2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871257

A Comparative Study on the Potential of Unsupervised Deep Learning-based Feature Selection in Radiomics*

Tobias Haueise[1,2], Annika Liebgott[3] and Bin Yang[3]

Abstract— **In Radiomics, deep learning-based systems for medical image analysis play an increasing role. However, due to the better explainability, feature-based systems are still preferred, especially by physicians. Often, high-dimensional data and low sample size pose different challenges (e.g. increased risk of overfitting) to machine learning systems. By removing irrelevant and redundant features from the data, feature selection is an effective way of pre-processing. The research in this study is focused on unsupervised deep learning-based methods for feature selection. Five recently proposed algorithms are compared regarding their applicability and efficiency on seven data sets in three different sample applications. It was found that deep learning-based feature selection leads to improved classification results compared to conventional methods, especially for small feature subsets.**

Clinical Relevance— **The exploration of distinctive features and the ability to rank their importance without the need for outcome information is a potential field of application for unsupervised feature selection methods. Especially in multiparametric radiology, the number of features is increasing. The identification of new potential biomarkers is important both for treatment and prevention.**

I. Introduction

Medical imaging technologies have become a core pillar of modern clinical diagnostics, enabling fast and reliable insights into a patient's body and thus accelerating the diagnostic process as well as increasing patient outcome significantly, e.g. in oncology. Due to the continuous advances, the complexity of the acquired imaging data has been increasing steadily regarding both data quality and quantity. While yielding great potential for all kinds of clinical research based on radiological images, the increasing amount of data poses a challenge when it comes to analyzing said data manually. Especially in the context of large epidemiological imaging studies, like the German National Cohort [1] or UK Biobank [2], the overwhelming amount of data calls for dedicated, automated image analysis methods to facilitate an efficient and reliable processing of the information contained in such data sets. Automated post-processing methods, including machine learning (ML), has hence become increasingly important within the medical imaging community. One approach to ML-based image analysis, which has been gaining attention over the past years, is Radiomics [3], where large numbers of statistical image features are extracted from radiological images and then used to analyze a data set statistically or train a classifier to solve a scientific problem [4–6].

However, high-dimensional data sets pose a challenge for ML models. On the one hand, there are often a lot of features which are highly correlated. On the other hand, a feature space that exceeds the number of samples in the data set significantly can easily lead to overfitting of the classifier. The latter is especially crucial in medical image analysis, as the number of available subjects is limited in most clinical applications. Moreover, the higher the number of features used to train a classifier, the harder it gets to explain how the ML algorithm decides, which is especially important in clinical applications involved in diagnostic processes leading to e.g. decisions about the treatment of a patient. Another drawback of too high dimensional feature spaces is a longer time for both training of a model and inference, which becomes increasingly important if the system in question is supposed to be a real-time application. All these issues make choosing the right features out of all extracted metrics in an efficient, automated way a crucial task in Radiomics-based image analysis.

Feature selection (FS) algorithms have been successfully used in many different ML applications for years, e.g. SFFS [7], ReliefF [8] or Fisher score [9]. Most of them are supervised algorithms requiring label information for the selection process. Unsupervised FS algorithms, e.g. Laplacian score [10] or UDFS [11], have the advantage of reducing the dimensionality of the feature space without the requirement of label information. More recently, with deep learning (DL) gaining more and more popularity amongst the ML community, several DL-based FS algorithms have been proposed. Such methods aim to exploit the ability of neural networks (NN) to autonomously identify meaningful structures within a given data set to find the most informative features in a feature space. In this study, we employed some promising DL-based FS techniques to different radiological ML applications in order to investigate the potential of such approaches for Radiomics-based research. In the past years, there have been a lot of studies on FS-related topics. However, to the best of our knowledge, there has not been a publication comparing conventional FS and DL-based FS focused on Radiomics applications at the point of submission of this work. Most Radiomics-related studies investigate the best FS methods for specific applications, e.g. Delzell et al. [12], Shakir et al. [13] and Zhang et al. [14]. Instead of drawing general conclusions about FS in Radiomics, they rather discuss the potential value of Radiomics for diagnostic or predictive purposes in the respective application. Kuncheva et al. recently investigated the stability of various FS techniques for very

*This work was not supported by any organization.

[1,2]T. Haueise is with the Section on Experimental Radiology, University Hospital Tuebingen, Tuebingen, Germany and the Institute of Diabetes Research and Metabolic Diseases at Helmholtz Munich, Tuebingen, Germany tobias.haueise@med.uni-tuebingen.de

[3]A. Liebgott is with the Institute of Signal Processing and System Theory, University of Stuttgart, Stuttgart, Germany annika.liebgott@iss.uni-stuttgart.de

[3]B. Yang is with the Institute of Signal Processing and System Theory, University of Stuttgart, Stuttgart, Germany bin.yang@iss.uni-stuttgart.de

small data sets with large numbers of features in a more general way, including some medical imaging applications [15]. However, in contrast to this study, [15] did not employ DL-based FS methods.

In light of current discussions about the feasibility of feature-based ML systems compared to DL-based systems, which require no explicit feature extractions but are harder to interpret, this work is intended to investigate whether DL-based FS can serve to combine the advantages of both methods to further improve feasibility and reliability of ML-based medical image analysis.

II. Materials and Methods

A. Deep learning-based feature selection

DL-based FS uses concepts from general DL research and applies them to the problem of FS, namely to select a subset from the columns of the feature matrix $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ with N samples and D features. Recently proposed unsupervised methods included in this study are: auto encoder inspired FS (AEFS) [16], deep FS using a teacher-student network (TSFS) [17], Attention-based FS (AFS) [18], FS using batch-wise attenuation and feature mask normalization (FM) [19] and Concrete auto encoders (CAE) [20].

Similar to conventional filter methods, AEFS, TSFS, AFS and FM compute weights for each input feature to create a ranking by relevance. Therefore, in case of AEFS, the nonlinear relationship of $\boldsymbol{X}$ is modeled using a single-layer auto encoder (AE) by solving the optimization problem (1)

$$\min_{\theta} \left\| \boldsymbol{X} - \widehat{\boldsymbol{X}} \right\|_{\mathrm{F}} + \lambda \| \boldsymbol{W}_1 \|_{2,1} \qquad (1)$$

where $\widehat{\boldsymbol{X}} = g(f(\boldsymbol{X}, \theta_1), \theta_2) = \phi(\boldsymbol{X}\boldsymbol{W}_1 + b_1)\boldsymbol{W}_2 + b_2$ is the output of the AE with tunable parameters $\theta = \theta_1 + \theta_2 = \{\boldsymbol{W}_{1,2}, b_{1,2}\}$. The $L_{2,1}$-norm regularization imposes sparse rows in $\boldsymbol{W}_1$. Similarly, in TSFS, the nonlinear transformation to the low-dimensional subspace is modeled by a "teacher" function $\widehat{\boldsymbol{X}} = \mathcal{F}(\boldsymbol{X})$, e.g. t-SNE [21] as proposed in [17]. The "student" network is a single hidden-layer NN, trained to recreate $\mathcal{F}$. The loss function is similar to (1) with $\boldsymbol{W}_{1,2}$ and $b_{1,2}$ being the parameters of the first or second transformation of the NN, respectively. To rank the features, a score $s = diag(\boldsymbol{W}_1 \boldsymbol{W}_1^{\mathrm{T}})$ is calculated and the input features are sorted in descending order according to s.

In an unsupervised setting, AFS combines the training of a single hidden-layer AE with the training of shallow NNs ("attention nets") for each input feature to determine its probability of selection [18]. The resulting output matrix of the attention nets $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ is multiplied elementwise with $\boldsymbol{X}$ such that the AE is trained on $\boldsymbol{X} \odot \boldsymbol{A}$. Therefore, the attention nets and the AE are trained jointly. Based on the same concept, FM computes a nonlinear transformation of $\boldsymbol{X}$, batch-wise average and a softmax normalization resulting in a feature mask $\boldsymbol{M} \in \mathbb{R}^{N \times D}$. Similarly, a single hidden-layer AE is subsequently trained on $\boldsymbol{X} \odot \boldsymbol{M}$. The learning objective of both methods is summarized in (2) where $\mathcal{L}(\cdot)$ denotes the mean squared error (MSE) reconstruction loss of the AE $g(\cdot)$ with parameters Θ_1 and the FS network $f(\cdot)$ with parameters Θ_2. $\Theta = \Theta_1 + \Theta_2$ summarizes the trainable parameters of both sub-networks and reflects their joint optimization.

$$arg \min_{\Theta} \mathcal{L}[g(f(\boldsymbol{X}, \Theta_1), \Theta_2)] \qquad (2)$$

In contrast to the other methods, CAE is selecting a specific number of features instead of generating a ranking. From an architectural point of view, CAE is a single hidden-layer AE, similar to AEFS. However, the hidden-layer uses the properties of the Concrete distribution [22] to implement a differentiable "Concrete selector" layer that directly parameterizes the selection of input features.

B. Data sets

The DL-based FS methods are compared on seven data sets derived from three tasks with different properties (number and type of image modalities, sample size, number of target regions) from the Medical Segmentation Decathlon (MSD) [23]. The usage of these data sets is in accordance to the guidelines of the Declaration of Helsinki. For each task, the data sets share the same dimensionality and number of target regions (number of classes in downstream classification task).

- Task 1 – Brain tumor: This task contains multiparametric MR data from 484 patients diagnosed with either glioblastoma or lower-grade glioma. The sequences used were native T1-weighted (T1w), post-Gadolinium (Gd) contrast T1-weighted (T1gd), native T2-weighted (T2w) and T2 Fluid-Attenuated Inversion Recovery (FLAIR). The corresponding target regions are the three tumor sub-regions, namely edema, enhancing, and non-enhancing tumor ($D_{1\text{-}4}$).
- Task 2 – Prostate delineation: This task contains multiparametric MR data from 30 patients and includes T2w MR images and the apparent diffusion coefficient (ADC) from diffusion-weighted MRI. The target regions are image background, peripheral zone and transition zone of the prostate ($D_{5\text{-}6}$).
- Task 3 – Pancreas cancer: This task contains portal-venous phase CT scans from 281 patients undergoing resection of pancreas masses. The target regions are pancreatic parenchyma and pancreatic mass (D_7).

For all data sets, PyRadiomics 3.0.1 [24] was used to extract 107 Radiomics features from the original gray-level images. The extracted features are following the feature definitions as described by the Imaging Biomarker Standardization Initiative (IBSI) [25]. Additional parameters (e.g. the fixed bin width during gray-level quantization) of the extractor have not been optimized.

C. Experimental setup

For all experiments, the data sets were split 5 times in 80% training data and 20% test data. Each FS algorithm is used to select feature subsets of different sizes using the training data of each split successively. The number of selected features N_{F} ranges from $N_{\mathrm{F}} = 5$ to the full set of 107 features in 14 steps. Potential hyperparameters of the FS methods (e.g. number of nodes in the hidden-layer of the auto encoder in AEFS, the number of nodes in the NN in TSFS) were not optimized and follow the default settings provided in the respective publications. The performance of the FS algorithms is evaluated in terms of the mean balanced classification accuracy (*BAcc*, mean of specificity and sensitivity) over all 5 runs in a classification task using three conceptually different classifiers on the hold-out test data sets. The selected classifiers are support vector machines (SVM), random forests (RF) and k-nearest neighbors (kNN). RF and kNN are used in their out-of-the-box implementation in scikit-learn 1.0.2 [26]. The SVM was implemented as a multi-class soft margin SVM with radial basis function kernel. For each data set, the kernel parameter γ and the soft-margin weight C have been optimized using a 5-fold cross-validated grid search on a logarithmic grid. The DL-based methods are compared to conventional unsupervised FS algorithms, namely Laplacian score (LAP) and principal feature analysis (PFA) [27]. As a baseline, a random selection (RND) of features of each subset size is included.

III. Results

Fig. 1 shows the mean *BAcc* of the three classifiers when selecting 25 features using the DL-based FS methods on four examples, representing the three MSD tasks. FS on both imaging modalities of Task 2 (D_5, D_6) shows larger influence of the classifier used compared to D_1 and D_7. Table I summarizes the mean *BAcc* and standard deviation of the three classifiers for all data sets, using DL-based FS methods, conventional methods as well as the random selection of 25 features. Additionally, mean *BAcc* using all extracted features is shown. Considering $D_{1\text{–}4}$, AFS, FM and CAE perform better than the conventional methods whereas TSFS and LAP show significantly lower performance than RND. In the case of D_5, features selected by CAE improve the mean *BAcc* by 5.8% compared to all features. Using AFS on D_6, 4.6% improvement compared to the full set of features can be reported. Considering D_7, all methods yield similar performance within the standard deviation. Fig. 2 shows the mean *BAcc* of the three classifiers in dependence of the number of selected features on D_1. AEFS, TSFS and LAP show similar low performance on the smallest subset, mean *BAcc* using features obtained from the other methods is up to 10% higher.

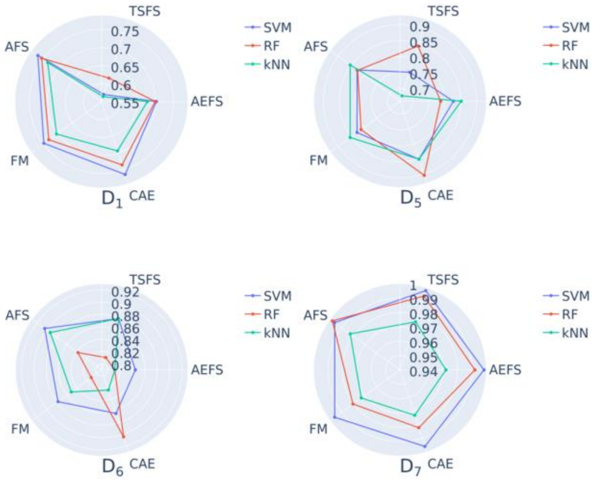

Figure 1: Comparison of mean *BAcc* of SVM, RF and kNN classifiers using DL-based FS methods to select 25 features. MSD Task 1 is represented by T1w MR images (D_1).

IV. Discussion

As shown in Table I, the application of DL-based FS can lead to competitive or better results compared to the conventional approaches ($D_{1\text{–}4}$) as well as the significant improvement compared to the full set of features ($D_{5\text{–}6}$). These data sets with a small number of samples compared to the number of features benefit most from the potential reduction of redundancy imposed by FS. The mean improvement is higher for some of the DL-based methods (TSFS, AFS, CAE) compared to the conventional methods. Considering D_7, all studied methods, including RND, yield surprisingly similar *BAcc* compared to the full set of features. The main difference of D_7 is the imaging modality, namely CT. It is possible that the distinctness of the extracted Radiomics features from CT scans is higher compared to MRI. Among the MRI data sets, only the post-Gd (D_2) reached a *BAcc* above 90%. The reproducibility and robustness of CT and MRI Radiomics features is subject of current research [28–30], their predictive capability should be further studied as non-uniform intensity scaling or noise in MRI could possibly harm the distinctiveness of the extracted features.

Following the methodology of [19] and studies investigating supervised FS techniques [31, 32], this study evaluates the performance of unsupervised FS methods in terms of a downstream classification task. However, especially in explorative analyses of potential new imaging biomarkers, studies requiring outcome information may lead to biased results. Comparing the features selected by the FS method with highest *BAcc* obtained from the two T2w data sets D_3 and D_6 over the five selections performed, a similarity of 84% is found. Comparing the similarity of the features selected by AFS across data sets, a mean similarity of equally 84% is achieved. CAE has a slightly higher similarity across data sets of 86%.

However, this study has some limitations. First, only a small number of features with low complexity, but according to the IBSI standard, was included. A higher number and complexity of the features, e.g. by including spectral transformations of the input images similar to [31, 32], could yield further differences between DL-based and conventional FS methods. Second, data set composition in terms of the

TABLE I. Mean balanced classification accuracy (%) and standard deviation (%) of the three classifiers using DL-based and conventional FS methods to select 25 features.

	D_1	D_2	D_3	D_4	D_5	D_6	D_7
RND	69.7 ± 3.2	84.3 ± 2.0	75.7 ± 4.0	72.5 ± 6.2	81.6 ± 1.7	85.0 ± 2.1	98.2 ± 0.7
all	75.4 ± 4.1	90.6 ± 3.5	82.2 ± 5.4	78.1 ± 6.2	81.1 ± 0.0	83.3 ± 3.1	98.7 ± 0.9
AEFS	69.3 ± 1.1	83.1 ± 2.7	74.2 ± 3.9	71.1 ± 4.1	82.1 ± 2.7	82.2 ± 1.6	98.8 ± 1.1
TSFS	58.4 ± 2.4	76.5 ± 3.1	69.4 ± 2.2	61.8 ± 2.9	75.9 ± 6.7	85.6 ± 3.1	98.9 ± 1.0
AFS	**75.2** ± 1.4	87.8 ± 2.9	80.0 ± 3.1	**77.3** ± 3.7	83.5 ± 1.3	**87.9** ± 2.9	**99.2** ± 0.7
FM	72.7 ± 1.8	**88.0** ± 2.3	79.5 ± 2.5	74.2 ± 3.5	83.0 ± 1.7	84.7 ± 2.7	98.3 ± 1.0
CAE	73.1 ±2.8	**88.0** ± 1.8	**80.7** ± 2.0	73.4 ± 3.1	**86.9** ± 2.5	86.4 ± 3.3	98.4 ± 0.9
LAP	65.8 ± 4.5	78.7 ± 1.8	67.5 ± 2.7	66.1 ± 4.3	78.4 ± 4.9	79.4 ± 3.2	98.2 ± 1.1
PFA	68.4 ± 3.2	85.3 ± 2.6	76.3 ± 2.6	72.5 ± 4.5	81.8 ± 2.0	84.9 ± 2.6	99.0 ± 0.9

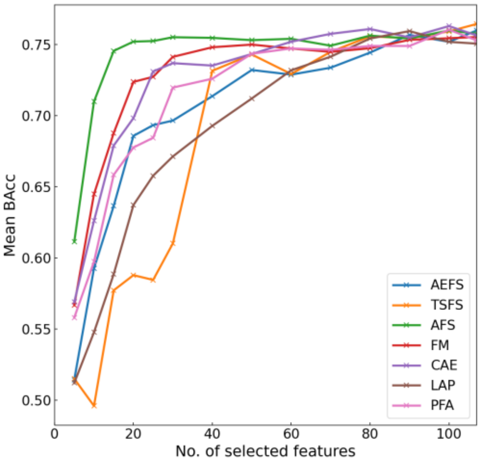

Figure 2: Mean *BAcc* of SVM, RF and kNN in dependence of the number of selected features by DL-based FS and conventional FS methods on D_1.

dimensionality of the feature matrix is limited since no larger imbalances between N and D were studied. Finally, a comprehensive study of the algorithms should include further relevant aspects like computational complexity and selection efficiency since FS is still "just" a preprocessing step in a Radiomics pipeline.

V. Conclusion

It was found that the presented unsupervised DL-based FS methods are competitive to other unsupervised conventional FS methods and can lead to improved classification results compared to conventional methods, especially for small feature subsets. The similarity of the selected features across different data sets is promising for future research on potential imaging biomarkers. A systematic comparison with conventional as well as DL-based supervised FS algorithms or the examination of the applicability of self-supervision to the task of feature selection can further pave the way for future applications in feature-based medical image analysis.

References

[1] F. Bamberg, H.-U. Kauczor, S. Weckbach, et al., „Whole-body MR imaging in the german national cohort: Rationale, design, and technical background," *Radiology*, vol. 277, no. 1, pp. 206–220, 2015.

[2] W. Ollier, T. Sprosen, T. Peakman, "Uk biobank: From concept to reality," *Pharmacogenomics*, vol. 6, no. 6, pp. 639–646, 2005.

[3] R. Gillies, P. Kinahan, H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data," *Radiology*, vol. 278, no. 2, pp. 151–169, 2015.

[4] R. Forghani, P. Savadjiev, A. Chatterjee, et al., "Radiomics and artificial intelligence for biomarker and prediction model development in oncology," *Comput and Struct Biotechnol J.*, vol. 17, pp. 995–1008, 2019.

[5] M. Avanzo, L. Wei, J. Stancanello, et al., "Machine and deep learning methods for radiomics," *Med Phys.*, vol. 47, no. 5, pp. e185–e202, 2020.

[6] A. Vial, D. Stirling, M. Field, et al., "The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review," *Transl Cancer Res*, vol. 7, no. 3, pp. 803–816, 2018.

[7] P. Pudil, F.J. Ferri, J. Novovicova, et al., "Floating search methods for feature selection with nonmonotonic criterion functions," in *Proc. of the 12th IAPR International Conference on Pattern Recognition*, 1994.

[8] I. Kononenko, E. Simec, M. Robnik-Sikonja, "Overcoming the myopia of inductive learning algorithms with relieff," *Applied Intelligence*, vol. 7, pp. 39–55, 1997.

[9] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, New York, 1973.

[10] X. He, D. Cai, P. Niyogi, "Laplacian score for feature selection," in *Proc. of the 18th International Conference on Neural Information Processing Systems*, Cambridge, MA, pp. 507–514, 2005.

[11] Y. Yang, H.T. Shen, Z. Ma, et al., "L2,1-norm regularized discriminative feature selection for unsupervised learning," in *Proc. of the Twenty-Second international joint conference on Artificial Intelligence*, Barcelona, Spain, pp. 1589–1594, 2011.

[12] D.A.P. Delzell, S. Magnuson, T. Peter, et al., "Machine learning and feature selection methods for disease classification with application to lung cancer screening image data," *Front Oncol.*, vol. 9, pp. 1393, 2019.

[13] H. Shakir, H. Rasheed, K. Rasool, et al., "Radiomic feature selection for lung cancer classifiers," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 5847–5855, 2020.

[14] Y. Zhang, A. Oikonomou, A.Wong, et al., "Radiomics-based prognosis analysis for non-small cell lung cancer," *Sci Rep*, vol. 7, no. 1, 2017.

[15] L.I. Kuncheva, C.E. Matthews, Á. Arnaiz-González, et al., "Feature selection from high-dimensional data with very low sample size: A cautionary tale," 2020.

[16] K. Han, Y. Wang, C. Zhang, et al., "Autoencoder inspired unsupervised feature selection," 2017.

[17] A. Mirzaei, V. Pourahmadi, M. Soltani, et al., "Deep feature selection using a teacher-student network," *Neurocomputing*, vol. 383, pp. 396–408, 2020.

[18] N. Gui, D. Ge, and Z. Hu, "AFS: An attention-based mechanism for supervised feature selection," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 3705–3713, 2019.

[19] Y. Liao, R. Latty, and B. Yang, "Feature selection using batch-wise attenuation and feature mask normalization," in *International Joint Conference on Neural Networks (IJCNN)*, 2021.

[20] M.F. Balin, A. Abid, and J. Zou, "Concrete autoencoders: Differentiable feature selection and reconstruction," in *Proc. of Machine Learning Research*, vol. 97, pp. 444–453, 2019.

[21] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[22] C.J. Maddison, A. Mnih, and Y.W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," arXiv:1611.00712, 2016.

[23] A.L. Simpson, M. Antonelli, S. Bakas, et al., „A large annotated medical image dataset for the development and evaluation of segmentation algorithms," arXiv:1902.09063, 2016.

[24] J.J.M. van Griethuysen, A. Fedorov, C. Parmar, et al., "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Res.*, vol. 77, no. 21, pp. 104 – 107, 2017.

[25] A. Zwanenburg, S. Leger, M. Vallières, et al., „Image biomarker standardisation initiative – feature definitions," 2016.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825–2830, 2011.

[27] I. Cohen, Q.T. Xiang, S. Zhou, et al., "Feature selection using principal feature analysis," 2002.

[28] L. Escudero Sanchez, L. Rundo, A.B. Gill, et al., "Robustness of radiomic features in CT images with different slice thickness, comparing liver tumor and muscle," *Sci Rep*, vol. 11, 8262, 2021.

[29] J. Lee, A. Steinmann, Y. Ding, et al., „Radiomics feature robustness as measured using an MRI phantom," *Sci Rep*, vol. 11, 3973, 2021.

[30] L.J. Jensen, D. Kim, T. Elegeti, et al., „Stability of Radiomic Features across Different Region of Interest Sizes – A CT and MR Phantom Study," *Tomography*, vol. 7, pp. 238–252, 2021.

[31] A. Liebgott, J. Steyer-Ege, T. Hepp, et al., "Feature reduction and selection: a study on their importance in the context of radiomics," in *Proc. of the Annual Meeting ISMRM*, 2019.

[32] A. Liebgott, S. Gatidis, V.C. Vu, et al., "Feature-based response prediction to immunotherapy of late-stage melanoma patients using pet/mr imaging," in *Proc. of the IEEE EUSIPCO*, 2020.