Universität Leipzig

Text

Description automatically generated

Masterarbeit

Bioinformatik

**Prediction of Immunotherapy Response in Melanoma Patients based on Machine Learning**

Christina Kuhn

Leipzig, April 2021

**Prediction of Immunotherapy Response in Melanoma Patients based on Machine Learning**

**Masterarbeit im Studiengang Master-Bioinformatik**

**des Fachbereichs Informatik und Mathematik**

**an der Universität Leipzig**

**Erstgutachter:**

**Prof.Torsten Schöneberg**

**Zweitgutachter:**

**Prof. Peter Stadler**

**Betreuer:**

**Dr.Susanne** **Horn**

Abstract

In recent years, immunotherapy with immune checkpoint blockade (ICB) has shown enormous success in the treatment of melanoma. However, reliably predicting a successful therapy while avoiding therapy options without benefit at baseline is still an unsolved issue. The aim of this master thesis is therefore to define statistical models that predict the success of immune checkpoint therapies in melanoma patient cohorts using neural networks. Since resistance to ICB is related to tumor environment and host immune factors, personalized models based on a patient's genomic setup could be decisive. However, complexity and high dimensionality resulting from the transcriptome data analysed here needs to be addressed with an automated machine learning algorithm. Models were based on Artificial Neural Networks (ANN) to predict the overall and progression free survival (OS and PFS, respectively) of melanoma patients undergoing anti-CTLA4 and anti-PD1/anti-PDL1 therapy. Measures of gene expression in Transcripts per Million (TPM) from bulk tumor RNA-sequencing data were used from five melanoma datasets. Clinical variables were included such as gender, age, and the type of therapy. The ANN was then optimised to achieve the highest possible accuracy in predicting the predefined survival outcome. Problems resulting from high-dimensional data, such as overfitting, were addressed using regularization and feature selection. As a result, the ANN-based model with feature selection was shown to have the ability to predict survival (PFS) to ICB therapy with up to 86% accuracy. ANN without feature selection, however with regularization, reached up to 72% accuracy for PFS and 71% for OS, respectively. To address the problem of small patient numbers and to test reproducibility, the model was trained and validated based on the combination of all five datasets. Since the combination did not lead to an improvement in prediction, follow-up studies are necessary, whereby the developed workflow can be used as a starting point for adaption to new datasets. In summary, the developed model may contribute to personalized therapy decisions in melanoma patients.

Zusammenfassung

In den letzten Jahren hat die Immuntherapie mit Immun-Checkpoint-Blockaden (ICB) bei der Behandlung von Melanom Patienten enorme Erfolge erzielt. Die zuverlässige Vorhersage eines Therapieerfolgs zu Beginn der Behandlung, bei gleichzeitiger Vermeidung von Therapieoptionen mit geringerer Wirksamkeit, ist jedoch noch eine ungelöste Frage. Ziel dieser Masterarbeit ist es daher, statistische Modelle zu definieren, die den Erfolg von Immun-Checkpoint-Therapien in Melanom-Patientenkohorten vorhersagen. Das Ansprechen ist nicht nur auf klinische Faktoren zurückzuführen, sondern auch auf die Tumorumgebung und die angeboreren immunologischen Faktoren des Wirtes. Personalisierte Modelle, die auf dem genomischen Profil eines Patienten basieren, können den Erfolg von Krebstherapien schon heute entscheidend verbessern. Die Komplexität und hohe Dimensionalität, die sich aus den hier analysierten Transkriptomdaten ergibt, wird in dieser Masterarbeit mit einem automatisierten maschinellen Lernalgorithmus addressiert. Hierfür wurde ein Model auf Basis eines Künstlichen Neuronalen Netzes (ANN) entwickelt, dass das Gesamtüberleben oder progressionsfreie Überleben (OS und PFS, respektive) von Melanompatienten vorhersagt, die eine anti-CTLA4 oder anti-PD1/anti-PDL1-Therapie erhielten. Genutzt wurden TPM Werte (Transcripts per Million) der RNA-Sequenzierungsdaten aus fünf Melanom-Datensätzen. Zusammen mit klinische Daten, wie zum Beispiel Alter, Geschlecht und Therapieform wurde das ANN so trainiert, dass eine möglichst hohe Genauigkeit bei der Vorhersage des vordefinierten Überlebensergebnisses erreicht wird. Probleme wie Überanpassung, resultierend aus den hochdimensionale Daten, wurden mit Hilfe von Regularisierung und Feature Selection adressiert. Abschließend wurde gezeigt, dass das ANN-basierte Modell mit Feature Selection in der Lage ist, das Ansprechen (PFS) auf die ICB-Therapie mit einer Genauigkeit von bis zu 86% vorherzusagen. ANN ohne Feature Selection, jedoch mit Regularisierung, erreichte Genauigkeiten von bis zu 71% für PFS bzw. 72% für OS. Um das Problem der geringen Patientenzahlen zu adressieren und die Reproduzierbarkeit zu testen, wurden die Vorhersagen auf die Kombination aller fünf Datensätze erweitert. Da die Kombination nicht zu einer Verbesserung der Vorhersage geführt hat, sind Folgeuntersuchungen notwendig, wobei der entwickelte Workflow aufgrund seiner Anpassungsfähigkeit als Ausgangspunkt direkt genutzt werden kann. Somit wurde im Zuge dieser Masterarbeit ein Model entwickelt, das im Besten Fall einen Beitrag zu personalisierten Therapieentscheidungen bei Melanompatienten leisten kann.

**Table of Contents**

[List of Figures I](#_Toc69294437)

[List of Tables II](#_Toc69294438)

[Abbreviations III](#_Toc69294439)

[1.0 Introduction 1](#_Toc69294440)

[1.1 ICB therapy of melanoma patients 1](#_Toc69294441)

[1.2 Neural Network based prediction of ICB therapy response 3](#_Toc69294442)

[1.3 Problems 4](#_Toc69294443)

[1.3.1 High Dimensional Data 4](#_Toc69294444)

[1.3.2 Overfitting 5](#_Toc69294445)

[1.3.3 Small Size Dataset 6](#_Toc69294446)

[1.3.4 Class Imbalance 6](#_Toc69294447)

[1.3.5 Study Reproducibility 6](#_Toc69294448)

[1.4 Objectives 8](#_Toc69294449)

[2. Methods 1](#_Toc69294450)

[2.1 Datasets 1](#_Toc69294451)

[2.1.1 Boston Housing Dataset 1](#_Toc69294452)

[2.1.2 Melanoma ICB Cohorts 1](#_Toc69294453)

[2.2 Programs 4](#_Toc69294454)

[2.2.1 Jupyter Notebook and Python Scripts 4](#_Toc69294455)

[2.2.3 Packages 4](#_Toc69294456)

[2.3 Workflow 6](#_Toc69294457)

[2.3.1 Pre-Processing of Data 6](#_Toc69294458)

[2.3.2 Feature Selection Technique 8](#_Toc69294459)

[2.3.3 Implementation of Neural Network 8](#_Toc69294460)

[2.3.4 Fine-Tuning 9](#_Toc69294461)

[2.3.5. Model Validation 10](#_Toc69294462)

[2.3.6 Applicability 10](#_Toc69294463)

[3. Results 11](#_Toc69294464)

[3.1 Sanity Check with Boston Housing Dataset 11](#_Toc69294465)

[3.2 Predictions on Liu Dataset 14](#_Toc69294466)

[3.2.1 Without Feature Selection 14](#_Toc69294467)

[3.2.2 With Feature Selection 16](#_Toc69294468)

[3.3. Predictions on Combination of Cohorts 18](#_Toc69294469)

[4. Discussion 21](#_Toc69294470)

[5. Outlook 27](#_Toc69294471)

[Bibliography i](#_Toc69294472)

[Appendix viii](#_Toc69294473)

[Selbstständigkeitserklärung xviii](#_Toc69294474)

# List of Figures

[Figure 1: Mathematical representation of ANN by Burke et al., 1997 3](#_Toc69299428)

[Figure 2: Overall workflow 8](#_Toc69299429)

[Figure 3: ANN workflow 5](file:///C:\Users\Christina\Desktop\Masterarbeit\AIforMT\Masterarbeit_12_04_21.docx#_Toc69299430)

[Figure 4: Neural Network Architecture 8](#_Toc69299431)

[Figure 5: Leader board Boston Housing Kaggle competition 2016 13](#_Toc69299432)

[Figure 6: Survival analysis of identified genes potential predictive for ICB efficacy 24](#_Toc69299433)

# List of Tables

[Table 1: Datasets of melanoma patients undergoing ICB therapy 2](#_Toc69299434)

[Table 2: Used Python packages with version and purpose 4](#_Toc69299435)

[Table 3: Pre-processing of datasets 7](#_Toc69299436)

[Table 4 Binary and continuous prediction of the Boston Housing dataset 12](#_Toc69299437)

[Table 5: Binary prediction on Liu et al. dataset with OS as endpoint 15](#_Toc69299438)

[Table 6: Occurrences of selected features by the RFE algorithm in the different train/test splits 17](#_Toc69299439)

[Table 7: Binary prediction on Liu et al. dataset after feature selection with PFS as endpoint 17](#_Toc69299440)

[Table 8: Binary prediction on the combination of cohorts with OS as endpoint 19](#_Toc69299441)

# Abbreviations

ANN Artificial Neural Network

CD8A Cluster of Differentiation 8a

CTLA4 Cytotoxic T Lymphocyte Antigen 4

DNA Deoxyribonucleic Acid

HLA-A Human Leukocyte Antigen

ICB Immune Checkpoint Blockade

L1 Lasso (Least Absolute Shrinkage and Selection Operator)

L2 Ridge

LDH Lactate Dehydrogenase

MAE Mean Absolute Error

MAPKi Mitogen-Activated Protein Kinase Inhibition

MEDV Median Value of Owner-Occupied Homes

(m)RNA (messenger-)Ribonucleic Acid

OS Overall Survival

PD1/ PDL1 Programmed Cell Death 1/ Programmed Cell Death Ligand 1

PFS Progression Free Survival

RECIST Response Evaluation Criteria in Solid Tumors

RFE Recursive Feature Elimination

RSME Root Mean Squared Error

SMOTE(-NC) Synthetic Minority Class Over Sampling (Numerical Categorical)

TPM Transcripts Per Million

# 1.0 Introduction

## 1.1 ICB therapy of melanoma patients

Recent studies have shown great progress in the therapy of melanoma with immune checkpoint blockade (ICB). In a cancer cell, a so-called checkpoint ensures that the immune response is not too strong and thus prevents the T cells from killing the cancer cell. Blocking these checkpoints with monoclonal antibodies therefore helps to initiate the immune response against the cancer cell. The use of antibodies targeting the checkpoint molecules cytotoxic T lymphocyte antigen 4 (CTLA4), programmed cell death 1 (PD1) and PD1 ligand 1 (PDL1) had early success in the clinic (Hodi et al., 2010; Robert et al., 2015; Wolchok et al., 2017). Approved therapies now include anti-PD1 (nivolumab or pembrolizumab) (Hugo et al., 2016) and anti-CTLA4 (ipilimumab) (Hodi et al., 2010) or a combination of anti-PD1 and anti-CTLA4 (Larkin et al., 2015).

However, not all patients currently benefit from those treatments which has fuelled a wave of research on molecular mechanisms of resistance to ICB (Liu, Schilling, et al., 2019; Spranger et al., 2015). Besides tumor-intrinsic and tumor-extrinsic biomarkers for response or resistance to ICB, patient genomic makeup is also considered for treatment decisions (Liu, Jenkins, et al., 2019). RNA-based studies have identified gene expression signatures that are linked to immune infiltration in various tumors within the tumor microenvironment which furthermore correlated with overall survival of the patients (Brown et al., 2014; Rooney et al., 2015). Those correlations show HLA-A and CD8A expression in correlation with overall survival which could be potential candidates for CTLA4-targeted antibodies (Brown et al., 2014). Rooney et al. (2015) suggest that neoantigens and viruses reveal known and novel mutations that enable tumors to resist immune attack. Whether those findings can serve as predictors of response remains unclear. Furthermore, it shows that prior failure of mitogen-activated protein kinase (MAPK)-targeted therapy is a negative factor for subsequent response to ICB in melanoma (Hugo et al., 2016).

More recent findings implicate that DNA- and RNA-level genomic information has predictive value for the clinical benefit of both anti-CTLA4 and anti-PD1 therapy (Hugo et al., 2016; Le et al., 2015; Rizvi et al., 2015; Rosenberg et al., 2016; Snyder et al., 2014; Van Allen et al., 2015). For example, since high mutational burden is associated with overall survival but not predictive of response to anti-PD1 therapy, Hugo et al. (2016) suggest that other genomic or non-genomic features also contribute to response. Detailed molecular characterization of large cohorts is needed to further identify a prognostic value for the daily routine patient care.

Finding predictors of responsiveness has remained a serious challenge, which is the main point for this study, because of the complexity of the immune response, the heterogenetic tumour environment, and the resulting high dimensional dataset. Those resulting big datasets need to be addressed with intelligent and automated methods, hence, associations between genomic and transcriptomic features and immune response have been the basis for the development of statistical models, such as multivariate predictive models by Liu, Schilling, et al. (2019). The multicollinearity of transcriptomic data calls for models that take nonlinear relationships into account. Furthermore, gene expression data underlies the problem of biological noise which is caused by the experimental setup. Deep neural networks recently show promising results in modeling these nonlinear relationships with respect to predicting cancer therapy response (Adam et al., 2020). Moreover, a study by Daoud & Mayo (2019) showed that neural network based approaches outperformed other machine learning tools in classifying gene expression samples in most of the studies.

Compared to the enormous growth of deep learning methods the use of deep learning in immunotherapy response prediction methods is almost absent (see appendix 1A). As shown in appendix 1B, the need for immunotherapy response prediction methods has increased greatly, whereas the application of deep learning has not increased proportionally. Especially applied to ICB therapy in melanoma patients where a neural network-based model has not yet been utilized, this could help decide whether a treatment is appropriate for an individual patient. Like Adam et al. (2020) stated, a deep neural network “with an effective inductive bias for genomics will allow the complex underlying cancer biology to be better modeled compared to linear models”.

## 1.2 Neural Network based prediction of ICB therapy response

An ANN consists of input nodes, hidden nodes, output nodes and weighted connections between the nodes. In the simplest so-called feed forward neural network all the nodes in the input layer are connected to the nodes in the adjacent layer. Each node of the input layer represents an input feature which can be gender of a patient or amount of expression of one gene. The input information is then combined by the neural network and patterns are detected to make predictions or to classify (Burke, 1994). Each node after the input layer sums up the information from the input nodes and uses an activation function to send the information to the adjacent layer. The neural network training algorithm used in this work is backpropagation. In backpropagation the error (difference between predicted outcome and true outcome) is propagated back from the output to the connection weights to adjust the weights in the direction of minimum error (Burke et al., 1997). A mathematical representation is presented below, where *hj* in (1) is the output of each of the hidden nodes *j*, *f* is a nonlinear transfer function, *wh* is the weight from predictor *i* to hidden node *j*, and *xi* is an input variable. The output *hj* of each hidden node is then used in the prediction of the network *oj*, where those hidden outputs are combined with the output weight *w0* and *g* represents a nonlinear activation function.

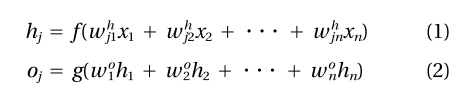


Figure 1: Mathematical representation of ANN by Burke et al., 1997

Based on next-generation sequencing data, the individual prediction of drug response is an untypical question and starting position for an ANN because of the high dimensional characteristics of the input features and the small number of samples. In recent studies the typical scientific approaches for drug response prediction consists of (1) quantification of drug response; (2) molecular feature selection or dimensionality reduction of the cellular measurements; (3) machine learning models fitting to predict drug response; and (4) model evaluation (Adam et al., 2020). In the case of this master thesis, the progression free survival (PFS) or the overall survival (OS) of a patient represents the quantification of drug response (1). PFS is defined as the length of time during and after the treatment that a patient lives with the disease without experiencing a predefined progressive event and OS as the length of time since the date of diagnosis that a patient is still alive. The molecular feature selection or dimensionality reduction (2) is a crucial step for finding patterns and is further addressed in section 1.3.1. Besides the gene expression and the given ICB treatment, other genomic or clinical features of melanoma patients are considered in recent studies like tumor mutational burden, purity, ploidy and heterogeneity of the tumor or sex, age and lactate dehydrogenase (LDH) value of the patient (Liu, Schilling, et al., 2019). Since current therapeutic strategy for melanoma treatment can include combined therapies (Domingues et al., 2018), prior treatments such as MAPK inhibition (MAPKi) may be included in the feature space if the information is available. The fitting and evaluation of the machine learning model (3 and 4) can then provide a sophisticated model that could be used clinically.

## 1.3 Problems

The development of a deep learning model to predict survival after ICB therapy leads to several problems arising from the inhomogeneous and high-dimensional data. Those problems are addressed in detail to specify the problems resulting from the main research question.

### 1.3.1 High Dimensional Data

A typical “gene expression dataset has the number of genes ranging from thousands to tens of thousands, while the number of tissue samples is less than several hundred” (Ahn, 2006). A dataset where the ratio of feature to sample is such can be described as a high dimensional low sample size dataset (Ahn, 2006). The problems occurring from high dimensional data is that a model easily results in overfitting data which leads to a poor generalization on an unknown dataset (Witten & Tibshirani, 2010). The enormous number of genes of a transcriptomic dataset includes noisy genes, making it difficult for statistical models to detect patterns based on a limited number of observations.

A beforehand feature selection “addresses [this] problem of generalization which leads to better accuracy” (Perez-Riverol et al., 2017) by filtering out the noise in the given data*.* Methods for feature selection, which are widely used in bioinformatics, are classified into filter, wrapper and embedded (Ma & Huang, 2008; Saeys et al., 2007). A filter-based approach separates the feature selection from the classifier construction whereas a wrapper is a combination of feature selection with the learning or classification step. For example, the Recursive Feature Elimination (RFE) algorithm conducts a wrapper feature selection by recursively eliminating the worst predictable feature based on a classification model. Embedded feature selection means the use of algorithms with built-in feature selection methods. Besides a beforehand feature selection, regularization methods, which can efficiently select features and control overfitting, became more popular in recent studies (Algamal & Lee, 2015b; Y. Liang et al., 2013; Torang et al., 2019). Regularization is one example of embedded feature selection. Moreover, models with regularization penalties like Lasso (L1), Ridge (L2) or Elastic Net belong to multivariate feature selection, thus are appropriate for gene expression data, whereas Elastic Net is a linear combination of Lasso and Ridge (Algamal & Lee, 2015a). It has been shown that Elastic Net outperforms Lasso and Ridge in extremely high dimensional problems (Zou & Hastie, 2005). There is no canonical way to choose an appropriate feature selection technique but extensive comparative analyses of machine learning models such as the ones from Costello et al. (2014) and Jang et al. (2014) recommend using Elastic Net or Ridge regression.

### 1.3.2 Overfitting

As previously described, high dimensional data can result in overfitting, which occurs when the network fits the training examples so perfectly that it is not capable of generalizing on the test dataset (Daoud & Mayo, 2019). This is because the informative gene signatures are hidden within the high dimensional transcriptomic space. Overfitting is also caused by using small number of training examples as indicated by Srivastava et al. (2014). There exist many different approaches to overcome overfitting in deep neural networks like using dropout (Srivastava et al., 2014), early stopping (Olson et al., 2018) and more regularization techniques (Wu et al., 2018).

### 1.3.3 Small Size Dataset

Because the size of the largest cohort undergoing ICB therapy is 121 patients (see 2.1.2), the available dataset for machine learning is small. This results in less training data and thus can lead to overfitting. It is crucial to keep in mind that a “small sized training sample can result in misclassification […] while the estimators may produce unstable and biased models” (Kourou et al., 2015). Since current data availability is limited in their size, named approaches for overfitting need to be applied to obtain stable models.

### 1.3.4 Class Imbalance

A problem for machine learning which occurs in a classification problem is that one class may have more elements than the other. This issue applies to the analysed datasets in this work, as there are often more patients with short survival than patients with long survival, so predicting OS/PFS is an imbalanced classification problem. If an overrepresented class has more examples the machine learning model is biased towards that class and consequently less accurate in predicting the underrepresented class (Daoud & Mayo, 2019). Danaee et al. (2017) used the Synthetic Minority Class Over Sampling (SMOTE) method in their deep learning approach to cancer detection to deal with class imbalance at the high-dimensional level. SMOTE is an approach where the minority class is over-sampled through creating synthetic minority class examples (Chwala et al., 2002).

### 1.3.5 Study Reproducibility

Validation of current study results is limited by the availability of comprehensive data (Liu, Schilling, et al., 2019), so collection and integration of existing data are critical. As Ali & Aittokallio (2019) noted, this requires common standards for data integration and sharing as well as infrastructure developments. Because gene expression is not always consistent due to its generations on different microarray platforms or batch effects, data integration is not straightforward (Geeleher et al., 2014). In addition to gene expression, the studies also differ in their clinical and genomic variables as well as their prediction endpoint. In the absence of an international standard for response measurement for preclinical in vitro models, comparison between response prediction methods remains difficult (Adam et al., 2020). To improve reproducibility, publication of used data and documentation of the methodologies used is necessary (Daoud & Mayo, 2019). Current network architectures are designed on a trial-and-error basis, as the decision on which architecture to choose depends on the task the neural network needs to perform. Recently, it has been shown that the success of architectural search techniques depends significantly on a careful design of the search space (Li & Talwalkar, 2019).

The suggestion that additional data from studies provides significant improvement in drug response prediction based on deep learning, calls for integration of different cohorts (Adam et al., 2020). It must be highlighted that validation for one cohort is hard and individual. When combining different cohorts, identical clinical and genomic features must be considered as well as normalization of gene expression values. The presented overview (figure 1) shows the applied procedure to predict the target variable. In this master thesis, the PFS/OS was chosen as the endpoint but applicability on other target variables such as Response Evaluation Criteria in Solid Tumors (RECIST)[[1]](#footnote-1) or the decision whether a combination or monotherapy leads to better results are imaginable as well.

A picture containing diagram

Description automatically generated

Figure 2: Overall workflow

The overall workflow of the model generation starts with the data preparation of the gene expression values and clinical variables of the single cohort or merged cohorts of melanoma patients. Merging cohorts requires additional preparation as the samples must be normalized, must have identical gene signatures and identical clinical variables. After training and validation of the ANN, the chosen target variable can be predicted.

## 1.4 Objectives

The aim of this master thesis is to develop a machine learning workflow that predicts the success of immune checkpoint therapies in melanoma patient cohorts. Individual therapy response prediction is here based on gene expression data in combination with additional genomic and clinical features. Overall survival (OS) and progression free survival (PFS) of the patients are the outcome variables of interest. Molecular biology is confronted with the issue of high-dimensional next-generation-sequencing data with small (patient) sample size, which is a limitation for machine learning. The presented workflow addresses the resulting problems such as overfitting and class imbalance. Furthermore, this thesis aims at automation of the workflow and finds high applicability to different cohorts.

# 2. Methods

## 2.1 Datasets

### 2.1.1 Boston Housing Dataset

The used Boston house price dataset is included in the Sklearn package. It consists of 506 instances with 13 numerical/categorical features and the target. The attributes include information such as location, environment, and crime rate. The dataset has been used to address a regression problem by predicting the median value of the homes. Original data can be viewed in the Sklearn manual.

### 2.1.2 Melanoma ICB Cohorts

All gene expression profiles used in this study were retrieved from published melanoma cohorts treated with different types of ICB therapies (table 1). Since size and type of therapy differ for each dataset, it is necessary to divide the cohorts by type of therapy and highlight the size of each subgroup. A patient in each cohort may have received either one immune checkpoint inhibitor (monotherapy), one inhibitor after the other (sequential therapy) or a combination of inhibitors (combination therapy). In addition, some melanoma patients have already undergone MAPKi therapy. For each dataset, the Transcripts Per Million (TPM) values of the gene expression and information on treatment and survival time were provided from the published datasets. Depending on the cohort, additional information about the patients or tumor environment was also provided. However, additional information is often only available for one cohort and not for the others, making it difficult to combine cohorts, for example, age is available for the Gide/Hugo/VanAllen cohort but is not available in the Liu/Riaz cohort. Since the Liu dataset containing 121 entries is the largest and the success of deep learning methods follows the collection of large, standardized datasets, this dataset was mainly used to test the ANN workflow.

Table 1: Datasets of melanoma patients undergoing ICB therapy

Presented are the different RNA-Seq datasets with number of samples (n) and number of samples per therapy type. The types of therapies, divided into monotherapy, sequential therapy, and combination therapy, were extracted from the original papers. The number of patients who received prior MAPKi therapy before the specific therapy is indicated in parentheses.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mono**  **Therapy** | | | **Sequential**  **Therapy** | | **Combination**  **Therapy** | |
| RNA-Seq  Data | **Nivo** | **Pembro** | **Ipi** | **1. Ipi**  **2. Nivo** | **1. Ipi**  **2. Pembro** | **Ipi+Nivo** | **Ipi+Pembro** |
| **Liu**  **(n=121)** | 30  (3 previous MAPKi) | 44  (4 previous MAPKi) |  | 21  (2 previous MAPKi) | 26  (8 previous MAPKi) |  |  |
| **Gide**  **(n=72)** | 9  (1 previous MAPKi) | 32  (13 previous MAPKi) |  |  |  | 7  (5 previous MAPKi) | 24  (13 previous MAPKi) |
| **Riaz**  **(n=78)** | 78 ([[2]](#footnote-2)) |  |  |  |  |  |  |
| **Hugo**  **(n=27)** |  | 27  (12 previous MAPKi) |  |  |  |  |  |
| **Van Allen**  **(n=42)** |  |  | 42  (7 previous MAPKi) |  |  |  |  |

Further information on the individual datasets is presented below.

#### 2.1.2.1 Liu Dataset

The Liu dataset was extracted from the Liu et al. publication in 2019 (Liu, Schilling, et al., 2019). Patients of this cohort received either pembrolizumab or nivolumab as monotherapy or ipilimumab and pembrolizumab/nivolumab as sequential therapy. Some of them also received a priori MAPKi treatment. Besides the 20849 gene expression values in TPM format, the following 6 clinical variables were chosen for the analysis: gender, heterogeneity, nonsynonymous mutation, stage of cancer, LDH elevated and type of treatment (pembrolizumab or nivolumab).

#### 2.1.2.2 Gide Dataset

The Gide dataset was extracted from the Gide et al. publication in 2019 (Gide et al., 2019). Patients of this cohort have undergone either monotherapy with pembrolizumab or nivolumab or combination therapy by adding ipilimumab as first agent, respectively. Patients with BRAFV600 variants were defined as patients with prior MAPKi[[3]](#footnote-3).

#### 2.1.2.3 Riaz Dataset

The Riaz dataset was extracted from the Riaz et al. publication in 2017 (Riaz et al., 2017). This cohort consists of patients undergoing nivolumab monotherapy. Information on prior MAPKi therapy was not available.

#### 2.1.2.4 Hugo Dataset

The Hugo dataset was extracted from the Hugo et al. publication in 2019 (Hugo et al., 2016). This cohort consists of patients undergoing pembrolizumab monotherapy. Patients with prior MAPKi were included.

#### 2.1.2.5 Van Allen Dataset

The Van Allen dataset was extracted from the Van Allen et.al publication in 2015 (van Allen et al., 2015). This cohort consists of patients undergoing ipilimumab monotherapy. Patients with BRAF inhibitor following ipilimumab were defined as patients with prior MAPKi.

## 2.2 Programs

### 2.2.1 Jupyter Notebook and Python Scripts

The presented ANN workflow is written using the open-source web application Jupyter Notebook (Jupyter core version:4.6.3, jupyter-notebook:6.1.4, https://jupyter.org/) in combination with the programming language Python (Version 3.7.3, https://www.python.org/). All scripts can be found on <https://github.com/chrissikath/ANNPredictionICBTherapy>.

### 2.2.3 Packages

For the development of the ANN workflow several Python packages were used. These are listed in table 2. All packages were installed via pip (20.3.3) and used on Debian GNU/Linux 10 (buster).

Table 2: Used Python packages with version and purpose

|  |  |  |
| --- | --- | --- |
| Package Name | Version | Purpose |
| tensorflow | 2.3.1 | Development of ANN workflow |
| keras | 2.2.4 | This Python interface is used to work with the TensorFlow backend |
| seaborn | 0.10. | Drawing statistical graphics like confusion matrix |
| scikit-learn | 0.23.2 | Used for dimensionality reduction, model selection, pre-processing and the general machine learning workflow |
| pandas | 1.1.0 | Handling of the datasets |
| imblearn | 0.7.0 | Upsampling method SMOTE |
| matplotlib | 3.0.2 | Required for seaborn |
| scipy | 1.1.0 | Required for numerical and scientific calculations |
| numpy | 1.16.2 | Required for numerical and scientific calculations |
| csv | 1.0 | Handling of csv files |
| argparse | 1.1 | User-friendly command-line interface |
| logging | 0.5.1.2 | Creation of the log of the software process |

## 2.3 Workflow

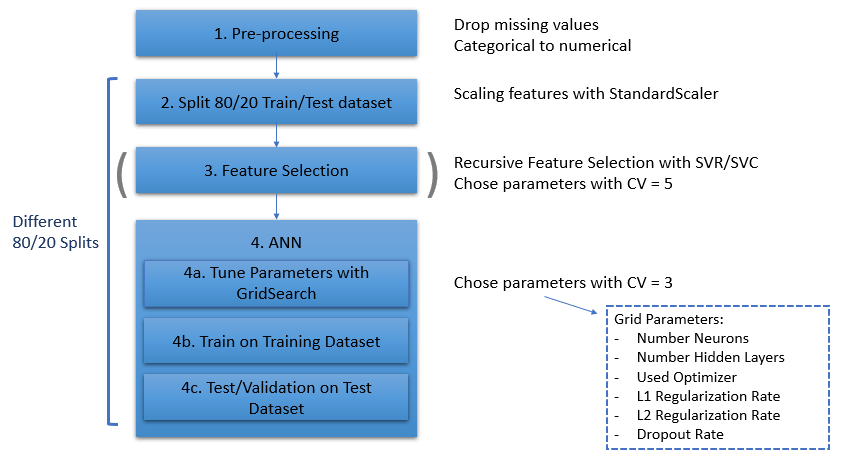


Figure 3: ANN workflow

The used ANN workflow can be divided into 4 parts. First, the data needs to be pre-processed (1) and split into a training and test dataset (2). The next step is optional and is used for dimension reduction with feature selection (3). The actual training and testing of the ANN consist of tuning the parameters with GridSearch (4a), training of the ANN with those parameters (4b) and the final validation of the model with the held back test dataset (4c).

### 2.3.1 Pre-Processing of Data

Data pre-processing is a crucial step in the workflow of ANN modelling. First, all samples with missing values in any feature were discarded. This leads to 506 samples for the Boston Housing dataset and 118 samples for the Liu dataset. Since the available data consists not only of numerical but also categorical variables such as treatment, all input variables must be converted to a numerical representation. In the Boston Housing dataset, all 13 characteristics are already numeric. All categorical characteristics in Liu are therefore transformed with LabelEncoder function by Sklearn. In the case of binary predictions, the result must be divided into two groups by a suitable threshold value. For our Sanity Check with the Boston housing dataset a MEDV (Median value of owner-occupied homes in $1000) smaller than 22 is considered as low-price houses and MEDV greater than 22 as high price houses. In the case of Liu, patients with PFS of less than 365 days or OS of less than 730 days are considered as short survivors and those with PFS of more than 365 or OS of more than 730 are considered long survivors. To deal with the problem of imbalanced data by dividing the data into two groups, the SMOTE-NC oversampling method can be applied. The Synthetic Minority Oversampling Technique (Numerical Categorical) upsamples the minor represented data by using a k-nearest neighbour algorithm to create synthetic data. SMOTE-NC was used because the data consists of categorical and continuous features. The upsampling is only applied to the training dataset and not to the test dataset.

To get all features to have the same scale and to speed up the gradient descent algorithm feature scaling is applied. There are two common ways to scale data: Min-Max-Scaling or Standardization. Apart from Min-Max-Scaling, Standardization is much less affected by outliers, for example, huge differences between gene expression values. For this case, the StandardScaler method provided by Sklearn is used. The StandardScaler assumes that data is normally distributed within each feature and scales them such that the distribution is centered around 0 with a standard deviation of 1. To avoid data leakage, the scaling is calculated on the training set and then used to transform the test set. An overview of the pre-processing steps is presented in the following table.

Table 3: Pre-processing of datasets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sample Size** | **Normalization** | **If Outcome Binary** | **Features** |
| Boston Housing Dataset | 506 samples | StandardScaler:  z = (x - u) / s | MEDV ≥ 22→ 1  MEDV < 22→ 0  (SMOTE-NC training data) | all 13 |
| Liu et al. | Drop NAs  From 121 to 118 samples | StandardScaler:  z = (x - u) / s | PFS ≥ 365 → 1  PFS < 365 → 0  Or  OS ≥ 730 → 1  OS < 730 → 0  (SMOTE-NC training data) | all 20849 genes with 'gender', 'heterogeneity', 'nonsynonymous mutation', 'stage', 'LDH elevated', 'treatment' |

### 2.3.2 Feature Selection Technique

For feature selection the RFE (Recursive Feature Selection) method provided by Sklearn is used. The external estimator that assigns weights to features is either the SVR (support vector regression) for continuous prediction or the SVC (support vector classification) for binary classification. The selected kernel must be linear, otherwise it is not possible to calculate the coefficients used to select the features. The penalty tuning parameter C is tuned using GridSearch with 3-fold cross-validation. The desired number of selected features can be chosen with the n\_features-flag. The default value is 50 features. Gene selection is exclusively performed on the training set to prevent data leakage.

### 2.3.3 Implementation of Neural Network

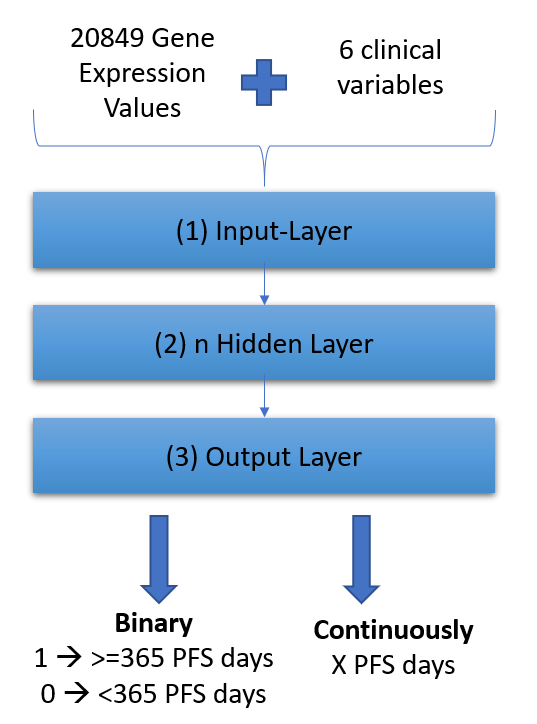


Figure 4: Neural Network Architecture

The neural network architecture takes the gene expression value and clinical features as input and predicts either classification in >= 365 PFS days or <365 (binary) or the exact number of PFS days (continuously). The features are processed from the input layer over n hidden layers to the output and final prediction layer

The architecture of the ANN consists of: (1) an input layer (genes + clinical variables) (2), one or more hidden layers with dropout and (3) an output layer. All neurons of the three layers are fully connected. For initialization of weights and biases Keras default values were used: the default kernel initializer is 'glorot\_uniform' and the default bias initializer is 'zeros'. The network was trained using the Adam optimizer with a variable learning rate.

***Input Layer***

The ANN takes the gene expression TPM value for each gene and the clinical variables from each patient as input after pre-processing as described in 2.3.1.

***Hidden Layer***

The hidden layer consists of a dense layer followed by a dropout layer with variable dropout rate. The used activation function for the input layer is ReLU (Rectified Linear Unit). For kernel regularization of the dense layer a variable L1 regularization and L2 regularization is applied. Number of neurons is held in a variable manner, too. The number of hidden layers (dense + dropout layer) can be chosen variably.

***Output layer***

The output layer has only one node. Depending on whether the predicted value is binary (e.g., PFS ≥ 365 or PFS <365) or continuous (e.g., PFS days) the activation and loss function are selected as follows: Binary: activation: sigmoid, loss: binary\_crossentropy or Continuous: activation: linear, loss: mse. As an optimizer the Adam algorithm and SGD with momentum and Nesterov momentum are implemented.

***Fine-Tuning Parameters***

The resulting variable parameters for the fine-tuning are the number of neurons per layer, the number of hidden layers, the learning rate, the L1 regularization rate, the L2 regularization rate, the kind of optimizer, and the dropout rate.

### 2.3.4 Fine-Tuning

Before fine-tuning the data is split into 80 and 20 percent of total data for model training and model testing, respectively, to test the model’s performance on an unseen dataset. Only the 80% training data (n=94) are used to tune and train the model. Validation with 20% (n=24) is performed as described in 2.3.5.

All variable parameters of the ANN are fine-tuned using GridSearchCV with CV=5 and a validation split of 20%. After finding the fine-tuning parameters the ANN is trained with these best parameters. The model is trained with shuffle mode to ensure that the training instances are independent and identically distributed (Géron, 2019, p. 153). To reduce training time and find the optimal learning rate, two Keras callbacks are used which monitor the loss on the validation split set: EarlyStopping and ReduceLROnPlateau. Early stopping stops the training when the validation loss no longer improves for 20 epochs (default value patience). For learning rate, ReduceLROnPlateau checks if validation loss is not improving for five epochs and reduces the learning rate by factor 0.1 until it reaches the minimal learning rate of 1-6.

### 2.3.5. Model Validation

For evaluation of the trained ANN, several 80/20 splits are performed with a random permutation cross-validator (ShuffleSplit). Different metrics are used depending on the desired outcome. For the binary classification besides accuracy the sensitivity, specificity, positive predictive value, negative predictive value, and confusion matrix is used. For the continuous outcome, the mean absolute error (MAE), root mean squared error (RMSE) and R² are used.

### 2.3.6 Applicability

To ensure applicability and increase reproducibility, the program can be adapted to different datasets via command-line arguments. Further information can be found in the README.md via the GitHub repository.

# 3. Results

## 3.1 Sanity Check with Boston Housing Dataset

To check whether the neural network has the ability to make predictions under good conditions, the workflow was tested with the well-known dataset "Boston Housing", where the aim is to predict house prices depending on 13 different characteristics. To simulate a binary prediction the target was divided into “expensive” and “cheap” house prices where the cut-off was 22 (corresponds $22.000). The chosen GridSearch parameters to tune were: number neurons = 8,16,32,64; number hidden layers = 1,2,3; optimizer = SGD, Adam; learning rate = 0.01; dropout rate = 0.0,0.2,0.4,0.6; L1 regularization rate = 0.0,0.2,0.4; L2 regularization rate = 0.0,0.2,0.4. **The constructed workflow archives an accuracy up to 95% over the five different 80/20 train/test set splits (table 1, left).**

The predictive power of the workflow depends on which split is chosen and there are still false positive and false negative predicted prices. Besides the accuracy, the best parameters chosen by the Grid Search also depend on the train/test split. The number of hidden layers goes from 1 to 3 layers and the number of neurons per layer varies between 32 and 64. Dropout rate was varying between 0.0 and 0.2. L1 and L2 regularization was chosen for each split without any regularization weight.

For a better comparison and since the Boston Housing dataset simulates a regression problem and not a binary decision, the predictive power of the workflow is tested with a continuous outcome (table 4, right). After five different train/test splits the neural network (similar GridSearch parameters) predicts a mean RMSE of 3.41, an MAE of 2.94 and an R² of 0.86. Again, the chosen best parameters depend on the specific train/test split, but no regularization was used in any of the five splits.

Table 4 Binary and continuous prediction of the Boston Housing dataset

Predictions on binary (left) and continuous (right) outcome for the Boston housing dataset are presented. For each of the five training/test splits, the calculated metrics and the best selected tuning parameters are presented. After finding the best parameters on the 80% training data of each split, the metrics on the 20% test data are presented. The mean and standard deviation of each metric is calculated after 30 runs with the best model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Binary Outcome** | | **Continuous Outcome** | |
|  | **Best Params** | **Metrics (mean +/- sd)** | **Best Params** | **Metrics (mean +/- sd)** |
| 0 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons:32  Number hidden: 3  Optimizer: Adam | **Accuracy 0.95 +/-(0.01)**  Sensitivity 0.94 +/-(0.01)  Specificity 0.96 +/-(0.01)  PPV 0.95 +/-(0.01)  NPV 0.95 +/-(0.01) | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons:32  Number hidden: 3  Optimizer: Adam | MAE 2.89 +/-(0.10)  **RMSE 4.65 +/-(0.12)**  R^2 0.73 +/-(0.01) |
| 1 | Dropout rate:0.2  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons:64  Number hidden: 2  Optimizer: SGD | **Accuracy 0.84 +/-(0.01)**  Sensitivity 0.85 +/-(0.01)  Specificity 0.84 +/-(0.01)  PPV 0.81 +/-(0.01)  NPV 0.87 +/-(0.01) | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons:64  Number hidden: 2  Optimizer: Adam | MAE 2.37 +/-(0.07)  **RMSE 3.03 +/-(0.09)**  R^2 0.91 +/-(0.01) |
| 2 | Dropout rate:0.2  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons:64  Number hidden: 2  Optimizer: Adam | **Accuracy 0.85 +/-(0.01)**  Sensitivity 0.81 +/-(0.01)  Specificity 0.88 +/-(0.02)  PPV 0.85 +/-(0.02)  NPV 0.85 +/-(0.01) | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons:64  Number hidden: 3  Optimizer: Adam | MAE 2.29 +/-(0.07)  **RMSE 2.90 +/-(0.09)**  R^2 0.90 +/-(0.01) |
| 3 | Dropout rate:0.2  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons:64  Number hidden: 3  Optimizer: Adam | **Accuracy 0.89 +/-(0.01)**  Sensitivity 0.87 +/-(0.02)  Specificity 0.90 +/-(0.02)  PPV 0.88 +/-(0.02)  NPV 0.89 +/-(0.01) | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons:64  Number hidden: 2  Optimizer: Adam | MAE 1.97 +/-(0.05)  **RMSE 2.55 +/-(0.08)**  R^2 0.92 +/-(0.00) |
| 4 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons:64  Number hidden: 1  Optimizer: adam | **Accuracy 0.90 +/-(0.01)**  Sensitivity 0.91 +/-(0.01)  Specificity 0.89 +/-(0.01)  PPV 0.88 +/-(0.01)  NPV 0.93 +/-(0.01) | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons:64  Number hidden: 2  Optimizer: adam | MAE 2.50 +/-(0.10)  **RMSE 3.92 +/-(0.16)**  R^2 0.83 +/-(0.01) |

**Compared to a competition from 2016 on Kaggle[[4]](#footnote-4) (figure 3) (which is an online community of data scientists and machine learning practitioners), in which 75 different teams took part, the RMSE of 3.41 is better than rank 7 (3.48) of the top 10 best results.**

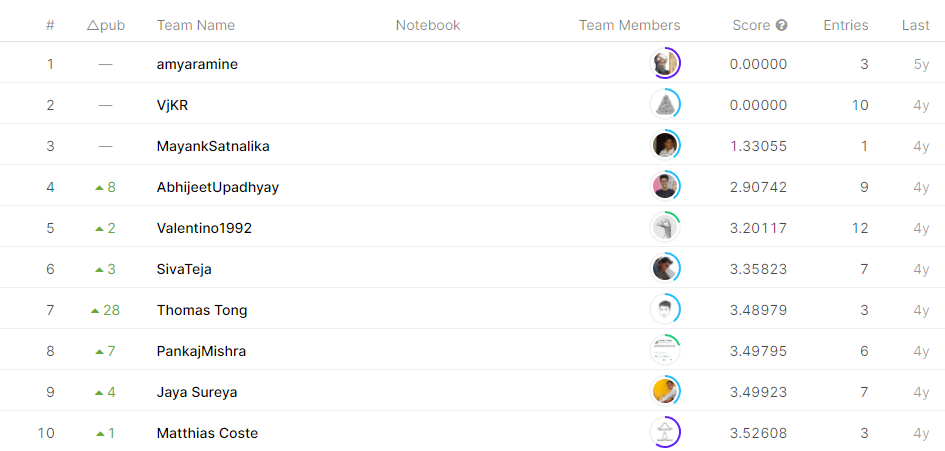


Figure 5: Leader board Boston Housing Kaggle competition 2016

The 10 best results from the final leader board from the 2016 competition have achieved best scores from 0.0 to 3.52 RSME. 75 Teams participated in the competition.

## 3.2 Predictions on Liu Dataset

In the following chapters, the developed workflow is now applied to the research question of individual therapy response prediction based on gene expression in combination with the described genomic and clinical features. To achieve the ultimate goal of a therapy decision based on a patient's genomic make-up, the value of the model needs to be tested in its ability to predict the outcome of a patient's ICB therapy based on their individual genomic make-up.

### 3.2.1 Without Feature Selection

Since the success of deep learning depends on the availability of large, standardized datasets, the Liu dataset is selected first as it contains the largest RNA sequencing dataset (see 2.1.2). Besides the 20849 gene expression values, the following 6 clinical variables were included as features: gender, heterogeneity, nonsynonymous mutation, stage of cancer, LDH elevated and treatment (pembrolizumab or nivolumab). For both, the binary as well as the continuous prediction of the PFS and OS five different splits were trained and tested with the same grid search parameters (number neurons= 3, 5, 10; number hidden layers= 1, 2, 3; L1 regularization rate= 0.0, 0.2, 0.4; L2 regularization rate= 0.0, 0.2, 0.4; learning rate= 0.01; dropout rate: 0.0, 0.2, 0.4; optimizer= Adam, SGD). The selection of appropriate parameters is based on a combination of testing and knowledge based on neural network predicting methods used in cancer prediction models summarized by (Daoud & Mayo, 2019). However, it was hard to find full details on network configuration, overfitting elimination techniques or concrete learning parameters since they were not provided in most publications.

In case of the continuous outcome, which means to predict the exact days of PFS/OS of a patient given the named features and gene expression profile, the ANNs result in the problem of “exploding gradients”. This happens when large updates to weights during training cause a numerical overflow or underflow. After using the gradient norm clipping function of Keras, the ANN overcomes the exploring gradients problem but still leads to unsatisfying results in the prediction metrics. The R² of all different splits are in all cases negative which means the model leads to arbitrarily worse results (attachment 3 and 4). Since the regression task is much more complicated to learn for an ANN and, as presented, leads to differences in the ANN structure, this thesis is focusing on the binary classification. Moreover, the interpretation of a “long” against “short” survivor is more applicable in clinical implementation when it comes to deciding whether a patient should receive a treatment or not.

Applied in the case of the binary outcome, up to 72% (OS) and up to 71% (PFS) of the test set were correctly predicted in the corresponding class (table 5, appendix 2). Besides the metrics, the best parameters chosen by the fine-tuning grid also differ for each split. Only L1 regularization was 0.0 for each of the five splits. As the test split is exceedingly small, other metrics should be used in addition to accuracy. It is worth mentioning that the specificity, for example, the ability to predict short survivors, with up to 93%, is much higher than the sensitivity with up to 49%, which predicts long survivors.

*Table 5: Binary prediction on Liu et al. dataset with OS as endpoint*

Predictions on binary classification based on OS for therapy response of patients from the Liu cohort are presented. For each of the five training/test splits, the calculated metrics and the best selected tuning parameters are presented. After finding the best parameters on the 80% training data of each split, the metrics on the 20% test data are presented. The mean and standard deviation of each metric is calculated after 30 runs with the best model with a confusion matrix example of the last validation run.

|  |  |  |  |
| --- | --- | --- | --- |
| **Split** | **Best Params** | **Metrics** | **Confusion Matrix** |
| 0 | Dropout rate:0.4  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 5  Number hidden: 2  Optimizer: Adam | **Accuracy 0.72 +/-(0.05)**  Sensitivity 0.30 +/-(0.17)  Specificity 0.93 +/-(0.07)  PPV 0.71 +/-(0.25)  NPV 0.73 +/-(0.04) |  |
| 1 | Dropout rate:0.4  L1 reg:0.0  L2 reg:0.2  Learning rate:0.01  Number neurons: 3  Number hidden: 2  Optimizer: Adam | **Accuracy 0.65 +/-(0.10)**  Sensitivity 0.49 +/-(0.14)  Specificity 0.73 +/-(0.14)  PPV 0.50 +/-(0.15)  NPV 0.74 +/-(0.06) |  |
| 2 | Dropout rate:0.2  L1 reg:0.0  L2 reg:0.4  Learning rate:0.01  Number neurons: 10  Number hidden: 3  Optimizer: SGD | **Accuracy 0.64 +/-(0.07)**  Sensitivity 0.34 +/-(0.10)  Specificity 0.79 +/-(0.08)  PPV 0.46 +/-(0.15)  NPV 0.71 +/-(0.04) |  |
| 3 | Dropout rate:0.4  L1 reg:0.0  L2 reg:0.4  Learning rate:0.01  Number neurons: 3  Number hidden: 1  Optimizer: SGD | **Accuracy 0.68 +/-(0.07)**  Sensitivity 0.45 +/-(0.15)  Specificity 0.82 +/-(0.07)  PPV 0.60 +/-(0.12)  NPV 0.72 +/-(0.06) |  |
| 4 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 5  Number hidden: 3  Optimizer: SGD | **Accuracy 0.63 +/-(0.04)**  Sensitivity 0.22 +/-(0.12)  Specificity 0.84 +/-(0.07)  PPV nan +/-(nan)  NPV 0.68 +/-(0.03) |  |

### 3.2.2 With Feature Selection

To overcome the problems of high dimensional data as described in 1.3.1, the feature selection is used to select 50 features before the actual training of the ANN. Since in the case of PFS prediction more features were present in the overlap of all five different splits, the selected features are presented here and the corresponding results for OS can be found in appendix 8. Consequently, the feature selected worked better in the case of PFS prediction where none feature selected were present in all five splits.

The results from the beforehand feature selection on PFS is presented in table 7. Two genes were present in all five different splits (*HIKESHI* and *LIN28A*) and four genes were present in four out of five splits (*EMP1, HSH2D, NHEJ1, PROC, HAGLR, NUP107*) (table 6). A list of all 50 features for each split can be found in appendix 5.

Table 6: Occurrences of selected features by the RFE algorithm in the different train/test splits

Presented are the selected features by the RFE algorithm based on SVC and their occurrences in each of the five different train/test splits.

|  |  |  |
| --- | --- | --- |
| **Item** | **Occurrences** | **Present In** |
| *HIKESHI* | 5 | Split 0, Split 1, Split 2, Split 3, Split 4 |
| *LIN28A* | 5 | Split 0, Split 1, Split 2, Split 3, Split 4 |
| *EMP1* | 4 | Split 0, Split 2, Split 3, Split 4 |
| *HSH2D* | 4 | Split 1, Split 2, Split 3, Split 4 |
| *NHEJ1* | 4 | Split 0, Split 1, Split 2, Split 3 |
| *PROC* | 4 | Split 1, Split 2, Split 3, Split 4 |

After the feature selection, the ANN is trained for each of the five splits with those selected 50 features on each split and the additional 6 clinical features. For reasons of comparability, the same grid as in 3.2.1 was used. The accuracy for binary prediction of PFS varied from 68% to 86% (table 7), while prediction of OS varied from 49% to 79% (appendix 8). Thus, not only was the workflow able to find better features for PFS than for OS, but it was also better at predicting patients’ PFS than OS after feature selection.

Table 7: Binary prediction on Liu et al. dataset after feature selection with PFS as endpoint

Predictions on binary classification after feature selection of 10 different genes for the Liu cohort are presented. For each of the 5 training/test splits, the calculated metrics and the best selected tuning parameters are presented. After finding the best parameters on the 80% training data of each split, the metrics on the 20% test data are presented. The mean and standard deviation of each metric is calculated after 30 runs with the best model.

|  |  |  |
| --- | --- | --- |
| **Split** | **Best Params** | **Metrics** |
| 0 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 3  Number hidden: 1  Optimizer: SGD | **Accuracy 0.86 +/-(0.04)**  Sensitivity 0.59 +/-(0.05)  Specificity 0.93 +/-(0.05)  PPV 0.72 +/-(0.17)  NPV 0.90 +/-(0.01**)** |
| 1 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 3  Number hidden: 1  Optimizer: Adam | **Accuracy 0.85 +/-(0.03)**  Sensitivity 0.50 +/-(0.11)  Specificity 0.94 +/-(0.03)  PPV 0.69 +/-(0.10)  NPV 0.88 +/-(0.02) |
| 2 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 5  Number hidden: 1  Optimizer: Adam | **Accuracy 0.76 +/-(0.02)**  Sensitivity 0.28 +/-(0.10)  Specificity 0.89 +/-(0.02)  PPV 0.40 +/-(0.08)  NPV 0.83 +/-(0.02) |
| 3 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 3  Number hidden: 3  Optimizer: SGD | **Accuracy 0.76 +/-(0.04)**  Sensitivity 0.47 +/-(0.13)  Specificity 0.84 +/-(0.04)  PPV 0.44 +/-(0.09)  NPV 0.86 +/-(0.03) |
| 4 | Dropout rate:0.4  L1 reg:0.0  L2 reg:0.2  Learning rate:0.01  Number neurons: 10  Number hidden: 2  Optimizer: Adam | **Accuracy 0.68 +/-(0.03)**  Sensitivity 0.19 +/-(0.05)  Specificity 0.81 +/-(0.04)  PPV 0.21 +/-(0.07)  NPV 0.79 +/-(0.01) |

Compared to the ANN without feature selection in section 3.2.1, there is an improvement for PFS prediction with feature selection since the best achieved accuracy up to 86% compared to 71% in the case without feature selection. The best working grid parameters showed that except for the last split no regularization was needed anymore. The number of neurons and the number of hidden layers used were not constant between the five splits.

## 3.3. Predictions on Combination of Cohorts

Addressing the problem of small size datasets, the enlargement of the training dataset is performed. For this purpose, the non-homogeneous cohorts presented in 2.1.2 were combined and the resulting dataset comprised 322 samples. Beforehand the raw counts of the cohorts were generated as described in the publication by Liu et al. (2019) because TPM does not account for differences in library composition or distinct levels of rRNA contamination of each cohort (Zhao2020). Due to different experimental conditions and/or different sequencing protocols, between sample normalization was then applied via the median of ratios method by DESeq2 to remove systematic and technical effects. As no clinical or genomic characteristic other than OS was available for all cohorts, predictions were based on gene expressions only. Treatment information about each patient in each cohort was integrated in a binary encoding.

For the binary prediction of OS five different splits were trained and tested with the following grid search parameters: number neurons= 3, 5, 10; number hidden layers= 1, 2, 3; L1 regularization rate= 0.0, 0.2, 0.4; L2 regularization rate= 0.0, 0.2, 0.4; learning rate= 0.01; dropout rate: 0.0, 0.2, 0.4; optimizer= Adam. The resulting prediction accuracy of OS varied between 61% to 66% (table 8). In all five splits, the specificity was above 70% but sensitivity was only up to 55%, showing that the prediction of short survivors worked better than that of long survivors. Except for L1 regularization, none of the grid search parameters were constant between the five splits, but L2 regularization was used in almost all splits. Compared to the prediction results on the Liu cohort only, metrics fluctuations were much less, but in general one cannot conclude a significant improvement of the prediction results when combining the cohorts.

*Table 8: Binary prediction on the combination of cohorts with OS as endpoint*

Predictions on binary classification of OS for the combination of all melanoma cohorts are presented. For each of the 5 training/test splits, the calculated metrics and the best selected tuning parameters are presented. After finding the best parameters on the 80% training data of each split, the metrics on the 20% test data are presented. The mean and standard deviation of each metric is calculated after 30 runs with the best model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Split** | **Best Params** | **Metrics** | **Confusion Matrix** |
| 0 | Dropout rate:0.2  L1 reg:0.0  L2 reg:0.2  Learning rate:0.01  Number neurons: 5  Number hidden: 3  Optimizer: Adam | **Accuracy 0.67 +/-(0.04)**  Sensitivity 0.40 +/-(0.13)  Specificity 0.81 +/-(0.09)  PPV 0.53 +/-(0.08)  NPV 0.73 +/-(0.03) |  |
| 1 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.4  Learning rate:0.01  Number neurons: 3  Number hidden: 2  Optimizer: Adam | **Accuracy 0.65 +/-(0.02)**  Sensitivity 0.33 +/-(0.11)  Specificity 0.82 +/-(0.07)  PPV 0.49 +/-(0.05)  NPV 0.71 +/-(0.02) |  |
| 2 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.2  Learning rate:0.01  Number neurons: 5  Number hidden: 1  Optimizer: SGD | **Accuracy 0.66 +/-(0.02)**  Sensitivity 0.35 +/-(0.06)  Specificity 0.82 +/-(0.03)  PPV 0.49 +/-(0.04)  NPV 0.71 +/-(0.02) |  |
| 3 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.2  Learning rate:0.01  Number neurons: 3  Number hidden: 3  Optimizer: Adam | **Accuracy 0.61 +/-(0.03)**  Sensitivity 0.25 +/-(0.07)  Specificity 0.80 +/-(0.06)  PPV 0.39 +/-(0.05)  NPV 0.68 +/-(0.01) |  |
| 4 | Dropout rate:0.2  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 3  Number hidden: 3  Optimizer: Adam | **Accuracy 0.60 +/-(0.04)**  Sensitivity 0.39 +/-(0.06)  Specificity 0.71 +/-(0.05)  PPV 0.41 +/-(0.05)  NPV 0.70 +/-(0.02) |  |

# 

# 4. Discussion

The identification of the best treatments for personalized response prediction does not only include biological but also computational challenges due to limited data availability and algorithmic shortcomings (Adam et al., 2020). The use of neural networks as a machine learning approach can provide a promising workflow for therapy response models based on microarray data (Daoud & Mayo, 2019). For the development of clinically relevant prediction methods, the reproducibility of the model is of immense importance. Since the architecture of the model depends on the given data, a variable neural network and data workflow is required. The Sanity Check presented here (see 3.1) proves the applicability of the developed workflow by comparing my RSME of 3.41 to rank 7 (3.48) of a Kaggle competition on a widely used dataset from 2016. Independent of binary or continuous outcome, the workflow can train a neural network based on a given grid through fine-tuning the parameters to achieve the best possible result. As a result, the workflow should be able to predict both binary and continuous outcomes if the data structure allows for this and an underlying structure can be identified.

Applied to the prediction of ICB therapy response, a deep neural network may open a new chapter for personalized therapy of melanoma patients. Investigations of the current data situation showed limitations of sample size in standardized datasets for melanoma patients treated with ICB therapy. Therefore, the main challenge for this work arose from dealing with shortcomings arising from cohorts of up to ~100 patients. The overview promoted in table 1 showed the variety of the different treatments and treatment combinations (see 2.1.2). As Adam noted in 2020, machine learning methods that incorporate the peculiarities of cancer biology are a huge area of work that should be addressed to take advantage of high-throughput data and computational advances from deep neural networks. Hence, this thesis addresses the development of a computational workflow which is shaped by the biological characteristics of the research question.

Predictions of PFS/OS for patients in the Liu cohort with ICB treatment pembrolizumab or nivolumab, as presented in 3.2.1, showed that the workflow was able to learn from the 20849 gene expression values in combination with the six clinical variables. The best result achieved a prediction accuracy of 72% for OS (table 5) and 71% for PFS (appendix 2). This is comparable to other ways of predicting anti-PD-1 therapy success in melanoma such as widely used biological predictors tumor heterogeneity, ploidy and purity or MHC-II HLA, LDH and presence of lymph node metastases. Equivalent results were produced when stratifying for ipilimumab-naive patients (ten-fold cross-validation AUC=0.73) and better results for ipilimumab-treated patients (five-fold cross-validation AUC=0.83) (Liu, Schilling, et al., 2019). Of note, responders were selected as a function of Response Evaluation Criteria in Solid Tumors (RECIST) metrics rather than PFS/OS and predictions were based on logistic regression. Comparisons with other approaches to predicting response to ICB treatment were not possible due to the lack of ANN models for melanoma patients. Noticeable, huge fluctuations in the metrics were observed. This is due to the small sample size of the Liu cohort. The difference of the best parameters chosen for each 80/20 split shows that the structure of the neural network depends on the provided training data. Without applying regularization (L1, L2 or dropout), the ANN showed severe overfitting when looking at validation metrics such as loss or accuracy. Therefore, regularization was applied in the grid parameters. Since L1 regularization rate was the only parameter without fluctuations it can be concluded that the Lasso regularization does not have any positive effect on the predictive ability although Lasso is proven to be a useful tool when applied on microarray data (Ghosh & Chinnaiyan, 2005). However, since dropout and L2 regularizations were also applied, there may be no need for L1 regularization. Algamal & Mohammed Ali (2017) stated three shortcomings when applying L1 regularization whereas the most important one might be that L1 regularization cannot perform effectively if there is high correlation among genes. Since this is the case for gene expression data where genes are forming a group when sharing the same biological pathway, this might be one reason why L1 regularization might not work (Zou & Hastie, 2005). This issue could be addressed with the successfully applied autoencoder which is another promising neural network architecture that is also addressing the problem of high dimensionality (Lin et al., 2017; Rampášek et al., 2019; Tan et al., 2016). The structure of an autoencoder compresses its input and tries to reconstruct the original data from the compressed representation. Because of its compression ability autoencoders are used for feature selection based on gene expression (Way & Greene, 2018). Way & Greene (2018) showed its usefulness by compressing a 5000-dimensional gene expression profile into 100 dimensions. Another neural network approach used by Chiu et al. (2019) was the prediction of prognosis of breast cancer by using Bayesian Neural Networks. They also address the problem of overfitting in a natural way by considering entire distributions of answers (Uusitalo, 2007).

As feature selection is a common approach in dealing with high dimensionality (Saeys et al., 2007), it was applied to test whether it leads to an improvement in OS/PFS prediction. The best-chosen parameters revealed that after feature selection no L1 or L2 regularization is needed anymore since the regularization rate was 0.0 for each of the splits (table 7 and appendix 8). Comparing PFS against OS the feature selection approach showed inconsistent improvement. Noticeably, the prediction for PFS with feature selection led to an improvement (from mean 62,4% to 78,2%) but not for OS (from mean 66,4% to 65,6%). The reason for this could be that the feature selection in the case of PFS prediction had a greater overlap compared to the OS prediction. A good gene signature should conclude genes that have a predictive value for each of the five splits. Therefore, the RFE based on the SVC was for the OS case not able to find a good gene signature which consequently led to worse results in the prediction. The RFE algorithm needs the weight of each feature given by the coefficients of the SVC to decide which feature is recursively removed. However, no coefficients are available for SVM with other than linear kernels in Sklearn. When we think of gene interactions and their multicollinearity, a linear kernel may not be suitable to model those interactions. In the context of working with gene expression data, multiple algorithms and reviews have been published to describe all existing methods for feature selection but the selection of the correct feature selection strategy remains an enormous problem (Perez-Riverol et al., 2017). Due to the diversity of feature selection methods available in bioinformatics, it is worth looking at other feature selection methods besides the used Recursive Feature Elimination based on SVM (Saeys et al., 2007). A feasible way could be dimension reduction via Principal Component Analysis (PCA) to reduce the input layer of the neural network. Like Prerez-Riverol et al. (2017), where univariate filtering methods were combined with correlation matrix and PCA strategies, a combination of PCA and deep learning could be a useful way to overcome the disadvantages of high dimensionality. This variety of feature selection methods shows the complexity but also the opportunities when working with transcriptomic data. However, the improvement in the case of PFS prediction with preceding feature selection showed that meaningful feature selection could have a valuable benefit for the model.

Chart

Description automatically generated

Figure 6: Survival analysis of identified genes potential predictive for ICB efficacy

Univariate survival analysis with A. Kaplan-Meier plots based on groups with high/low mRNA expression level and survival analysis with B. multivariate Cox proportional-hazard models including the six clinical variables. Given are the hazard ratio and its p-value.

In order to be able to classify the relevance of the selected features, the genes were examined for their strength as predictors for PFS (appendix 6). Of those, only mRNA expression of HIKESHI, H2HD and LIN28A is associated with the survival of the patients (figure 4A). Interestingly, by analysing the interaction of the genes in combination with the six clinical features on survival, using multivariate Cox proportional-hazards models, HIKESH and LIN28A lose their predictive value (figure 4B). However, in the multivariate analysis HSH2D, which is involved in T-cell activation, showed a positive effect on survival. Therefore, HSH2D may serve as a biomarker with some predictive value for ICB therapy.

A major problem addressed in this work is that recent studies lack external validation of the performance of their models (Kourou et al., 2015). The reproducibility of the results requires validation in larger and independent cohorts (Liu, Schilling, et al., 2019). This was the last step of the analysis where the different melanoma cohorts were combined (see 3.3). The predictions based on the larger dataset led to fewer fluctuations in their metrics and thus to an improved informative value. The accuracy in the case of OS prediction reached results up to 67% (before in Liu cohort was up to 71%). Specificity, which means the prediction of short survivors, was higher with up to 79% (before in Liu was up to 93%) compared to the sensitivity with up to 55% (before in Liu was up to 49%) which can be explained by the fact that short survivors are overrepresented in the dataset. As running the developed workflow on the combined datasets could not achieve better than 67% accuracy, combining cohorts may still require a larger sample size due to the high variability in their data structure. This is due to the different treatment and pre-treatment variants and should be investigated when more data are available including either ICB monotherapy only or ICB combination therapy. For this purpose, better data sharing and the public release of used models is necessary (Adam et al., 2020). A further difference between the combined cohorts was the library preparation protocol used for mRNA sequencing. Genes may be under/over-represented depending on, for example, the rRNA depletion protocol used.

At the present time, treatment combinations of ICB and little available data make it difficult to achieve a final result, thus further investigations are necessary, for example, stratifying for treatment combinations. Steps in the reproducibility of ICB treatment response prediction have been achieved through the development of a variable workflow, the integration of different melanoma cohorts based on their different treatments and the pooling of available clinical and genomic features. The variability of the workflow allows it to be applied to diverse and differently structured datasets and the GridSearch function could relieve the burden on humans by reducing the workload and enable clinical applicability.

In summary, this thesis presents a workflow with proven performance if the available data provide useful information for the regression or classification problem. By incorporating several approaches such as regularization and feature selection, the workflow handles high dimensional low sample size data which is crucial for therapy response prediction based on transcriptomic data. This master thesis furthermore illustrates the problems of class imbalance and data heterogeneity arising from therapy prediction based on transcriptomic data and the need for further research handling low sample size and high dimension data. The workflow developed in this master thesis addresses the existing gap of deep learning based prediction of clinical benefit from ICB therapy of melanoma patients and thus serves as a starting point for personalized treatment decision-making of melanoma patients undergoing ICB therapy.

# 5. Outlook

There still exist unsolved problems with ICB therapy response prediction, however, also opportunities to address the complexity of cancer transcriptomic data with clinical annotations. Most importantly, the limited sample size of melanoma patients treated with ICB is restricting the potential of therapy response prediction models. However, it is only a matter of time until more data will be available. In the meantime, the workflow could be adapted by including more datasets that are publicly available such as high-throughput gene expression data from the Gene Expression Omnibus (GEO) database. Moreover, datasets from other cancer types could be used to achieve larger training sets. Combining different cohorts with different treatments and different RNA sequencing workflow remains a difficult challenge. Thus, collecting high quality datasets in large-scale open-access databases could lead to a meaningful way to increase accessibility (Niu et al., 2020). Another method dealing with small sample size is imaginable like the use of pretraining or transfer learning. Thereby the algorithm aims to improve its performance via transferring knowledge from auxiliary data of a related task. S. Liang et al. (2018) showed improved performance using unlabelled data from other sources to pretrain their model which consequently significantly increased the number of samples available. Convolutional Neural Networks and their ability in transfer learning are also powerful when it is difficult to obtain a large number of available omics data (Saeys et al., 2007).

Since this thesis focused on the prediction of PFS/OS as the endpoint of therapy response, it would be interesting to use other endpoints such as RECIST. However, grouping RECIST criteria into a binary variable is non-trivial due to the existence of mixed responders. Even IC50 values have been used in some drug response prediction studies, with the task framed as a regression problem (Sakellaropoulos et al., 2019). To date, there exists no international standard to measure response for preclinical in vitro models which prohibits fair comparisons between response prediction models (Adam et al., 2020).

Besides the choice of a suitable endpoint, there exist many input features worth looking at. The summary of Costello et al. (2014) showed leading trends for successful methods. It turns out that besides the ability to model nonlinear relationships between input and output, incorporation of prior knowledge such as biological pathways lead to better results than without (Adam et al., 2020). Ways to incorporate could be with graph-convolutional networks or conditional scaling of the input features according to the domain knowledge. Most computational feature selection methods also do not consider the nature of the biological data. Like Yousef et al. (2021) proposes, incorporating domain knowledge from external biological resources during gene selection should be included additionally to statistical metrics. Prior biological knowledge not only from transcriptomic data can lead to improvement of interpretability and predictive performance.

There are various ways ahead to improve specifics of personalized therapy response prediction. By incorporating other dimensionality reduction methods, transfer learning and other machine learning approaches into an automated workflow, as presented in this master thesis, deep learning could offer significant improvements over traditional machine learning approaches. With additional data from future studies and the use of related datasets, deep learning could be used as a clinical decision-making tool for ICB therapy in melanoma patients.

# Bibliography

Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. *Npj Precision Oncology*, *4*(1). https://doi.org/10.1038/s41698-020-0122-1

Ahn, J. (2006). *High Dimension, Low Sample Size Data Analysis*.

Algamal, Z. Y., & Lee, M. H. (2015a). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in Biology and Medicine*, *67*, 136–145. https://doi.org/10.1016/j.compbiomed.2015.10.008

Algamal, Z. Y., & Lee, M. H. (2015b). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in Biology and Medicine*, *67*, 136–145. https://doi.org/10.1016/j.compbiomed.2015.10.008

Algamal, Z. Y., & Mohammad Ali, H. T. (2017). An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression. *Electronic Journal of Applied Statistical Analysis*, *10*(1), 242–256. https://doi.org/10.1285/i20705948v10n1p242

Ali, M., & Aittokallio, T. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. In *Biophysical Reviews* (Vol. 11, Issue 1, pp. 31–39). Springer Verlag. https://doi.org/10.1007/s12551-018-0446-z

Brown, S. D., Warren, R. L., Gibb, E. A., Martin, S. D., Spinelli, J. J., Nelson, B. H., & Holt, R. A. (2014). Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Research*, *24*(5), 743–750. https://doi.org/10.1101/gr.165985.113

Burke, H. B. (1994). Artificial neural networks for cancer research: Outcome prediction. *Seminars in Surgical Oncology*, *10*(1), 73–79. https://doi.org/10.1002/ssu.2980100111

Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell, F. E., Marks, J. R., Winchester, D. P., & Bostwick, D. G. (1997). Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, *79*(4), 857–862. https://doi.org/10.1002/(SICI)1097-0142(19970215)79:4<857::AID-CNCR24>3.0.CO;2-Y

Chiu, Y. C., Chen, H. I. H., Zhang, T., Zhang, S., Gorthi, A., Wang, L. J., Huang, Y., & Chen, Y. (2019). Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Medical Genomics*, *12*(Suppl 1). https://doi.org/10.1186/s12920-018-0460-9

Chwala, N. v., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*(2), 321–357. https://doi.org/10.1002/eap.2043

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Ammad-Ud-Din, M., Hintsanen, P., Khan, S. A., Mpindi, J. P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., Collins, J. J., Gallahan, D., Singer, D., Saez-Rodriguez, J., … van Westen, G. J. P. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, *32*(12), 1202–1212. https://doi.org/10.1038/nbt.2877

Danaee, P., Ghaeini, R., & Hendrix, D. A. (2017). A deep learning approach for cancer detection and relevant gene identification. *Pacific Symposium on Biocomputing*, *0*(212679), 219–229. https://doi.org/10.1142/9789813207813\_0022

Daoud, M., & Mayo, M. (2019). A survey of neural network-based cancer prediction models from microarray data. In *Artificial Intelligence in Medicine* (Vol. 97, pp. 204–214). Elsevier B.V. https://doi.org/10.1016/j.artmed.2019.01.006

Domingues, B., Lopes, J., Soares, P., & Populo, H. (2018). Melanoma treatment in review. *ImmunoTargets and Therapy*, *Volume 7*, 35–49. https://doi.org/10.2147/itt.s134842

Geeleher, P., Cox, N. J., & Huang, R. S. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology*, *15*(3), R47. https://doi.org/10.1186/gb-2014-15-3-r47

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow - Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition (2019)* (2nd ed.). O’Reilly Media.

Ghosh, D., & Chinnaiyan, A. M. (2005). Classification and selection of biomarkers in genomic data using LASSO. *Journal of Biomedicine and Biotechnology*, *2005*(2), 147–154. https://doi.org/10.1155/JBB.2005.147

Gide, T. N., Quek, C., Menzies, A. M., Tasker, A. T., Shang, P., Holst, J., Madore, J., Lim, S. Y., Velickovic, R., Wongchenko, M., Yan, Y., Lo, S., Carlino, M. S., Guminski, A., Saw, R. P. M., Pang, A., McGuire, H. M., Palendira, U., Thompson, J. F., … Wilmott, J. S. (2019). Distinct Immune Cell Populations Define Response to Anti-PD-1 Monotherapy and Anti-PD-1/Anti-CTLA-4 Combined Therapy. *Cancer Cell*, *35*(2), 238-255.e6. https://doi.org/10.1016/j.ccell.2019.01.003

Hodi, F. S., O’Day, S. J., McDermott, D. F., Weber, R. W., Sosman, J. A., Haanen, J. B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J. C., Akerley, W., van den Eertwegh, A. J. M., Lutzky, J., Lorigan, P., Vaubel, J. M., Linette, G. P., Hogg, D., Ottensmeier, C. H., Lebbé, C., … Urba, W. J. (2010). Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. *New England Journal of Medicine*, *363*(8), 711–723. https://doi.org/10.1056/nejmoa1003466

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., Seja, E., Lomeli, S., Kong, X., Kelley, M. C., Sosman, J. A., Johnson, D. B., Ribas, A., & Lo, R. S. (2016). Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*, *165*(1), 35–44. https://doi.org/10.1016/j.cell.2016.02.065

Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H., & Margolin, A. A. (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pacific Symposium on Biocomputing*, 63–74. https://doi.org/10.1142/9789814583220\_0007

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. v., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. In *Computational and Structural Biotechnology Journal* (Vol. 13, pp. 8–17). Elsevier. https://doi.org/10.1016/j.csbj.2014.11.005

Larkin, J., Chiarion-Sileni, V., Gonzalez, R., Grob, J. J., Cowey, C. L., Lao, C. D., Schadendorf, D., Dummer, R., Smylie, M., Rutkowski, P., Ferrucci, P. F., Hill, A., Wagstaff, J., Carlino, M. S., Haanen, J. B., Maio, M., Marquez-Rodas, I., McArthur, G. A., Ascierto, P. A., … Wolchok, J. D. (2015). Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. *New England Journal of Medicine*, *373*(1), 23–34. https://doi.org/10.1056/nejmoa1504030

Li, L., & Talwalkar, A. (2019). Random search and reproducibility for neural architecture search. *ArXiv*, 1–20.

Liang, S., Ma, A., Yang, S., Wang, Y., & Ma, Q. (2018). A Review of Matched-pairs Feature Selection Methods for Gene Expression Data Analysis. In *Computational and Structural Biotechnology Journal* (Vol. 16, Issue 1, pp. 88–97). Oxford University Press. https://doi.org/10.1016/j.csbj.2018.02.005

Liang, Y., Liu, C., Luan, X. Z., Leung, K. S., Chan, T. M., Xu, Z. ben, & Zhang, H. (2013). Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. *BMC Bioinformatics*, *14*(1), 1–12. https://doi.org/10.1186/1471-2105-14-198

Lin, C., Jain, S., Kim, H., & Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Research*, *45*(17), e156. https://doi.org/10.1093/nar/gkx681

Liu, D., Jenkins, R. W., & Sullivan, R. J. (2019). Mechanisms of Resistance to Immune Checkpoint Blockade. In *American Journal of Clinical Dermatology* (Vol. 20, Issue 1, pp. 41–54). Springer International Publishing. https://doi.org/10.1007/s40257-018-0389-y

Liu, D., Schilling, B., Liu, D., Sucker, A., Livingstone, E., Jerby-Amon, L., Zimmer, L., Gutzmer, R., Satzger, I., Loquai, C., Grabbe, S., Vokes, N., Margolis, C. A., Conway, J., He, M. X., Elmarakeby, H., Dietlein, F., Miao, D., Tracy, A., … Schadendorf, D. (2019). Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nature Medicine*, *25*(12), 1916–1927. https://doi.org/10.1038/s41591-019-0654-5

Ma, S., & Huang, J. (2008). Penalized feature selection and classification in bioinformatics. In *Briefings in Bioinformatics* (Vol. 9, Issue 5, pp. 392–403). Oxford University Press. https://doi.org/10.1093/bib/bbn027

Niu, P. H., Zhao, L. L., Wu, H. L., Zhao, D. B., & Chen, Y. T. (2020). Artificial intelligence in gastric cancer: Application and future perspectives. In *World Journal of Gastroenterology* (Vol. 26, Issue 36, pp. 5408–5419). Baishideng Publishing Group Co. https://doi.org/10.3748/wjg.v26.i36.5408

Olson, M., Wyner, A. J., & Berk, R. (2018). Modern neural networks generalize on small data sets. *Advances in Neural Information Processing Systems*, *2018-Decem*, 3619–3628.

Perez-Riverol, Y., Kuhn, M., Vizcaíno, J. A., Hitz, M. P., & Audain, E. (2017). Accurate and fast feature selection workflow for high-dimensional omics data. *PLoS ONE*, *12*(12). https://doi.org/10.1371/journal.pone.0189875

Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2019). Dr.VAE: Improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics*, *35*(19), 3743–3751. https://doi.org/10.1093/bioinformatics/btz158

Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., Hodi, F. S., Martín-Algarra, S., Mandal, R., Sharfman, W. H., Bhatia, S., Hwu, W. J., Gajewski, T. F., Slingluff, C. L., Chowell, D., Kendall, S. M., Chang, H., Shah, R., Kuo, F., … Chan, T. A. (2017). Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*, *171*(4), 934-949.e15. https://doi.org/10.1016/j.cell.2017.09.028

Robert, C., Schachter, J., Long, G. v., Arance, A., Grob, J. J., Mortier, L., Daud, A., Carlino, M. S., McNeil, C., Lotem, M., Larkin, J., Lorigan, P., Neyns, B., Blank, C. U., Hamid, O., Mateus, C., Shapira-Frommer, R., Kosh, M., Zhou, H., … Ribas, A. (2015). Pembrolizumab versus Ipilimumab in Advanced Melanoma. *New England Journal of Medicine*, *372*(26), 2521–2532. https://doi.org/10.1056/nejmoa1503093

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., & Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*, *160*(1–2), 48–61. https://doi.org/10.1016/j.cell.2014.12.033

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. In *Bioinformatics* (Vol. 23, Issue 19, pp. 2507–2517). Bioinformatics. https://doi.org/10.1093/bioinformatics/btm344

Sakellaropoulos, T., Vougas, K., Narang, S., Koinis, F., Kotsinas, A., Polyzos, A., Moss, T. J., Piha-Paul, S., Zhou, H., Kardala, E., Damianidou, E., Alexopoulos, L. G., Aifantis, I., Townsend, P. A., Panayiotidis, M. I., Sfikakis, P., Bartek, J., Fitzgerald, R. C., Thanos, D., … Gorgoulis, V. G. (2019). A Deep Learning Framework for Predicting Response to Therapy in Cancer. *Cell Reports*, *29*(11), 3367-3373.e4. https://doi.org/10.1016/j.celrep.2019.11.017

Spranger, S., Bao, R., & Gajewski, T. F. (2015). Melanoma-intrinsic β-catenin signalling prevents anti-tumour immunity. *Nature*, *523*(7559), 231–235. https://doi.org/10.1038/nature14404

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Tan, J., Hammond, J. H., Hogan, D. A., & Greene, C. S. (2016). ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions. *MSystems*, *1*(1). https://doi.org/10.1128/msystems.00025-15

Torang, A., Gupta, P., & Klinke, D. J. (2019). An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets. *BMC Bioinformatics*, *20*(1), 433. https://doi.org/10.1186/s12859-019-2994-z

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, *203*(3–4), 312–318. https://doi.org/10.1016/j.ecolmodel.2006.11.033

van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Geukes, M. H., Goldinger, S. M., Utikal, J., Hassel, J. C., Weide, B., Kaehler, K. C., Loquai, C., Mohr, P., & Gutzmer, R. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*, *350*(6257), 207–211. https://doi.org/10.1126/science.aad0095.Genomic

Way, G. P., & Greene, C. S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pacific Symposium on Biocomputing*, *0*(212669), 80–95. https://doi.org/10.1142/9789813235533\_0008

Witten, D. M., & Tibshirani, R. (2010). Survival analysis with high-dimensional covariates Daniela. *Stat Methods Med Res.*, *19*(1), 29–51. https://doi.org/10.1177/0962280209105024.Survival

Wolchok, J. D., Chiarion-Sileni, V., Gonzalez, R., Rutkowski, P., Grob, J.-J., Cowey, C. L., Lao, C. D., Wagstaff, J., Schadendorf, D., Ferrucci, P. F., Smylie, M., Dummer, R., Hill, A., Hogg, D., Haanen, J., Carlino, M. S., Bechter, O., Maio, M., Marquez-Rodas, I., … Larkin, J. (2017). Overall Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *New England Journal of Medicine*, *377*(14), 1345–1356. https://doi.org/10.1056/nejmoa1709684

Wu, S., Jiang, H., Shen, H., & Yang, Z. (2018). Gene selection in cancer classification using sparse logistic regression with L1/2 regularization. *Applied Sciences (Switzerland)*, *8*(9), 1–12. https://doi.org/10.3390/app8091569

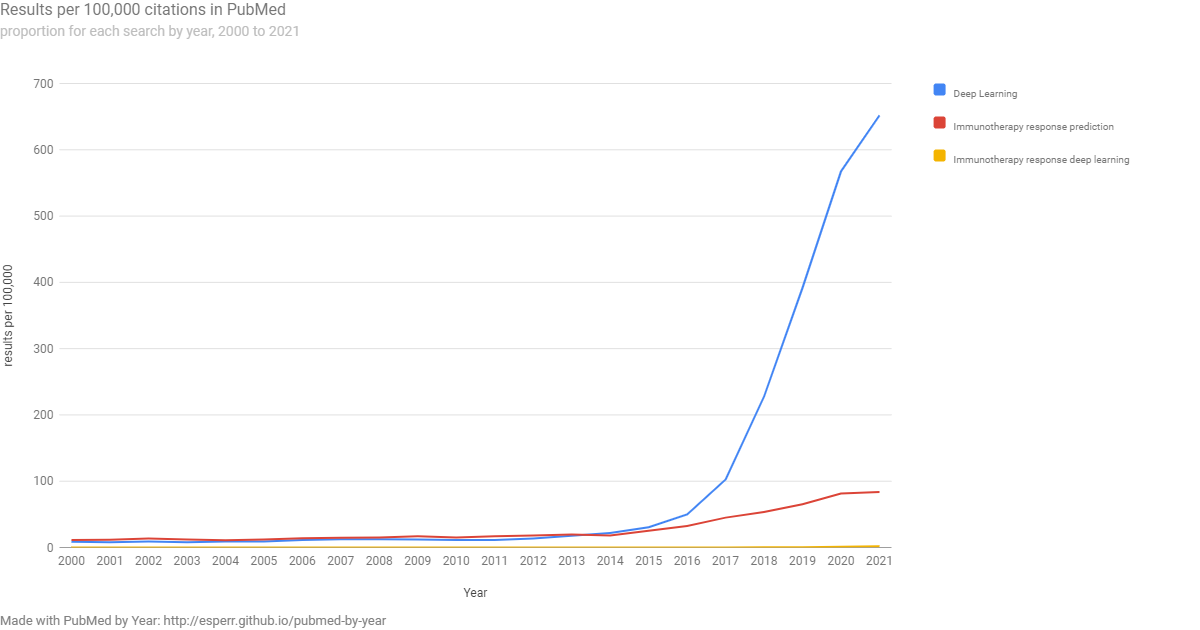
Yousef, M., Kumar, A., & Bakir-Gungor, B. (2021). Application of biological domain knowledge based feature selection on gene expression data. In *Entropy* (Vol. 23, Issue 1, pp. 1–15). MDPI AG. https://doi.org/10.3390/e23010002

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

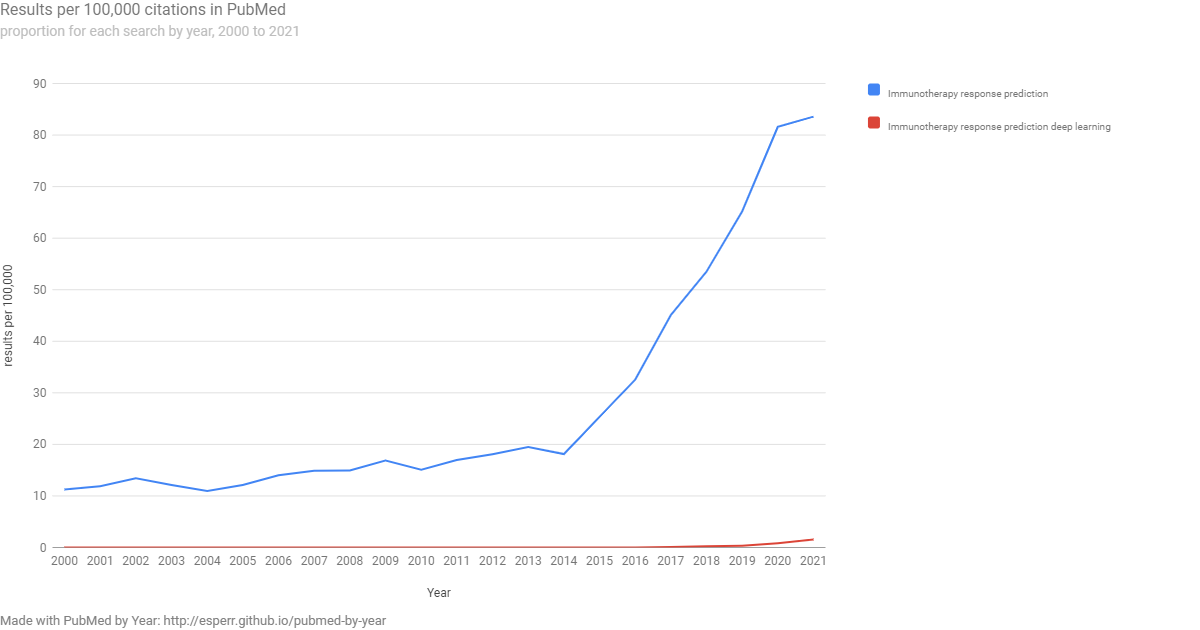
# Appendix

Appendix 1: PubMed publications[[5]](#footnote-5) of deep learning and immunotherapy response prediction related articles from 2000 to 2021

A: Deep learning (blue) vs. Immunotherapy response prediction (red) vs. Immunotherapy response deep learning (yellow)



B: Immunotherapy response prediction (blue) vs. Immunotherapy response deep learning (red)



Appendix 2: Binary prediction on Liu et al. dataset with PFS as endpoint

Predictions on binary classification based on PFS for therapy response of patients from the Liu cohort are presented. For each of the five training/test splits, the calculated metrics and the best selected tuning parameters are presented. After finding the best parameters on the 80% training data of each split, the metrics on the 20% test data are presented. The mean and standard deviation of each metric is calculated after 30 runs with the best model with a confusion matrix example of the last validation run.

|  |  |  |  |
| --- | --- | --- | --- |
| **Split** | **Best Params** | **Metrics** | **Confusion Matrix** |
| 0 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 3  Number hidden: 1  Optimizer: SGD | **Accuracy 0.71 +/-(0.06)**  Sensitivity 0.48 +/-(0.09)  Specificity 0.85 +/-(0.07)  PPV 0.66 +/-(0.11)  NPV 0.73 +/-(0.04) |  |
| 1 | Dropout rate:0.2  L1 reg:0.0  L2 reg:0.2  Learning rate:0.01  Number neurons: 5  Number hidden: 3  Optimizer: SGD | **Accuracy 0.52 +/-(0.04)**  Sensitivity 0.33 +/-(0.00)  Specificity 0.63 +/-(0.07)  PPV 0.35 +/-(0.05)  NPV 0.61 +/-(0.03) |  |
| 2 | Dropout rate:0.4  L1 reg:0.0  L2 reg:0.4  L earning rate:0.01  Number neurons: 3  Number hidden: 3  Optimizer: SGD | **Accuracy 0.53 +/-(0.02)**  Sensitivity 0.21 +/-(0.04)  Specificity 0.72 +/-(0.02)  PPV 0.31 +/-(0.05)  NPV 0.60 +/-(0.02) |  |
| 3 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 3  Number hidden: 1  Optimizer: Adam | **Accuracy 0.65 +/-(0.07)**  Sensitivity 0.46 +/-(0.19)  Specificity 0.77 +/-(0.12)  PPV 0.56 +/-(0.14)  NPV 0.71 +/-(0.07) |  |
| 4 | Dropout rate:0.4  L1 reg:0.0  L2 reg:0.4  Learning rate:0.01  Number neurons: 5  Number hidden: 3  Optimizer: SGD | **Accuracy 0.59 +/-(0.03)**  Sensitivity 0.22 +/-(0.00)  Specificity 0.81 +/-(0.05)  PPV 0.42 +/-(0.08)  NPV 0.63 +/-(0.02) |  |

Appendix 3: Continuous prediction on Liu et al. dataset with OS as endpoint

Predictions on continuous classification based on OS for therapy response of patients from the Liu cohort are presented. For each of the five training/test splits, the calculated metrics and the best selected tuning parameters are presented. After finding the best parameters on the 80% training data of each split, the metrics on the 20% test data are presented. The mean and standard deviation of each metric is calculated after 30 runs with the best model.

|  |  |  |
| --- | --- | --- |
| **Split** | **Best Params** | **Metrics** |
| 0 | Dropout rate:0.4  L1 reg:0.2  L2 reg:0.4  Learning rate:0.01  Number neurons: 10  Number hidden: 2  Optimizer: Adam | MAE 250.95 +/-(29.03)  RMSE 320.10 +/-(31.42)  **R^2 -0.28 +/-(0.29)** |
| 1 | Dropout rate:0.0  L1 reg:0.4  L2 reg:0.2  Learning rate:0.01  Number neurons: 10  Number hidden: 1  Optimizer: Adam | MAE 598.49 +/-(127.48)  RMSE 830.38 +/-(209.81)  **R^2 -3.00 +/-(2.06)** |
| 2 | Dropout rate:0.0  L1 reg:0.2  L2 reg:0.4  Learning rate:0.01  Number neurons: 10  Number hidden: 2  Optimizer: Adam | MAE 438.81 +/-(58.63)  RMSE 570.52 +/-(64.58)  **R^2 -1.25 +/-(0.52)** |
| 3 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.4  Learning rate:0.01  Number neurons: 5  Number hidden: 1  Optimizer: Adam | MAE 387.51 +/-(67.52)  RMSE 505.26 +/-(86.02)  **R^2 -0.50 +/-(0.52)** |
| 4 | Dropout rate:0.2  L1 reg:0.4  L2 reg:0.2  Learning rate:0.01  Number neurons: 10  Number hidden: 2  Optimizer: Adam | MAE 395.81 +/-(37.84)  RMSE 555.09 +/-(105.92)  **R^2 -0.79 +/-(0.74)** |

Appendix 4: Continuous prediction on Liu et al. dataset with PFS as endpoint

Predictions on continuous classification based on PFS for therapy response of patients from the Liu cohort are presented. For each of the five training/test splits, the calculated metrics and the best selected tuning parameters are presented. After finding the best parameters on the 80% training data of each split, the metrics on the 20% test data are presented. The mean and standard deviation of each metric is calculated after 30 runs with the best model.

|  |  |  |
| --- | --- | --- |
| **Split** | **Best Params** | **Metrics** |
| 0 | Dropout rate: 0.0  L1 reg: 0.2  L2 reg: 0.0  Learning rate: 0.01  Number neurons: 10  Number hidden: 2  Used optimizer: Adam | MAE 278.83 +/-(10.75)  RMSE 325.94 +/-(14.58)  **R^2 -0.31 +/-(0.12)** |
| 1 | Dropout rate: 0.0  L1 reg: 0.4  L2 reg: 0.2  Learning rate: 0.01  Number neurons: 10  Number hidden: 3  Used optimizer: Adam | MAE 494.26 +/-(57.33)  RMSE 689.50 +/-(78.32)  **R^2 -0.89 +/-(0.45)** |
| 2 | Dropout rate: 0.0  L1 reg: 0.2  L2 reg: 0.0  Learning rate: 0.01  Number neurons:10  Number hidden: 2  Used optimizer: Adam | MAE 344.83 +/-(30.44)  RMSE 460.07 +/-(35.12)  **R^2 -0.57 +/-(0.25)** |
| 3 | Dropout rate: 0.0  L1 reg: 0.0  L2 reg: 0.2  Learning rate: 0.01  Number neurons: 10  Number hidden: 2  Used optimizer: Adam | MAE 348.03 +/-(20.53)  RMSE 475.10 +/-(27.49)  **R^2 -0.05 +/-(0.12)** |
| 4 | Dropout rate: 0.0  L1 reg: 0.0  L2 reg: 0.4  Learning rate: 0.01  Number neurons: 10  Number hidden: 3  Used optimizer: Adam | MAE 278.33 +/-(33.70)  RMSE 365.61 +/-(73.54)  **R^2 -0.30 +/-(0.62)** |

Appendix 5: Whole list of occurrences of top 50 selected features after RFE with PFS as endpoint

Presented are the selected features by RFE algorithm based on SVC and their occurrences in each of the five different train/test splits.

|  |  |  |
| --- | --- | --- |
| **Item** | **Occurrences** | **Present in** |
| HIKESHI | 5 | Split 0, Split 1, Split 2, Split 3, Split 4 |
| LIN28A | 5 | Split 0, Split 1, Split 2, Split 3, Split 4 |
| EMP1 | 4 | Split 0, Split 2, Split 3, Split 4 |
| HSH2D | 4 | Split 1, Split 2, Split 3, Split 4 |
| NHEJ1 | 4 | Split 0, Split 1, Split 2, Split 3 |
| PROC | 4 | Split 1, Split 2, Split 3, Split 4 |
| AGXT | 3 | Split 0, Split 2, Split 3 |
| ALDH7A1 | 3 | Split 0, Split 1, Split 4 |
| ARHGAP31 | 3 | Split 0, Split 1, Split 2 |
| DNAJC9 | 3 | Split 1, Split 2, Split 3 |
| IGLV2.11 | 3 | Split 0, Split 1, Split 4 |
| SCN8A | 3 | Split 2, Split 3, Split 4 |
| TXK | 3 | Split 0, Split 3, Split 4 |
| C3orf22 | 2 | Split 1, Split 4 |
| CCDC173 | 2 | Split 0, Split 3 |
| CPB1 | 2 | Split 3, Split 4 |
| DBNDD2 | 2 | Split 3, Split 4 |
| EMP2 | 2 | Split 2, Split 3 |
| FBXO32 | 2 | Split 1, Split 3 |
| HPDL | 2 | Split 0, Split 2 |
| IDH1 | 2 | Split 2, Split 3 |
| KHDRBS3 | 2 | Split 0, Split 2 |
| KYAT3 | 2 | Split 0, Split 1 |
| NOP2 | 2 | Split 1, Split 2 |
| OLR1 | 2 | Split 3, Split 4 |
| PKD2L2 | 2 | Split 2, Split 4 |
| POU2AF1 | 2 | Split 0, Split 1 |
| PPARG | 2 | Split 2, Split 3 |
| RDH16 | 2 | Split 2, Split 4 |
| RHOB | 2 | Split 2, Split 3 |
| RNA5S9 | 2 | Split 0, Split 1 |
| RPL23AP75 | 2 | Split 1, Split 4 |
| RUNX3 | 2 | Split 0, Split 3 |
| TAC3 | 2 | Split 2, Split 3 |
| TCP10L | 2 | Split 0, Split 1 |

Appendix 6: Analysis of best six selected features by SVM as potential marker for ICB response

Presented are function and pathway[[6]](#footnote-6) of the six selected features (table 7). Further analysis for value as a potential marker of response to ICB is based on <https://doc.hornlab.org/shiny/sc_Tirosh/> to show gene expression in malignant and non-malignant cells (tumor microenvironment) and <https://doc.hornlab.org/shiny/cru337phenotime/> for survival analysis using the Kaplan-Meier model based on high or low gene expression and Cox proportional hazards model for multivariate survival analysis involving the 6 clinical features.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gene | Function | Pathway | Tumor Microenvironment | Kaplan-Meier survival analysis of PFS (only Pembro, only Nivo) | Multivariate Cox Proportional Hazards model with 6 clinical features |
| *HIKESHI* | nuclear import carrier | cellular response to heat stress, cellular senescence | in 7/15 higher significant higher in malignant cells | significant survival **p=0.0078247**  (p=0.11093, **p=0.057524**)  Cut-off =  20 TPM | Hazard ratio =1.01156  p=0.13308 |
| *LIN28A* | inhibits processing and regulates translation of mRNAs that control developmental timing, pluripotency, and metabolism | translational control, embryonic and induced pluripotent stem cell differentiation | in 10/17 significant higher in non-malignant | significant survival  **p=0.*038422***  (**p=0.0027303**, p=0.98852)  Cut-off =  0.01 TPM | Hazard ratio =0.81997  p=0.15318 |
| *EMP1* | epithelial membrane protein 1, tumor associated membrane protein, expressed in mammary tumor and T cell lymphoma | NA | in 10/17 significant higher in malignant | n.s. p=0.82247  (p=0.62578, p=0.7706)  Cut-off =  200 TPM | Hazard ratio =0.99958  p=0.23560 |
| *HSH2D* | May be a modulator of the apoptotic response through its ability to affect mitochondrial stability | NA | in 10/17 significant higher in non-malignant cells | significant survival  **p=0.033845**  (p=0.54542, **p=0.015381)**  Cut-off=  1 TPM | Hazard ratio = 0.88809  **p=0.00288 \*\*** |
| *NHEJ1* | DNA repair protein | DNA double-strand break repair | in 8/15 significant higher in malignant cells | n.s. p=0.80101  (p=0.76346, p=0.65939)  Cut-off=  15TPM | Hazard ratio = 1.00655  p=0.62938 |
| *PROC* | protein C is a vitamin K-dependent serine protease that regulates blood | formation of fibrin clot, gamma carboxylation, hypusine formation and arylsulfatase activation | not expressed in Tirosh | n.s. p=0.57588  (p=0.88114, p=0.6663)  Cut-off =  0.3 TPM | Hazard ratio =0.59157  p=0.06093 |

Appendix 7: Whole list of occurrences of top 50 selected features after RFE with OS as endpoint

Presented are the selected features by RFE algorithm based on SVC and their occurrences in each of the five different train/test splits.

|  |  |  |
| --- | --- | --- |
| **Item** | **Occurrences** | **Present in** |
| ARHGAP6 | 4 | Split 0, Split 1, Split 3, Split 4 |
| RPL23AP75 | 4 | Split 0, Split 1, Split 2, Split 4 |
| AHSP | 3 | Split 2, Split 3, Split 4 |
| ANKRD20A4 | 3 | Split 2, Split 3, Split 4 |
| ASB2 | 3 | Split 0, Split 3, Split 4 |
| FOXD3.AS1 | 3 | Split 0, Split 2, Split 3 |
| HAGLR | 3 | Split 0, Split 1, Split 3 |
| NUP107 | 3 | Split 0, Split 1, Split 4 |
| C12orf29 | 2 | Split 0, Split 3 |
| CA8 | 2 | Split 2, Split 3 |
| CPNE7 | 2 | Split 0, Split 4 |
| DYNC2H1 | 2 | Split 0, Split 4 |
| EFCAB14P1 | 2 | Split 2, Split 3 |
| FAM57A | 2 | Split 0, Split 4 |
| FCMR | 2 | Split 1, Split 4 |
| GAPDHP51 | 2 | Split 0, Split 4 |
| GLULP4 | 2 | Split 1, Split 3 |
| IGHV3.53 | 2 | Split 1, Split 2 |
| IGLV2.11 | 2 | Split 0, Split 1 |
| IGLV2.5 | 2 | Split 2, Split 3 |
| LINC00167 | 2 | Split 1, Split 2 |
| LINC00526 | 2 | Split 2, Split 3 |
| LINC00680 | 2 | Split 1, Split 3 |
| MC1R | 2 | Split 3, Split 4 |
| MMP24OS | 2 | Split 0, Split 1 |
| NT5C | 2 | Split 1, Split 4 |
| PPARG | 2 | Split 1, Split 2 |
| PRDM7 | 2 | Split 1, Split 3 |
| PTPN13 | 2 | Split 1, Split 2 |
| SPATA18 | 2 | Split 2, Split 4 |
| TAX1BP3 | 2 | Split 1, Split 2 |
| TMEM221 | 2 | Split 0, Split 4 |
| WASH7P | 2 | Split 0, Split 3 |
| ZFAT | 2 | Split 0, Split 4 |
| ZNRF2P1 | 2 | Split 0, Split 1 |

Appendix 8: Binary prediction on Liu et al. dataset after feature selection with OS as endpoint

Predictions on binary classification after feature selection of 10 different genes for the Liu cohort are presented. For each of the 5 training/test splits, the calculated metrics and the best selected tuning parameters are presented. After finding the best parameters on the 80% training data of each split, the metrics on the 20% test data are presented. The mean and standard deviation of each metric is calculated after 30 runs with the best model.

|  |  |  |
| --- | --- | --- |
| **Split** | **Best Params** | **Metrics** |
| 0 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 5  Number hidden: 1  Optimizer: Adam | **Accuracy 0.79 +/-(0.03)**  Sensitivity 0.40 +/-(0.07)  Specificity 0.99 +/-(0.03)  PPV 0.94 +/-(0.11)  NPV 0.77 +/-(0.02) |
| 1 | Dropout rate:0.4  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 10  Number hidden: 2  Optimizer: SGD | **Accuracy 0.49 +/-(0.05)**  Sensitivity 0.35 +/-(0.11)  Specificity 0.56 +/-(0.07)  PPV 0.29 +/-(0.07)  NPV 0.64 +/-(0.04) |
| 2 | Dropout rate:0.2  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 5  Number hidden: 3  Optimizer: SGD | **Accuracy 0.67 +/-(0.04)**  Sensitivity 0.41 +/-(0.12)  Specificity 0.80 +/-(0.05)  PPV 0.51 +/-(0.07)  NPV 0.73 +/-(0.04) |
| 3 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.0  Learning rate:0.01  Number neurons: 10  Number hidden: 1  Optimizer: Adam | **Accuracy 0.62 +/-(0.04)**  Sensitivity 0.17 +/-(0.06)  Specificity 0.84 +/-(0.06)  PPV 0.36 +/-(0.12)  NPV 0.67 +/-(0.02) |
| 4 | Dropout rate:0.0  L1 reg:0.0  L2 reg:0.2  Learning rate:0.01  Number neurons: 3  Number hidden: 1  Optimizer: SGD | **Accuracy 0.71 +/-(0.00)**  Sensitivity 0.38 +/-(0.00)  Specificity 0.88 +/-(0.00)  PPV 0.60 +/-(0.00)  NPV 0.74 +/-(0.00) |

# Selbstständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung angegebener Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann. Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Leipzig, 16.04.2021

Christina Kuhn

1. A way to measure how well a cancer patient responds to treatment. The types of responses a patient can have are a complete response (CR), a partial response (PR), progressive disease (PD), and stable disease (SD). [↑](#footnote-ref-1)
2. Some of the patients received ipilimumab as initial treatment, but unfortunately it was not specified which sample. [↑](#footnote-ref-2)
3. The study reported that a subgroup of (17) patients had a previous MAPKi, but the information about which patients was not disclosed. [↑](#footnote-ref-3)
4. https://www.kaggle.com/c/boston-housing/leaderboard [↑](#footnote-ref-4)
5. Generated with https://esperr.github.io/pubmed-by-year/ [↑](#footnote-ref-5)
6. Information about function and pathways were extracted from https://www.genecards.org/. [↑](#footnote-ref-6)