects for old challenges. *Genes & development*, 24(18):1967–2000, 2010.
- [114] Ronglai Shen, Adam B Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009.
- [115] Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal. Cancer statistics, 2013. *CA: a cancer journal for clinicians*, 63(1):11–30, 2013.
- [116] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2016. *CA: a cancer journal for clinicians*, 66(1):7–30, 2016.
- [117] Dennis J Slamon, Brian Leyland-Jones, Steven Shak, Hank Fuchs, Virginia Paton, Alex Bajamonde, Thomas Fleming, Wolfgang Eiermann, Janet Wolter, Mark Pegram, et al. Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that over-expresses her2. *New England Journal of Medicine*, 344(11):783–792, 2001.
- [118] DJ Slamon. Human breast cancer: correlation of relapse and. *Science*, 3798106(177):235, 1987.
- [119] Gordon K Smyth et al. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1):3, 2004.
- [120] NATIONAL AUDUBON SOCIETY. The audubon society field guide to north american mushrooms. alfred a, 1981.
- [121] Lincoln D Stein. Integrating biological databases. *Nature Reviews Genetics*, 4(5):337–345, 2003.

- [122] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pages 861–870. International Society for Optics and Photonics, 1993.
- [123] MH Eileen Tan, Jun Li, H Eric Xu, Karsten Melcher, and Eu-leong Yong. Androgen receptor: structure, role in prostate cancer and drug discovery. *Acta Pharmacologica Sinica*, 36(1):3–23, 2015.
- [124] Joseph H Taube, Jason I Herschkowitz, Kakajan Komurov, Alicia Y Zhou, Supriya Gupta, Jing Yang, Kimberly Hartwell, Tamer T Onder, Piyush B Gupta, Kurt W Evans, et al. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proceedings of the National Academy of Sciences*, 107(35):15449–15454, 2010.
- [125] Barry S Taylor, Nikolaus Schultz, Haley Hieronymus, Anuradha Gopalan, Yonghong Xiao, Brett S Carver, Vivek K Arora, Poorvi Kaushik, Ethan Cerami, Boris Reva, et al. Integrative genomic profiling of human prostate cancer. *Cancer cell*, 18(1):11–22, 2010.
- [126] AR Thorner, Katherine A Hoadley, JS Parker, S Winkel, RC Millikan, and Charles M Perou. In vitro and in vivo analysis of b-myb in basal-like breast cancer. *Oncogene*, 28(5):742–751, 2009.
- [127] Scott A Tomlins, Mohammed Alshalalfa, Elai Davicioni, Nicholas Erho, Kasra Yousefi, Shuang Zhao, Zaid Haddad, Robert B Den, Adam P Dicker, Bruce J Trock, et al. Characterization of 1577 primary prostate cancers reveals novel biological and clinicopathologic insights into molecular subtypes. *European urology*, 68(4):555–567, 2015.
- [128] Scott A Tomlins, Bharathi Laxman, Saravana M Dhanasekaran, Beth E Helgeson, Xuhong Cao, David S Morris, Anjana Menon, Xiaojun Jing, Qi Cao, Bo Han, et al. Distinct classes of chromosomal rearrangements create oncogenic ets gene fusions in prostate cancer. *Nature*, 448(7153):595–599, 2007.
- [129] Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, et al. Recurrent fusion of tmrss2 and ets transcription factor genes in prostate cancer. *science*, 310(5748):644–648, 2005.

- [130] William N Venables and Brian D Ripley. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.
- [131] Ying-Wooi Wan, Genevera I Allen, and Zhandong Liu. Tcga2stat: simple tcga data access for integrated statistical analysis in r. *Bioinformatics*, page btv677, 2015.
- [132] Fulton Wang and Cynthia Rudin. Causal falling rule lists. *arXiv preprint arXiv:1510.05189*, 2015.
- [133] Xiao-Song Wang, Sunita Shankar, Saravana M Dhanasekaran, Bushra Ateeq, Atsuo T Sasaki, Xiaojun Jing, Daniel Robinson, Qi Cao, John R Prensner, Anastasia K Yocum, et al. Characterization of kras rearrangements in metastatic prostate cancer. *Cancer discovery*, 1(1):35–43, 2011.
- [134] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.
- [135] Rick Weiss. Nih launches cancer genome project. *Washington Post*, 2005.
- [136] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.
- [137] Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.
- [138] Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer, 1997.
- [139] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable bayesian rule lists. *arXiv preprint arXiv:1602.08610*, 2016.
- [140] Sungyong You, Beatrice S Knudsen, Nicholas Erho, Mohammed Alshalaifa, Mandeep Takhar, Hussam Al-deen Ashab, Elai Davicioni, R Jeffrey Karnes, Eric A Klein, Robert B Den, et al. Integrated classification of prostate cancer reveals a novel luminal subtype with poor outcome. *Cancer Research*, 76(17):4948–4958, 2016.

- [141] Min Yu, Aditya Bardia, Ben S Wittner, Shannon L Stott, Malgorzata E Smas, David T Ting, Steven J Isakoff, Jordan C Ciciliano, Marissa N Wells, Ajay M Shah, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *science*, 339(6119):580–584, 2013.
- [142] Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390, 2000.
- [143] Mehdi Zarrei, Jeffrey R MacDonald, Daniele Merico, and Stephen W Scherer. A copy number variation map of the human genome. *Nature Reviews Genetics*, 16(3):172–183, 2015.
- [144] Zhenhuan Zhang, Hiroko Yamashita, Tatsuya Toyama, Hiroshi Suguri, Yoshiaki Ando, Keiko Mita, Maho Hamaguchi, Yasuo Hara, Shunzo Kobayashi, and Hirotaka Iwase. Ncor1 mrna is an independent prognostic factor for breast cancer. *Cancer letters*, 237(1):123–129, 2006.
- [145] Chang-Qi Zhu, Gilda da Cunha Santos, Keyue Ding, Akira Sakurada, Jean-Claude Cutz, Ni Liu, Tong Zhang, Paula Marrano, Marlo Whitehead, Jeremy A Squire, et al. Role of kras and egfr as biomarkers of response to erlotinib in national cancer institute of canada clinical trials group study br. 21. *Journal of clinical oncology*, 26(26):4268–4275, 2008.
- [146] Yitan Zhu, Peng Qiu, and Yuan Ji. Tcga-assembler: open-source software for retrieving and processing tcga data. *Nature methods*, 11(6):599–600, 2014.

# Abbreviations

**AKT1** AKT Serine/Threonine Kinase 1. 91

**APC** APC, WNT Signaling Pathway Regulator. 92, 93

**API** application programming interface. 18, 20–24, 67, 68, 83, 84, *Glossary:* API

**AR** Androgen Receptor. 3, 62, 64, 72, 73, 75, 92

**ATM** ATM Serine/Threonine Kinase. 92, 93

**bagging** Bootstrap Aggregation. 33, 35, *Glossary:* Bootstrap Aggregation

**bp** base pairs. 16, 86, *Glossary:* bp

**BRAF** B-Raf Proto-Oncogene, Serine/Threonine Kinase. 91, 93

**BRCA1** BRCA1, DNA Repair Associated. 91, 92

**BRCA2** BRCA2, DNA Repair Associated. 91, 92

**BRL** Bayesian Rule List. 69, 70

**CCNB1** Cyclin B1. 7, 58, 70, 75

**CCND1** Cyclin D1. 93

**CCNE1** Cyclin E1. 59

**CDK12** Cyclin Dependent Kinase 12. 92

**CDKN1B** Cyclin Dependent Kinase Inhibitor 1B. 91, 92

**CDKN2A** Cyclin Dependent Kinase Inhibitor 2A. 92, 93

**CHD1** Chromodomain Helicase DNA Binding Protein 1. 91, 92

- CI** confidence interval. 37, 89, *Glossary*: confidence intervals
- CNV** Copy Number Variation. 7, 10, 16, 17, 25, 51, 53, 55, 56, 58, 61, 64, 71, 72, *Glossary*: CNV
- CRAN** Comprehensive R Archive Network. *Glossary*: CRAN
- CSV** Comma Separated Values. 21, *Glossary*: CSV
- CTNNB1** Catenin Beta 1. 91
- DBMS** Database Management System. 21, *Glossary*: DBMS
- DNA** deoxyribonucleic acid. 10, 15, 83, 86, *Glossary*: DNA
- EDL** evolutionary decision list. 3, 8, 13, 30, 38, 43, 44, 48, 50, 53, 56, 60–62, 64, 69–73, 75, *Glossary*: EDL
- EGFR** Epidermal Growth Factor Receptor. 51, 54
- EMT** epithelial-to-mesenchymal transition. 58, 59, *Glossary*: epithelial-to-mesenchymal transition
- ERB2** Erb-B2 Receptor Tyrosine Kinase 2. 7, 34, 58, 71, 75
- ERF** ETS2 Repressor Factor. 93
- ERG** ERG, ETS Transcription Factor. 10, 17, 64, 72, 90
- ESR1** Estrogen Receptor 1. 7, 58, 70, 71, 75
- ETS** ETS transcription factor family. 64, 72, *Glossary*: ETS transcription factor family
- ETV1** ETS Variant 1. 17, 64, 90
- ETV4** ETS Variant 4. 90
- FAM175A** Family With Sequence Similarity 175 Member A. 92
- FANCC** Fanconi Anemia Complementation Group C. 92, 93
- FANCD2** Fanconi Anemia Complementation Group D2. 92
- FGFR1** Fibroblast Growth Factor Receptor 1. 53, 55
- FLI1** Fli-1 Proto-Oncogene, ETS Transcription Factor. 90

**FN** false negatives. 40, *Glossary*: false negatives

**FNA** Fine-needle aspiration. 5, *Glossary*: FNA

**FOXA1** Forkhead Box A1. 58, 71, 91

**FOXC1** Forkhead Box C1. 58, 59, 71

**FP** false positives. 40, 64, *Glossary*: false positives

**GATA3** GATA Binding Protein 3. 8, 16, 71

**GNAS** GNAS Complex Locus. 93

**HRAS** HRas Proto-Oncogene, GTPase. 91

**HTTP** Hypertext Transfer Protocol. 21, 22, *Glossary*: HTTP

**IDH1** Isocitrate Dehydrogenase (NADP(+)) 1, Cytosolic. 91

**INDEL** insertion/deletion. 15–17, *Glossary*: INDEL

**IT** Information Technology. 17

**JSON** JavaScript Object Notation. 21, 23, *Glossary*: JSON

**KDM6A** Lysine Demethylase 6A. 91

**KMT2A** Lysine Methyltransferase 2A. 91

**KMT2C** Lysine Methyltransferase 2A. 91, 92

**KMT2D** Lysine Methyltransferase 2D. 91

**KRAS** KRAS Proto-Oncogene, GTPase. 51, 54

**lasso** least absolute shrinkage and selection operator. *Glossary*: least absolute shrinkage and selection operator

**MAF** Mutations Annotation Format. *Glossary*: MAF

**MAP** Maximum a posteriori estimation. 31, *Glossary*: MAP

**MAP3K1** Mitogen-Activated Protein Kinase Kinase Kinase 1. 8, 16, 71

- mCRPC** metastatic, castration-resistant prostate cancer. 3, 10, 60–62, 64, 72, 73, 75, 90
- MED12** Mediator Complex Subunit 12. 91
- MIA** Melanoma Inhibitory Activity. 71
- MLH1** MutL Homolog 1. 92, 93
- MLPH** Melanophilin. 59
- mm** millimeter. 9
- mRNA** micro Ribonucleic acid. *Glossary:* mRNA
- MSH2** MutS Homolog 2. 92, 93
- MYBL2** MYB Proto-Oncogene Like 2. 58, 71
- NAT1** N-Acetyltransferase 1. 58, 59
- NCD80** NDC80, Kinetochore Complex Component. 58
- NCI** National Cancer Institute. 11
- NCOR1** Nuclear Receptor Corepressor 1. 3, 62, 64, 72, 73, 75, 92, 93
- NCOR2** Nuclear Receptor Corepressor 1. 92, 93
- NGS** next generation sequencing. 15, 17, *Glossary:* NGS
- NHGRI** National Human Genome Research Institute. 11
- NIH** National Institute of Health. 86
- PAM50** Prediction Analysis of Microarray. 7, *Glossary:* backward elimination
- PCA** principal component analysis. 30, *Glossary:* Principal component analysis
- PDA** Pushdown Automaton. 36, *Glossary:* pda
- PgR** Progesterone Receptor. 7, 58, 59, 70, 75
- PIK3CA** Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha. 8, 16, 53, 55, 64, 71, 90, 91

<i>Abbreviations</i>	99
----------------------	----

- PIK3CB** Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Beta. 62, 64, 91
- PIK3R1** Phosphoinositide-3-Kinase Regulatory Subunit 1. 91, 93
- PNMT** Phenylethanolamine N-Methyltransferase. 58, 71
- PSA** prostate-specific antigen. 9
- PTEN** Phosphatase And Tensin Homolog. 64, 91, 92
- RAD51C** RAD51 Paralog C. 93
- RAF1** Raf-1 Proto-Oncogene, Serine/Threonine Kinase. 91, 93
- RB1** RB Transcriptional Corepressor 1. 92
- REST** Representational State Transfer. 20, 21, *Glossary:* REST
- RNA** Ribonucleic acid. 17, 25, *Glossary:* RNA
- RPKM** Reads Per Kilobase per Million. 25, *Glossary:* RPKM
- RSEM** RNA-Seq by Expectation Maximization. 25, *Glossary:* RSEM
- SAM** Sequence Alignment Map. 21, *Glossary:* SAM
- SD** standard deviation. 5, 44, 45, 47, 48, 51, 57, 62, 71, *Glossary:* standard deviation
- SETD2** SET Domain Containing 2. 91
- SFRP1** Secreted Frizzled Related Protein 1. 58
- SNP** single nucleotide polymorphism. 10, 15–17, *Glossary:* SNP
- SPOP** Speckle Type BTB/POZ Protein. 64, 91
- SPOPL** Speckle Type BTB/POZ Protein Like. 93
- SQL** Structured Query Language. 85, *Glossary:* SQL
- SVM** Support Vector Machine. 30, 33, 34, 43, 56, 62, 70, 71, *Glossary:* SVM
- TCGA** The Cancer Genome Atlas. 3, 11, 18, 21, 51, 55, 60, 61, 67, 68, *Glossary:* TCGA

***TMEM45B*** Transmembrane Protein 45B. 58

***TMPRSS2*** Transmembrane Protease, Serine 2. 10, 17, 64, 72

**TN** true negatives. 40, *Glossary*: true negatives

**TP** true positives. 37, 40, 59, 65, *Glossary*: true positives

***TP53*** Tumor Protein P53. 3, 51, 52, 62, 64, 72, 73, 75, 91, 92

***TSHZ3*** Teashirt Zinc Finger Homeobox 3. 51, 52

***TTF1*** Transcription Termination Factor 1. 51, 54

***URL*** Uniform Resource Locator. 21, 22, *Glossary*: URL

**US** United States. 7, 9

***VCF*** Variant Call Format. 21, *Glossary*: VCF

***VHL*** Von Hippel-Lindau Tumor Suppressor. 51, 52

***ZBTB16*** Zinc Finger And BTB Domain Containing 16. 93

***ZFHX3*** Zinc Finger Homeobox 3. 64, 92, 93

***ZMYM3*** Zinc Finger MYM-Type Containing 3. 92

# Glossary

**alignment** An alignment (or sequence alignment) is the process of arranging two sequences to each other. Often short reads are aligned to a reference genome. 15

**API** An API is a set of definitions, allowing the automated interaction between two systems. 18

**backward elimination** A set of 50 genes, by which expression status a cancer subtypes can be determined. 7

**boosting** A meta algorithm to build a strong classifier from several weak ones. 35

**Bootstrap Aggregation** A technique used by machine learning models to increase stability and accuracy.  $n'$  samples are drawn from  $n$  with replacement, a model is build using  $n'$  samples. This procedure is repeated  $k$  times.. 33

**bp** A base pair consists out of two nucleobases, bound to each other. They form the building blocks of DNA double helix. 16

**branch** A container within the git version control system. Often one master branch serving the software product and several branches for development exist. 23

**C++** A programming language, allowing more efficient computations than R. 13

**cardinality** The number of unique elements in a set. The cardinality of  $S = a, a, b, c$  would be 3, as there are 3 unique elements. 26, 31, 36, 39, 44

**cBioPortal** A portal to provide visualization, analysis and download of large-scale cancer genomics data sets. v, 18–20, 61, 68

**CNV** The number of gene copies in a genome, differing from two. 7

**confidence intervals** An interval yielding the precision of an estimated parameter. 37

**confusion matrix** A matrix indication the number of false and correct predicted samples, compared to their true label. 40

**cron-job** A small script which is run by a pre-defined interval. 23

**CSV** A file type, where information is stored in comma separated columns. 21

**DBMS** A system to manage multiple databases of the same type. 21

**DNA** A molecule that carries instructions for reproduction, growth and development all living organisms. 10

**EDL** An EDL builds a statistical classifier by the means of evolutionar computing. 3

**ensemble method** A machine learning method, which utilizes multiple models to perform a classification or regression task. 32

**epithelial-to-mesenchymal transition** A process in which cells lose their cell polarity and cell-cell adhesion, allowing them to become mesenchymal stem cells. 58

**ETS transcription factor family** Also known as E26 transformation-specific is a family genes, known to be associated with leukemia and several cancer disease. 64

**expected accuracy** The accuracy any random classifier is expected to achieve. 41

**false negatives** The number of positive samples being predicted negative. 40

**false positives** The number of negative samples being predicted positive. 40

**Firebrowse** A tool on top of the Firehose Pipeline to provide online access to analytical results over an API. 3, 18, 20–23, 68, 85

**FirebrowseR** A client to the Firebrowse software, enabling automatic data integration into the R environment. 3, 20, 23–25, 43, 51, 53–55, 61, 67, 75

**Firehose Pipeline** A data portal to systematize the analyses of TCGA data sets. 18–21, 23, 51, 53, 55, 56, 67, 68, 84

**gene fusion** The event that a new gene is formed from two previously separated genes. 17

**genotype** Set of amino acids to be translated into an organism phenotype. 15, 17, 85

**gini index** A measure for the impurity of a set. 31, 32

**grid search** A method used for parameter optimization. For two parameters, the two vectors  $x$  and  $y$ , containing the parameters to test, span a grid. This results in  $x * y$  models which need to be evaluated. 43

**HTTP** A protocol used for machine-machine communication, commonly used in the world wide web. 21

**inter-rater reliability** The degree of agreement among two raters. 14

**JSON** A compact and human readable file format, created for data exchange between applications. 21

**label** The variable to predict in a classification task. Also referred to as  $y$  in formulas. 14, 15

**MAP** A Bayesian parameters estimator, based on the a-priori distribution. 31

**NGS** A modern and efficient way to determine the order of nucleotides, compared to Sanger sequencing. 15

**noSQL** A database which is not structured as an Structured Query Language (SQL) database, mostly key/value stores. 17

**pda** A pushdown automaton is a deterministic automaton, with an additional stack. 36

**phenotype** Set of characteristics the genotype is expressed into. 17, 85

**predictor** A samples attribute, used in the context of classification problems. 14

**Principal component analysis** A PCA creates a set of uncorrelated variables from a set of correlated variables. 30

**R** A statistical computing environment, used for calculations and plotting. 13

**Rcpp** An extension to the R programming language, to execute C++ code within the R environment. 13

**reads** Short fragments (50-200 bp) of the amino acid sequence. 15

**REST** A programming paradigm for distributed systems, enabling the inter-machine communication over a set pre-defined commands. 20

**RNA** Similar to DNA, but single stranded. 17

**RPKM** A method of quantifying gene expression from RNA sequencing data by normalizing for total read length and the number of sequencing reads. 25

**RSEM** A software package for estimating gene and isoform expression levels from RNA sequencing data. 25

**rule support** The percentage of samples covered by a rule. 36, 44

**SAM** A file format to store reads mapped to a reference in. 21

**Sanger sequencing** The first known method to determine the nucleotides order in a DNA molecule. 85

**SNP** The variation of a single base pair in read, which occurs in >1% of all samples within a population. 10

**SQL** A standardized language to work with databases. 85

**standard deviation** A measure indicating the amount of variation within a set of values. 5

**SVM** A kernel based machine learning model. 30

**TCGA** A data portal run by the National Institute of Health (NIH), to provide public genomic data sets. 3, 19, 23, 85

**transcript** A transcript is a segment of a gene, which is translated by a single RNA. 17

**translation** The process of translating a genes amino acid sequence in a protein. 15

**true negatives** The number of false labels being predicted as negative. 40

**true positives** The number of true labels being predicted as true. 37

**unit tests** Requirements to programmed functions, which need to be fulfilled before the software product can be finalized. 23

**URL** An unique identifier locate an resource on the network. 21

**VCF** A file format used store mutation data. 21

**Web-TCGA** An online platform to visualize and analyze genomic data sets provided by the TCGA and Firehose Pipeline. 3, 21, 23, 25, 43, 51, 53–55, 75



# Appendix A

## Appendix

### A.1 Test Data Decision List

In this section decision lists for each test data set (described in 3.2 and 4.1), generated over all test samples, are given. Trailing numbers indicate the rules precision and their CI in braces.

#### A.1.1 The Tic Tac Toe Decision List

```
IF V2=x V1=x V3=x THEN win 1 (0.95 - 1)
ELSE IF V8=x V7=x V9=x THEN win 1 (0.95 - 1)
ELSE IF V9=x V5=x V1=x THEN win 1 (0.96 - 1)
ELSE IF V6=x V5=x V4=x THEN win 1 (0.95 - 1)
ELSE IF V4=x V1=x V7=x THEN win 1 (0.95 - 1)
ELSE IF V8=x V5=x V2=x THEN win 1 (0.95 - 1)
ELSE IF V6=x V3=x V9=x THEN win 1 (0.95 - 1)
ELSE IF V7=x V5=x V3=x THEN win 1 (0.96 - 1)
ELSE loss
```

#### A.1.2 The Titanic Decision List

```
IF class = Third THEN survival=No 0.75 (0.71 - 0.78)
IF gender=Female & age=Adult THEN survival=Yes 0.92 (0.88 - 0.95)
IF age=Child THEN survival=Yes 1 (0.88 - 1)
ELSE No
```

### A.1.3 The Mushrooms Decision List

```

IF odor=none & bruises=no & stalk-surface-above-ring=smooth THEN edible 1 (1 - 1)
ELSE IF gill-size=narrow & gill-attachment=free & gill-spacing=close THEN poisonous 1 (1 - 1)
ELSE IF odor=foul & veil-color=white THEN poisonous 1 (1 - 1)
ELSE IF stalk-shape=enlargingenlarging & population=several THEN poisonous 1 (0.63 - 1)
ELSE IF gill-size=narrow & bruises=no THEN poisonous 1 (0.97 - 1)
ELSE IF gill-color=white & odor=no THEN edible 1 (0.99 - 1)
ELSE IF habitat=woods & bruises=nof veil-color=white THEN poisonous 1 (0.95 - 1)
ELSE IF stalk-surface-above-ring=fibrous & gill-attachment=free THEN edible 1 (0.9 - 1)
ELSE edible

```

### A.1.4 The Cars Database decision lists

```

IF maint=low & safety=low THEN unacc 1 (0.97 - 1)
ELSE IF persons=2 THEN unacc 1 (0.99 - 1)
ELSE IF lug-boot=med & persons=2 THEN unacc 0 (0 - 0)
ELSE IF maint=vhigh & buying=high THEN unacc 1 (0.95 - 1)
ELSE IF buying=low & maint=low good 0.48 (0.33 - 0.63)
ELSE IF safety=med & buying=low THEN acc 0.64 (0.52 - 0.75)
ELSE IF maint=low & buying=med good 0.48 (0.33 - 0.63)
ELSE IF maint=high & buying=vhigh THEN unacc 1 (0.95 - 1)
ELSE IF safety=low THEN unacc 1 (0.98 - 1)
ELSE IF doors=2 & lug-boot=small THEN unacc 0.71 (0.55 - 0.84)
ELSE IF maint=vhigh & persons=2 THEN unacc 0 (0 - 0)
ELSE IF maint=vhigh & buying=vhigh THEN unacc 1 (0.92 - 1)
ELSE IF safety=high & maint=vhigh THEN acc 1 (0.92 - 1)
ELSE IF safety=high & lug-boot=small THEN acc 0.89 (0.77 - 0.96)
ELSE IF safety=high & buying=high THEN acc 1 (0.93 - 1)
ELSE IF buying=med & maint=med THEN acc 0.66 (0.49 - 0.8)
ELSE IF buying=low vgood 0.81 (0.64 - 0.93)
ELSE IF lug-boot=small THEN unacc 1 (0.92 - 1)
ELSE IF buying=low & safety=high THEN unacc 0 (0 - 0)
ELSE IF doors=4 & safety=low THEN unacc 0 (0 - 0)
ELSE IF safety=high & persons=4 THEN acc 1 (0.86 - 1)
ELSE IF safety=med & lug-boot=big THEN acc 1 (0.94 - 1)
ELSE IF doors=2 & safety=med THEN unacc 1 (0.77 - 1)
ELSE IF buying=low & maint=high THEN unacc 0 (0 - 1)
ELSE IF doors=3 & persons=4 THEN unacc 1 (0.59 - 1)
ELSE THEN acc

```

## A.2 Common Altered Genes in Prostate Cancer

In the following the preselected features for the prostate cancer classification task are listed. Frequently occurring gene fusions, somatic mutations and copy number alterations for primary, as well as for mCRPC patients are taken into account. The lists are based on publications from Perner et al [97], Cooper et al [23], Pflueger et al [101], Tomlins et al [128], Taylor et al [125] and Wang et al [133].

- *ERG*
- *ETV1*
- ETS Variant 4 (*ETV4*)
- Fli-1 Proto-Oncogene, ETS Transcription Factor (*FLI1*)
- *PIK3CA*

- *PIK3CB*
- Phosphoinositide-3-Kinase Regulatory Subunit 1 (*PIK3R1*)
- *SPOP*
- *FOXA1*
- Mediator Complex Subunit 12 (*MED12*)
- Isocitrate Dehydrogenase (NADP(+)) 1, Cytosolic (*IDH1*)
- Lysine Methyltransferase 2A (*KMT2A*)
- Lysine Methyltransferase 2A (*KMT2C*)
- Lysine Methyltransferase 2D (*KMT2D*)
- Lysine Demethylase 6A (*KDM6A*)
- SET Domain Containing 2 (*SETD2*)
- Chromodomain Helicase DNA Binding Protein 1 (*CHD1*)
- *TP53*
- *PTEN*
- *PIK3CA*
- *PIK3CB*
- *PIK3R1*
- B-Raf Proto-Oncogene, Serine/Threonine Kinase (*BRAF*)
- HRas Proto-Oncogene, GTPase (*HRAS*)
- Catenin Beta 1 (*CTNNB1*)
- AKT Serine/Threonine Kinase 1 (*AKT1*)
- BRCA1, DNA Repair Associated (*BRCA1*)
- BRCA2, DNA Repair Associated (*BRCA2*)
- Cyclin Dependent Kinase Inhibitor 1B (*CDKN1B*)
- (*RAF1*)

- APC, WNT Signaling Pathway Regulator (*APC*)
- RB Transcriptional Corepressor 1 (*RB1*)
- Zinc Finger MYM-Type Containing 3 (*ZMYM3*)
- ATM Serine/Threonine Kinase (*ATM*)
- Cyclin Dependent Kinase 12 (*CDK12*)
- Fanconi Anemia Complementation Group C (*FANCC*)
- Fanconi Anemia Complementation Group D2 (*FANCD2*)
- *AR*
- *NCOR1*
- Nuclear Receptor Corepressor 1 (*NCOR2*)
- MutL Homolog 1 (*MLH1*)
- MutS Homolog 2 (*MSH2*)
- Cyclin Dependent Kinase Inhibitor 2A (*CDKN2A*)
- *KMT2C*
- *ZFHX3*
- *PTEN*
- *TP53*
- *CHD1*
- *BRCA1*
- *BRCA2*
- *CDKN1B*
- *RB1*
- *CDK12*
- *FANCD2*
- Family With Sequence Similarity 175 Member A (*FAM175A*)

- *FANCC*
- RAD51 Paralog C (*RAD51C*)
- Speckle Type BTB/POZ Protein Like (*SPOPL*)
- (*ZBTB16*)
- *NCOR1*
- *NCOR2*
- *PIK3R1*
- *BRAF*
- *RAF1*
- *APC*
- *ATM*
- *MLH1*
- *MSH2*
- *CDKN2A*
- Cyclin D1 (*CCND1*)
- *ZFHX3*
- GNAS Complex Locus (*GNAS*)
- ETS2 Repressor Factor (*ERF*)



January 25, 2018

## Personal data

Date of birth	$xx^{th}$ of xxxx 19XX, Herne - GER
Family status	engaged
Nationality	german

## Experience

11/2015–present	<b>Research Fellow</b> , <i>University Hospital Schleswig-Holstein</i> , Lübeck, GER, Department of Pathology.
11/2010–11/2015	<b>Research Fellow</b> , <i>University Hospital Bonn</i> , Bonn, GER, Section of Prostate Cancer Research.
05/2010–09/2010	<b>Software Developer, Bachelors Thesis</b> , <i>Pixelpark AG</i> , Cologne, GER. Evaluating, conceiving and enhancing data structures for mobile Augmented Reality Browsers.
10/2009–05/2010	<b>Software Developer</b> , <i>Rheni GmbH</i> , Sankt Augustin, GER. Designing, planning and developing of web bases applications using Groovy On Grails
07/2009–10/2009	<b>Internship</b> , <i>Pixelpark AG</i> , Cologne, GER. Conception, design and planning of mobile applications based on the Android platform

## Education

05/2014–Present	<b>PhD Student in Computational Biology</b> , <i>University of Bonn</i> , Faculty of Mathematics and Natural Sciences, Bonn, GER. Title of thesis: Predicting Rules for Cancer Subtype Classification using Grammar-Based Genetic Programming on various Genomic Data Types
09/2010–06/2013	<b>Master of Science in Computer Science</b> , <i>Bonn-Rhine-Sieg University</i> , University of Applied Sciences, Sankt-Augustion, GER. Masters Thesis "Identification of functional interactions between distant metastasis and primary tumor on whole exome next generation sequencing data", in german
07/2008–09/2008	<b>Visiting Student</b> , <i>York University</i> , Faculty of Science & Engineering, Toronto, CA. Theory of Computation and Programming Language Fundamentals
09/2007–09/2010	<b>Bachelor of Science in Computer Science</b> , <i>Bonn-Rhine-Sieg University</i> , University of Applied Sciences, Sankt-Augustion, GER. Bachelor Thesis "Development of a communication interface for augmented reality servers and mobile augmented reality browsers", in german
08/2004–06/2007	<b>Adolph-Kolping-Berufskolleg</b> , <i>Vocational School</i> , Brakel, GER. Vocational- and Technical Diploma

## Publications

- Deng, M., J. Bragelmann, I. Kryukov, N. Saraiva-Agostinho, and S. Perner, "FirebrowseR: an R client to the Broad Institute's Firehose Pipeline," *Database (Oxford)*, vol. 2017, 2017. [PubMed Central:PMC5216271] [DOI:10.1093/database/baw160] [PubMed:28062517].
- M. Nientiedt, Deng, M., D. Schmidt, S. Perner, S. C. Muller, and J. Ellinger, "Identification of aberrant

tRNA-halves expression patterns in clear cell renal cell carcinoma," *Sci Rep*, vol. 6, p. 37158, Nov 2016. [PubMed Central:PMC5121638] [DOI:10.1038/srep37158] [PubMed:27883021].

J. Bragelmann, N. Klumper, A. Offermann, A. von Massenhausen, D. Bohm, **Deng, M.**, A. Queisser, C. Sanders, I. Syring, A. S. Merseburger, W. Vogel, E. Sievers, I. Vlasic, J. Carlsson, O. Andren, P. Brossart, S. Duensing, M. A. Svensson, Z. Shaikhbrahim, J. Kirfel, and S. Perner, "Pan-Cancer Analysis of the Mediator Complex Transcriptome Identifies CDK19 and CDK8 as Therapeutic Targets in Advanced Prostate Cancer," *Clin. Cancer Res.*, Sep 2016. [DOI:10.1158/1078-0432.CCR-16-0094] [PubMed:27678455].

A. Queisser, S. Hagedorn, H. Wang, T. Schaefer, M. Konantz, S. Alavi, **Deng, M.**, W. Vogel, A. von Massenhausen, G. Kristiansen, S. Duensing, J. Kirfel, C. Lengerke, and S. Perner, "Ecotropic viral integration site 1, a novel oncogene in prostate cancer," *Oncogene*, Sep 2016. [DOI:10.1038/onc.2016.325] [PubMed:27617580].

A. von Massenhausen, C. Sanders, B. Thewes, **Deng, M.**, A. Queisser, W. Vogel, G. Kristiansen, S. Duensing, A. Schrock, F. Bootz, P. Brossart, J. Kirfel, L. Heasley, J. Bragelmann, and S. Perner, "MERTK as a novel therapeutic target in head and neck cancer," *Oncotarget*, vol. 7, pp. 32678–32694, May 2016. [PubMed Central:PMC5078043] [DOI:10.18632/oncotarget.8724] [PubMed:27081701].

A. von Massenhausen, **Deng, M.**, H. Billig, A. Queisser, W. Vogel, G. Kristiansen, A. Schrock, F. Bootz, F. Goke, A. Franzen, L. Heasley, J. Kirfel, J. Bragelmann, and S. Perner, "Evaluation of FGFR3 as a Therapeutic Target in Head and Neck Squamous Cell Carcinoma," *Target Oncol*, vol. 11, pp. 631–642, Oct 2016. [DOI:10.1007/s11523-016-0431-z] [PubMed:27053219].

I. Syring, N. Klumper, A. Offermann, M. Braun, **Deng, M.**, D. Boehm, A. Queisser, A. von Massenhausen, J. Bragelmann, W. Vogel, D. Schmidt, M. Majores, A. Schindler, G. Kristiansen, S. C. Muller, J. Ellinger, Z. Shaikhbrahim, and S. Perner, "Comprehensive analysis of the transcriptional profile of the Mediator complex across human cancer types," *Oncotarget*, vol. 7, pp. 23043–23055, Apr 2016. [PubMed Central:PMC5029609] [DOI:10.18632/oncotarget.8469] [PubMed:27050271].

**Deng, M.**, J. Bragelmann, J. L. Schultze, and S. Perner, "Web-TCGA: an online platform for integrated analysis of molecular cancer data sets," *BMC Bioinformatics*, vol. 17, p. 72, Feb 2016. [PubMed Central:PMC4744375] [DOI:10.1186/s12859-016-0917-9] [PubMed:26852330].

S. Schrodter, M. Braun, I. Syring, N. Klumper, **Deng, M.**, D. Schmidt, S. Perner, S. C. Muller, and J. Ellinger, "Identification of the dopamine transporter SLC6A3 as a biomarker for patients with renal cell carcinoma," *Mol. Cancer*, vol. 15, p. 10, Feb 2016. [PubMed Central:PMC4736613] [DOI:10.1186/s12943-016-0495-5] [PubMed:26831905].

J. Ellinger, J. Alam, J. Rothenburg, **Deng, M.**, D. Schmidt, I. Syring, H. Miersch, S. Perner, and S. C. Muller, "The long non-coding RNA Inc-ZNF180-2 is a prognostic biomarker in patients with clear cell renal cell carcinoma," *Am J Cancer Res*, vol. 5, no. 9, pp. 2799–2807, 2015. [PubMed Central:PMC4633906] [PubMed:26609485].

**Deng, M.**, J. J. Blondeau, D. Schmidt, S. Perner, S. C. Muller, and J. Ellinger, "Identification of novel differentially expressed lncRNA and mRNA transcripts in clear cell renal cell carcinoma by expression profiling," *Genom Data*, vol. 5, pp. 173–175, Sep 2015. [PubMed Central:PMC4584005] [DOI:10.1016/j.gdata.2015.06.016] [PubMed:26484251].

J. J. Blondeau, **Deng, M.**, I. Syring, S. Schrodter, D. Schmidt, S. Perner, S. C. Muller, and J. Ellinger, "Identification of novel long non-coding RNAs in clear cell renal cell carcinoma," *Clin Epigenetics*, vol. 7, p. 10, 2015. [PubMed Central:PMC4326488] [DOI:10.1186/s13148-015-0047-7] [PubMed:25685243].

C. A. Brownstein, A. H. Beggs, N. Homer, B. Merriman, T. W. Yu, K. C. Flannery, E. T. DeChene, M. C. Towne, S. K. Savage, E. N. Price, I. A. Holm, L. J. Luquette, E. Lyon, J. Majzoub, P. Neupert, D. McCallie, P. Szolovits, H. F. Willard, N. J. Mendelsohn, R. Temme, R. S. Finkel, S. W. Yum, L. Medne, S. R. Sunyaev, I. Adzhubei, C. A. Cassa, P. I. de Bakker, H. Duzkale, P. Dworzycski, W. Fairbrother, L. Francioli, B. H. Funke, M. A. Giovanni, R. E. Handsaker, K. Lage, M. S. Lebo, M. Lek, I. Leshchiner, D. G. MacArthur, H. M. McLaughlin, M. F. Murray, T. H. Pers, P. P. Polak, S. Raychaudhuri, H. L. Rehm, R. Soemedi, N. O. Stitzel, S. Vestecka, J. Supper, C. Gugenmus, B. Klocke, A. Hahn, M. Schubach, M. Menzel, S. Biskup, P. Freisinger, **Deng, M.**, M. Braun, S. Perner, R. J. Smith, J. L. Andorf, J. Huang, K. Ryckman, V. C. Sheffield, E. M. Stone, T. Bair, E. A. Black-Ziegelbein, T. A. Braun, B. Darbro, A. P. DeLuca, D. L. Kolbe, T. E. Scheetz, A. E. Shearer, R. Sompallae, K. Wang, A. G. Bassuk, E. Edens, K. Mathews, S. A. Moore, O. A. Shchelochkov, P. Trapane, A. Bossler, C. A. Campbell, J. W. Heusel, A. Kwitek, T. Maga, K. Panzer, T. Wassink, D. Van Daele, H. Azaiez, K. Booth, N. Meyer, M. M. Segal, M. S. Williams, G. Tromp, P. White, D. Corsmeier, S. Fitzgerald-Butt, G. Herman, D. Lamb-Thrush, K. L. McBride, D. Newsom, C. R. Pierson, A. T. Rakowsky, A. Maver, L. Lovre?i?, A. Palanda?i?,

B. Peterlin, A. Torkamani, A. Wedell, M. Huss, A. Alexeyenko, J. M. Lindvall, M. Magnusson, D. Nilsson, H. Stranneheim, F. Taylan, C. Gilissen, A. Hoischen, B. van Bon, H. Yntema, M. Nelen, W. Zhang, J. Sager, L. Zhang, K. Blair, D. Kural, M. Cariaso, G. G. Lennon, A. Javed, S. Agrawal, P. C. Ng, K. S. Sandhu, S. Krishna, V. Veeramachaneni, O. Isakov, E. Halperin, E. Friedman, N. Shomron, G. Glusman, J. C. Roach, J. Caballero, H. C. Cox, D. Mauldin, S. A. Ament, L. Rowen, D. R. Richards, F. A. San Lucas, M. L. Gonzalez-Garay, C. T. Caskey, Y. Bai, Y. Huang, F. Fang, Y. Zhang, Z. Wang, J. Barrera, J. M. Garcia-Lobo, D. Gonzalez-Lamuno, J. Llorca, M. C. Rodriguez, I. Varela, M. G. Reese, F. M. De La Vega, E. Kiruluta, M. Cargill, R. K. Hart, J. M. Sorenson, G. J. Lyon, D. A. Stevenson, B. E. Bray, B. M. Moore, K. Eilbeck, M. Yandell, H. Zhao, L. Hou, X. Chen, X. Yan, M. Chen, C. Li, C. Yang, M. Gunel, P. Li, Y. Kong, A. C. Alexander, Z. I. Albertyn, K. M. Boycott, D. E. Bulman, P. M. Gordon, A. M. Innes, B. M. Knoppers, J. Majewski, C. R. Marshall, J. S. Parboosingh, S. L. Sawyer, M. E. Samuels, J. Schwartzenzuber, I. S. Kohane, and D. M. Margulies, "An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge," *Genome Biol.*, vol. 15, p. R53, Mar 2014. [PubMed Central:PMC4073084] [DOI:10.1186/gb-2014-15-3-r53] [PubMed:24667040].

R. Menon, **Deng, M.**, K. Ruenuver, A. Queisser, M. Peifer, M. Pfeifer, A. Offermann, D. Boehm, W. Vogel, V. Scheble, F. Fend, G. Kristiansen, N. Wernert, N. Oberbeckmann, S. Biskup, M. A. Rubin, Z. Shaikhbrahim, and S. Perner, "Somatic copy number alterations by whole-exome sequencing implicates YWHAZ and PTK2 in castration-resistant prostate cancer," *J. Pathol.*, vol. 231, pp. 505–516, Dec 2013. [DOI:10.1002/path.4274] [PubMed:24114522].

R. Menon, **Deng, M.**, D. Boehm, M. Braun, F. Fend, D. Boehm, S. Biskup, and S. Perner, "Exome enrichment and SOLiD sequencing of formalin fixed paraffin embedded (FFPE) prostate cancer tissue," *Int J Mol Sci*, vol. 13, no. 7, pp. 8933–8942, 2012. [PubMed Central:PMC3430274] [DOI:10.3390/ijms13078933] [PubMed:22942743].