Feature-based Response Prediction to Immunotherapy of late-stage Melanoma Patients Using PET/MR Imaging

Annika Liebgott*, Sergios Gatidis†, Viet Chau Vu†, Tobias Haueise*, Konstantin Nikolaou† and Bin Yang*
*Institute of Signal Processing and System Theory, University of Stuttgart, Germany
† Department of Diagnostic and Interventional Radiology, University Hospital of Tübingen, Germany
Email: annika.liebgott@iss.uni-stuttgart.de

***Abstract*—The treatment of malignant melanoma with immunotherapy is a promising approach to treat advanced stages of the disease. However, the treatment can cause serious side effects and not every patient responds to it. This means, crucial time may be wasted on an ineffective treatment. Assessment of the possible therapy response is hence an important research issue. The research presented in this study focuses on the investigation of the potential of medical imaging and machine learning to solve this task. To this end, we extracted image features from multi-modal images and trained a classifier to differentiate non-responsive patients from responsive ones.**

***Index Terms*—Support Vector Machine, Random Forest, PET/MR imaging, Therapy Response Prediction**

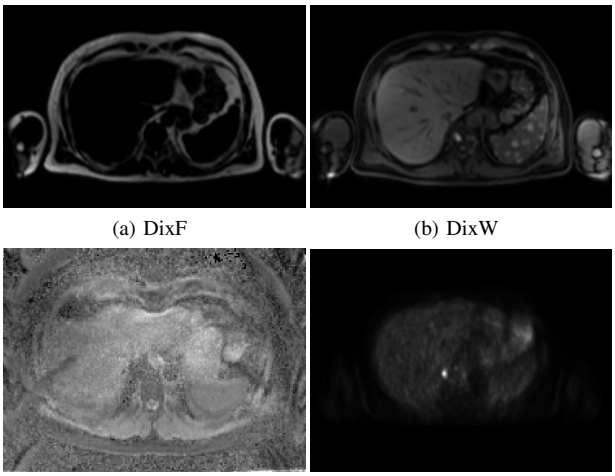
I. Motivation

Malignant melanoma has shown increasing worldwide incidence over the last decades [1]–[3]. Although the prognosis is very good when caught early (up to 99% 5-year survival rate), it is a very aggressive type of cancer. Once cancerous cells have advanced beyond the skin barrier, they spread quickly throughout the whole body. At this point, conventional clinical intervention (i.e. tumor resection, chemo or radiation therapy) often fails to eliminate all tumor cells, resulting in high recurrance rates and ultimately low survival rates for advanced stage melanoma patients ($< 10\%$ for stage IV melanoma).

In recent years, however, immunotherapy has been sucessfully used to treat cancer (including malignant melanoma) in cases, where conventional therapy is likely to fail, leading to James P. Allison and Tasuko Honjo winning the Nobel Prize in Medicine for their pioneer work on this topic in 2018 [4]. The basic idea behind all kinds of cancer immunotherapy is stimulating the patient's immune system to identify and fight malignant cells. One type of immunotherapy, which was also used in the clinical study this work is based on, utilizes so-called immune checkpoint inhibitors. By blocking certain receptors on t-cells, these special antibodies prevent cancer cells from using the receptors to disguise themselves as healthy cells and hence avoid being attacked by the immune system.

Therapy with immune checkpoint inhibitors has lead to a significant improvement in patient outcome, especially when combined with e.g. radiation therapy. The treatment has shown the potential to slow down, stop or completely reverse the progress of the disease [5]. While the positive effects observed in a significant number of patients are promising, there are also some issues which often lead to immunotherapy being not the first choice of treatment. For instance, the stimulation of the immune system with checkpoint inhibitors can inflict severe side effects. The main concern, however, is that only part of the patients respond to the treatment and in other cases the disease continues to progress, in the worst case wasting crucial time with an ineffective therapy.

This research was conducted with the generous support of Vector Stiftung.



(a) DixF (b) DixW

(c) ADC (d) PET

Fig. 1: Abdominal example images of the 4 different imaging modalities used in this study.

Hence, a major issue is finding out what differentiates responsive from non-responsive patients, as well as trying to predict the individual therapy response potential. Our research focuses on using hybrid positron emission tomograpy (PET) and magnetic resonance imaging (MRI) combined with machine learning (ML) to predict therapy response by analyzing images acquired at multiple time instances during treatment. In the past years, a couple of related studies have been published proposing to use ML approaches. To the best of our knowledge, none of these studies used a similar approach to ours. In 2019, Trebeschi et al. proposed a ML-based approach with computed tomography (CT) images [6], whereas we use combined PET/MR imaging. We decided to use these images

in our study, because the MR images visualize different tissues more detailed than CT and the PET images add information about the metabolic activity within the body, capturing for example possible locations of metastases. Sun et al. also published a study involving classification of CT images and in addition combined them with other prior knowledge (RNA sequencing) [7]. Other research including machine learning did not use radiological images as input data, but other clinical examinations (e.g. h&e stain [8], genetic analyses [9]–[11] or blood tests [12]).

As a first step, we focus on identifying non-responsive patients (i.e. disease progresses further under therapy) by a radiomics-based ML approach. Radiomics [13] is a term commonly used to describe research in the radiological scientific comunity which involves the extraction of mathematical representations from images to solve a problem, often combined with ML. For this study, we extracted features based on gray-level variations from 72 multi-modal images of the liver, spleen and vertebrae of 24 late-stage melanoma patients treated with immunotherapy and trained support vector machines (SVM) and random forests (RF) to differentiate between non-responsive and responsive patients. In addition, we calculated 13 metrics directly derived from the radiological imaging modalities, which are commonly used in clinical research, and compared classifier performances between them and the radiomics features we extracted. This work is intended as a first feasibility study, investigating feature-based ML approaches as they are on the one hand better interpretable by clinicians than deep learning methods and on the other hand easier to train on our comparably small dataset.

II. Dataset

Our dataset consists of PET/MR images from 24 patients diagnosed with advanced malignant melanoma. The imaging data was acquired at three times over the course of treatment: prior to treatment start, two weeks as well as two months after starting immunotherapy. For every patient and time instance, MRI and PET examinations were performed, resulting in the acquisition of the following images:

- fat-weighted Dixon MR images (DixF)
- water-weighted Dixon MR images (DixW)
- apparent diffusion coefficient maps (ADC maps)
- FDG-18 PET images

Fig. 1 shows the 4 acquired image types on the example of abdominal images of one patient. The optical difference between the DixF and DixW images is caused by varying spin precession rates of atoms in fat and water molecules. By changing parameters in the MR acquisition sequence, different tissue types in the human body can be emphasized, i.e. fat is bright while water is dark in the DixF images and the opposite holds for DixW images. ADC maps are used in clinical diagnostics to differentiate between normal and unhealthy structures in the body by merging diffusion weighted MR images aquired with different diffusion weights. Diffusion weighted imaging is a special technique of T2 weighted MRI exploiting the random motion of water molecules. As the diffusion is influenced by interactions of the molecules with obstacles (e.g. cell membranes), different tissues result in varying diffusion. The intuition behind using ADC maps is that it eliminates the T2 weighting that is otherwise inherent to the acquisition of a diffusion weighted image, leading to a less disturbed view of the diffusion. PET images are used to display the metabolic activity throughout the body. To this end, a positron-emitting tracer is administered to the patient. A particularly popular tracer for clinical PET scans is the radioisotope fluorine-18 synthesized into fludeoxyglucose (FDG-18), which was also used in our study. The tracer then travels through the body and releases positrons upon decay, which create gamma radiation when they encounter electrons and annihilate. A computer calculates a three-dimensional image based on the concentration of gamma particles, which is measured by a detector ring located around the patient. A higher concentration indicates an area of high metabolic activity. Since tumor cells are normally growing fast and hence exhibit an increased metabolic acitivity compared to healthy cells, they consume more of the tracer and hence appear brighter in the PET image. The glucose uptake in FDG-18 PET is measured in standardized uptake values (SUV).

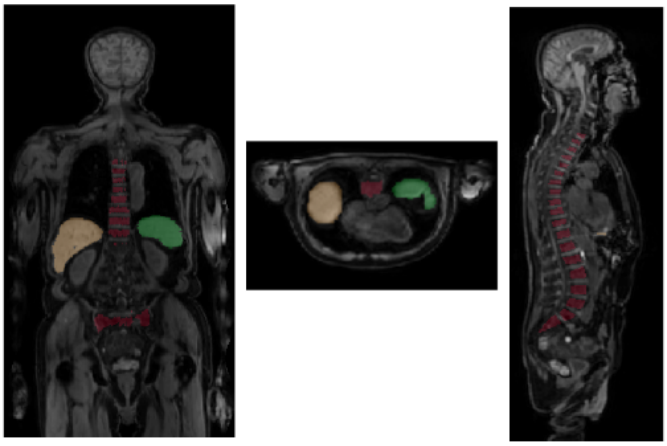


Fig. 2: Example of the segmantations used to extract the organs of interest (liver, spleen and vertebrae).

The images have been annotated by doctors according to the patients' response to the treatment. 11 of the patients exhibited a progressing disease (PD) under immunotherapy, 4 responded completely (i.e. the cancer vanished) and the rest showed mixed responses. As we intended to identify the PD patients, we combined all non-PD patients into one class, leading to a binary classification task.

As our cohort of patients is relatively small and we did not want individual physiological traits to influence our results too much, we only used the liver, spleen and vertebrae as input for our classifiers (see Fig. 2). These organs have been selected for two reasons: Firstly, they exhibit limited variability between individuals regarding their shape, so the risk of accidental correlations between the patients' physiology and the therapy response labels is only minimal. Secondly, they have proven to be closely related to the immune system, so they are likely

TABLE I: Overview of the feature extraction algorithms from ImFEATbox we used in our study.

feature group	# features
intensity	7
histogram	6
gradient	81
gray-level co-occurence matrix	672
gray-level run length	44
total # features	*810*

to be affected by immunotherapy or exhibit characteristics related to how susceptible a patient is to immunotherapy. The segmentation of the organs has been performed by experienced radiologists on the DixW images. The resulting masks have then been transferred to the DixF images as well as their corresponding PET images and ADC maps by resampling them to the respective coordinate systems.

III. Methods

A. Modality features

In our initial experiments, we used only features which can be directly derived from the acquired imaging modalities and are commonly used in clinical research. For each organ we calculated the following metrics:

- Mean ADC value
- Mean SUV
- Maximum SUV
- Fat percentage
- Organ volume

As the volume of the vertebrae and the fat percentage of the spleen should not vary between acquisitions, these two metrics were not calculated, leading to a total of $5{+}4{+}4 = 13$ features. They were calculated for each examination, i.e. three times for each of the 24 patients, resulting in 72 samples. The dataset consisting of these features is denoted as $\mathcal{D}_{\mathrm{M}}$ in the following sections.

B. Texture features

For further experiments, we extracted 810 features based on gray-level variation from each organ using the MATLAB toolbox ImFEATbox [14]. The used feature extraction algorithms and the number of resulting features per algorithm are given in Table I. As the toolbox is designed to extract features from 2D images, we calculated the mean of each feature per organ to have the features describe the charateristic of the whole organ rather than smaller sections captured by a transversal 2D slice. We extracted the features from DixF and DixW images, PET images and the ADC maps for all three organs, leading to a total number of $810 \cdot 4 \cdot 3 = 3240$ features per examination. As feature-based methods are often refered to as "radiomics" [13] in the radiological scientific community, the resulting dataset is denoted as $\mathcal{D}_{\mathrm{R}}$.

C. Feature reduction and selection

After feature extraction, we employed either principal component analysis (PCA, [15]), a feature transform method, or sequential forward floating selection (SFFS, [16]), a feature selection technique, to our datasets to reduce the feature dimensionality and avoid overfitting.

In addition to these well-known methods, we investigated the effect of two novel feature selection techniques based on deep learning. The first approach is an **attention-based feature selection (AFS)** [17]. AFS consists of two modules, the attention module and the learning module. Features are first compressed into a lower-dimensional representation to eliminate noise, outliers and redundancy. Afterwards, every feature is assigned a shallow neural network to determine its probability of selection. The probabilities of each feature build an attention matrix, which is then multiplied with the original features to weight them with their importance. The weighted feature matrix is used as input for the learning module, which can be any kind of neural network. During training, the attention matrix is iteratively adjusted until the final attention weight is used to select the most important features.

The second technique called **concrete autoencoder (CAE)** [18] uses a concrete selector layer with a number of nodes corresponding to the number of features to be selected in the encoding part of an autoencoder. This layer is based on concrete random variables and selects stochastic linear combinations of input features, which converge to a discrete set of features by producing a continuous relaxation of the one-hot vector. A temperature parameter $T \in (0, \infty)$, which decays exponentially during training, serves to controll to which extent it is relaxed. As T approaches zero, the selector layer outputs exactly one feature per node. The decoder of the concrete autoencoder can be any neural network suitable for the problem to be solved using the selected features.

D. Classifiers

We investigated the performance of support vector machines [19] and random forests [20] to classify PD patients from responsive patients. The SVM was implemented as a binary soft-margin SVM with radial basis function kernel. To determine the kernel parameter γ and the soft-margin weight C, we used a grid search and 10-fold cross-validation. Results showed that a combination of $\gamma = 2^{-7}$ and $C = 2^{9}$ is a sensible choice for our task. For the random forest, we choose to construct a forest consisting of 100 trees based on initial experiments with varying tree sizes.

IV. Experiments and results

A. Experimental setup

First, we compared the modality-based features to the radiomics features as well as the influence of PCA and SFFS on the performance of SVM and RF. We trained both classifiers on either $\mathcal{D}_{\mathrm{M}}$, $\mathcal{D}_{\mathrm{R}}$ or a combination of both using PCA and SFFS for feature reduction. Moreover, we aimed to investigate the value of the features derived from the individual organs

(liver, spleen, vertebrae). To this end, we trained our classifiers on subsets of $\mathcal{D}_{\mathrm{R}}$ containing only the features derived from one organ as well as the whole set including all organs. As $\mathcal{D}_{\mathrm{M}}$ only contains 5 features extracted from the liver and 4 each from spleen and vertebrae and good classification results on this dataset were only achieved when the SFFS selected features from all three organs, we did not conduct experiments with individual organs for this dataset. To achieve more stable results and reduce overfitting, we conducted each experiment 5 times with a different random split in 70% training and 30% test set. We used the mean balanced accuracy calculated over all 5 runs, denoted as A_{b}, to evaluate our trained models.

Subsequently, we investigated whether a more modern approach to the feature selection problem, namely AFS or CAE, was able to increase performance. For these analyses, we only used the radiomics features extracted from the liver as input (due to the poor performance of classifiers trained with the other organs). To evaluate the impact of AFS and CAE, we once again trained an SVM and an RF as in the previous experiments, repeating each experiment 5 times with different training and test set.

B. Classification results

An overview of our results for the experiments combining datasets $\mathcal{D}_{\mathrm{M}}$ and $\mathcal{D}_{\mathrm{R}}$ with PCA or SFFS is given in Table II. Using the clinical features worked best when employing SFFS to $\mathcal{D}_{\mathrm{M}}$ and training an RF ($A_{\mathrm{b}} = 84.89\%$). For the radiomics features, using the organs spleen and vertebrae as input resulted in poor performance of both classifiers, regardless of which feature reduction method was employed. Combining all three organs also resulted in lower accuracies than using features derived from the liver as input. The best overall results of $A_{\mathrm{b}} = 86.90\%$ were achieved when using only the liver and combining SVM and SFFS for the training using either $\mathcal{D}_{\mathrm{R}}$ or a combination of $\mathcal{D}_{\mathrm{R}}$ and $\mathcal{D}_{\mathrm{M}}$ as input data.

Table III shows the results from our second batch of experiments conducted with only features derived from the liver images as well as AFS and CAE for feature selection. In combination with the SVM, both methods resulted in an increased balanced accuracy of $A_{\mathrm{b}} = 88.90\%$ (AFS) and $A_{\mathrm{b}} = 91.11\%$ (CAE). When using an RF classifier, only CAE was able to increase performance to $A_{\mathrm{b}} = 94.40\%$, which is the best overall accuracy in our experiments.

TABLE II: Overview of the results for training on the mean features calculated per acquisition from all imaging modalities. The number of samples is 72, the number of features is 15 for $\mathcal{D}_{\mathrm{M}}$ and 3240 times the number of organs for $\mathcal{D}_{\mathrm{R}}$.

feature set	organs	classifier	reduction	A_{b}
$\mathcal{D}_{\mathrm{M}}$	liver, spleen, vertebrae	RF	SFFS	84.89%
$\mathcal{D}_{\mathrm{M}}$	liver, spleen, vertebrae	RF	PCA	67.12%
$\mathcal{D}_{\mathrm{M}}$	liver, spleen, vertebrae	SVM	SFFS	80.56%
$\mathcal{D}_{\mathrm{M}}$	liver, spleen, vertebrae	SVM	PCA	72.18%
$\mathcal{D}_{\mathrm{R}}$	liver, spleen, vertebrae	RF	SFFS	68.80%
$\mathcal{D}_{\mathrm{R}}$	liver, spleen, vertebrae	RF	PCA	80.35%
$\mathcal{D}_{\mathrm{R}}$	liver, spleen, vertebrae	SVM	SFFS	73.94%
$\mathcal{D}_{\mathrm{R}}$	liver, spleen, vertebrae	SVM	PCA	62.46%
$\mathcal{D}_{\mathrm{R}}$	liver	RF	SFFS	77.52%
$\mathcal{D}_{\mathrm{R}}$	liver	RF	PCA	85.01%
$\mathcal{D}_{\mathrm{R}}$	liver	SVM	SFFS	86.90%
$\mathcal{D}_{\mathrm{R}}$	liver	SVM	PCA	72.81%
$\mathcal{D}_{\mathrm{R}}$	spleen	RF	SFFS	56.89%
$\mathcal{D}_{\mathrm{R}}$	spleen	RF	PCA	65.92%
$\mathcal{D}_{\mathrm{R}}$	spleen	SVM	SFFS	43.23%
$\mathcal{D}_{\mathrm{R}}$	spleen	SVM	PCA	56.46%
$\mathcal{D}_{\mathrm{R}}$	vertebrae	RF	SFFS	60.02%
$\mathcal{D}_{\mathrm{R}}$	vertebrae	RF	PCA	66.89%
$\mathcal{D}_{\mathrm{R}}$	vertebrae	SVM	SFFS	51.34%
$\mathcal{D}_{\mathrm{R}}$	vertebrae	SVM	PCA	57.68%
$\mathcal{D}_{\mathrm{M}}$ + $\mathcal{D}_{\mathrm{R}}$	liver	RF	SFFS	68.19%
$\mathcal{D}_{\mathrm{M}}$ + $\mathcal{D}_{\mathrm{R}}$	liver	RF	PCA	82.13%
$\mathcal{D}_{\mathrm{M}}$ + $\mathcal{D}_{\mathrm{R}}$	liver	SVM	SFFS	86.90%
$\mathcal{D}_{\mathrm{M}}$ + $\mathcal{D}_{\mathrm{R}}$	liver	SVM	PCA	85.01%

TABLE III: Overview of the results for training on the mean features calculated per acquisition from all imaging modalities. The number of samples is 72, the number of features is 3240. N_s denotes the number of selected features

feature set	organs	classifier	reduction	A_{b}	N_s
$\mathcal{D}_{\mathrm{R}}$	liver	RF	AFS	73.33%	8
$\mathcal{D}_{\mathrm{R}}$	liver	RF	CAE	94.40%	14
$\mathcal{D}_{\mathrm{R}}$	liver	SVM	AFS	88.90%	7
$\mathcal{D}_{\mathrm{R}}$	liver	SVM	CAE	91.11%	17

V. Discussion

In our first round of experiments, we found that combining an SVM with SFFS trained on radiomics features extracted from the patients livers exclusively works best. When using the clinical features only for training, employing SFFS to the input data gives better results than PCA and choosing an RF for training results in higher accuracy than an SVM.

By using radiomics features compared to using only metrics derived directly from imaging modalities, we were able to increase the performance of the SVM. We found that when combining $\mathcal{D}_{\mathrm{M}}$ and $\mathcal{D}_{\mathrm{R}}$, features from $\mathcal{D}_{\mathrm{M}}$ never got selected by SFFS when combining it with an SVM. The performance of the RF even decreases, regardless of which feature selection method is used. Only for a combination of SVM and PCA, the classifier seems to benefit from the additional information provided by $\mathcal{D}_{\mathrm{M}}$. When using the spleen or vertebrae as input data or combining all three organs, classifier performance decreased significantly for both SVM and RF as well as PCA and SFFS.

Regarding the two feature selection methods based on deep learning, we found that AFS resulted in a slightly higher accuracy when combining it with an SVM. However, when using it in combination with an RF, the acurracy was significantly lower than when employing SFFS or PCA. The CAE on the other hand increased the performance of both RF and SVM significantly compared to PCA and SVM, leading to the best overall balanced accuracy of $A_{\mathrm{b}} = 94.40\%$ in this study when combining this method with the RF.

As CAE resulted in the best balanced accuracy for both RF and SVM, we investigated the difference in selected features between CAE and SFFS, which lead to the best classifier performance in our first experiments. We found that CAE chose mainly features derived from the gray-level co-occurence matrix (GLCM, 70% of selected features), follwed by run length (17%), gradient (11%) and histogram (2%) features, whereas SFFS resulted in a selection of histogram (35%), gradient (30%), GLCM (20%) and intensity (15%) features. We attribute this to the fact that SFFS is designed to find the best combination of features in a trial-and-error way by going iteratively trough the feature matrix starting with the first feature. If several features in one iteration would result in the same accuracy if added to the already selected features, the first one is picked. The method is hence more likely to be biased towards features located at the beginning of the feature matrix than the CAE. AFS, which also resulted in a slightly increased accuracy compared to SFFS when compared with an SVM, selected even more GLCM features (88%), followed by gradient (7%) and run length (5%) features.

As we could observe a better perfomance using CAE and AFS (at least in combination with the SVM) for feature selection, we feel that using deep learning for classification could further improve our results. Our future research will hence include the investigation of convolutional neural network architectures as a method to train an end-to-end model without the need to perform explicit feature extraction and feature reduction. Moreover, the task we solved in this study was only intended as a first proof of concept. The main goal of our research is finding a way to predict treatment response for individual patients in the early stages of immunotherapy. Hence, we plan to investigate methods to detect PD patients based only on their first or first and second examination.

Although classification accuracy could still be improved, our results indicate that, in general, predicting therapy response based on radiological imaging should be feasible. However, a major drawback in our current work is the very small amount of patients we were able to use for training. To be able to draw more general conclusions from our findings, our approach needs to be validated on a larger dataset, preferrably from varying geographic regions. This is especially necessary, as the patients included in our study were all treated in the same clinic and hence stem from a comparably small area. As the physiology of people can vary depending on their origin, including patients from other geographic regions would increase the generalization capabilities of our trained models.

VI. Conclusion

In this study, we presented an approach to predict treatment response of late-stage melanoma patients to immunotherapy with feature-based machine learning. We were able to differentiate between responsive patients and patients exhibiting a progressive disease under treatment with a balanced classification accuracy of $A_{\mathrm{b}} = 94.40\%$ with our best model, which combined a concrete autoencoder for feature selection with a random forest for classification. Moreover, we found that features derived from the liver of patients contain more significant information than features extracted from the liver or spleen. Using radiomics features resulted in an increased accuracy compared to using only clinical features derived directly from the imaging modalities. However, although these findings are promising, they need to be validated on a larger and preferrably more diverse dataset to be able to draw more general conclusions.

References

[1] Cancer Research UK, "Melanoma Skin Cancer Statistics," Accesed: 2020-02-17.

[2] "Krebs in Deutschland für 2013/14, Robert-Koch-Institut Berlin," Tech. Rep., 2017.

[3] Matthias Augustin and Gerd Glaeske, "TK hautkrebsreport 2019," Tech. Rep., 5 2019.

[4] The Nobel Assembly, ," online, 2018, https://www.nobelprize.org/prizes/medicine/2018/summary.

[5] Iwona Lugowska, Pawel Teterycz, and Piotr Rutkowski, "Immunotherapy of melanoma," *Współczesna Onkologia*, vol. 2018, no. 1, pp. 61–67, 2018.

[6] Stefano Trebeschi, Silvia Drago, Nicolai Birkbak, and et al., "Predicting response to cancer immunotherapy using non-invasive radiomic biomarkers," *Annals of oncology : official journal of the European Society for Medical Oncology*, vol. 30, 03 2019.

[7] Roger Sun, Elaine Limkin, Maria Vakalopoulou, and et al., "A radiomics approach to assess tumour-infiltrating cd8 cells and response to anti-pd-1 or anti-pd-l1 immunotherapy: an imaging biomarker, retrospective multicohort study," *The Lancet Oncology*, vol. 19, 08 2018.

[8] Zarmeena Dawood, Nicolas Coudray, Randie Kim, and et al., "Prediction of response and toxicity to immune checkpoint inhibitor therapies (ici) in melanoma using deep neural networks machine learning.," *Journal of Clinical Oncology*, vol. 36, pp. 9529–9529, 05 2018.

[9] Shipra Gandhi, Sarabjot Pabla, and Mary et al. Nesline, "Algorithmic prediction of response to checkpoint inhibitors: Hyperprogressors versus responders.," 06 2017.

[10] Carl Morrison, Sarabjot Pabla, Jeffrey Conroy, and et al., "Predicting response to checkpoint inhibitors in melanoma beyond pd-l1 and mutational burden," *Journal for ImmunoTherapy of Cancer*, vol. 6, 12 2018.

[11] Mark D. M. Leiserson, Vasilis Syrgkanis, Amy Gilson, Miroslav Dudik, Sharon Gillett, and et al., "A multifactorial model of t cell expansion and durable clinical benefit in response to a PD-l1 inhibitor," *PLOS ONE*, vol. 13, no. 12, pp. e0208422, Dec. 2018.

[12] Carsten Krieg, Malgorzata Nowicka, Silvia Guglietta, and et al., "High-dimensional single-cell analysis predicts response to anti-pd-1 immunotherapy," *Nature Medicine*, vol. 24, 02 2018.

[13] Robert Gillies, Paul Kinahan, and Hedvig Hricak, "Radiomics: Images Are More than Pictures, They Are Data," *Radiology*, vol. 278, pp. 151169, 11 2015.

[14] A. Liebgott, T. Küstner, H. Strohmeier, and et al., "ImFEATbox: a toolbox for extraction and analysis of medical image features," *IJCARS*, Sep 2018.

[15] Svante Wold, Kim Esbensen, and Paul Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37 – 52, 1987, Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[16] Pavel Pudil, Francesc J. Ferri, Jana Novovicová, and Josef Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5)*, vol. 2, pp. 279–283 vol.2, 1994.

[17] Ning Gui, Danni Ge, and Ziyin Hu, "AFS: An attention-based mechanism for supervised feature selection," in *AAAI Conference on Artificial Intelligence (AAAI-19)*, Honolulu, Hawaii, USA, 1 2019.

[18] Abubakar Abid, Muhammad Fatih Balin, and James Zou, "Concrete autoencoders for differentiable feature selection and reconstruction," *ArXiV*, 1 2019.

[19] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.

[20] Leo Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.