Graph-based approach applied to large radiomics feature dataset

Alessandro Ceresi

March 8, 2024

1 Introduction

Medical imaging is an indispensable tool in diagnosing and characterizing various diseases. In the realm of oncology, particularly in the study of multiple myeloma, Computed Tomography (CT) scans provide very important details from the prognostic point of view [1]. They are able to highlight lesions in the bone structure and bone marrow that are crucial for the understanding of the disease progression, as they are used for its staging with standardized criteria as R-ISS and IMPeTUs [2, 3].

The emerging field of radiomics focuses on extracting and analyzing a multitude of quantitative features from medical images, providing a rich source of information for predictive modeling. In a radiomics pipeline, from medical images are extracted patterns and properties with complex mathematical evaluations on the image grey-levels and described with numerical features. The number of feature extracted from a medical image can be a very large number, in order to make sure that all the possible relations between the image and the annotations are unveiled. This limitation can hinder the application of specific methods and models, as high-dimensional datasets may experience issues such as overfitting or challenging convergence. Hence, reducing and selecting variables are crucial in those cases. There are many algorithms for reducing the features in a high-dimensionality dataset, and each one has its own strengths and weaknesses. Traditional methods for selecting variables in constructing multidimensional feature datasets often rely on the performance of individual features. However, such approaches may encounter limitations when dealing with complex datasets. Graph-based approaches for feature pruning is an interesting and promising process in this. When executing an algorithm that assesses variables independently, it may show less than optimal performance. However, when these variables are combined in a two-dimensional space, it typically enhances the linear distinction between different classes. This highlights the necessity for advanced techniques that can capture relationships in higher dimensions.

This study leverages the DNetPRO algorithm [4], showcasing its efficacy in generating signatures based on lower-dimensional features. Graph theory, in this context, serves as a potent tool for unveiling intricate associations and dependencies among variables, particularly in the realm of feature selection and classification tasks. Unlike conventional methods that often assume linear separability, the inherent complexity of biological data, such as gene-expression patterns, necessitates a more flexible and adaptive approach. Graphs, with their capacity to represent nonlinear and higher-dimensional interactions through nodes and edges, encapsulate the multifaceted relationships among features. As a result, the graph structure captures the intricate web of connections, dependencies, and co-regulatory mechanisms present in biological systems. This complexity is vital for achieving optimal classification performance, as it acknowledges the reality that grey-level pixel interactions typically follow intricate patterns, including non-linear or multi-dimensional behaviors. By embracing such complexity, the graph-based methodology allows for the identification of discriminative features that contribute significantly to classification accuracy, even in scenarios where traditional linear separation may fall short.

Complex separation surfaces recognize and use the complexity of radiomics datasets, offering a more holistic and adaptive approach to feature selection and classification in the context of high-dimensional, interconnected variables.

2 Materials and methods

Here the data used for the work will be explained. Moreover all the processes implemented in the pipeline will be explained and accurately discussed in order to clear the reasons of the choices made.

Input Data

The input data for this study consists of radiomics features extracted from computed tomography (CT) images of spine segmentation in 93 patients diagnosed with multiple myeloma. These features, totaling 982, were obtained, employing PyRadiomics library [5], from the original image and from the image result of the application of few wavelet filters proposed by the library. The radiomics features provide a comprehensive representation of the spine segmentation in CT images, capturing diverse patterns and characteristics relevant to multiple myeloma patients.

Graph-based approach

The graph-based approach presented is a method employed for feature selection and classification in the context of radiomics data. The methodology draws upon graph theory concepts to construct a fully connected symmetric weighted network, where nodes represent individual radiomics features, and edges capture the classification performance of feature pairs. The steps it performs to reduce the number of feature of high-dimensionality dataset are the following:

- selection of the feature couple;
- evaluation of the classification model with Cross-Validation on the couple of features;
- creation the fully connected symmetric weighted network, with as nodes the features connected by an edge weighted with the performance of the model with just the two features it connects;
- removing of edges with a weight lower than a selected threshold;
- remove the pendant nodes;
- reduce the dataset features with the selection performed.

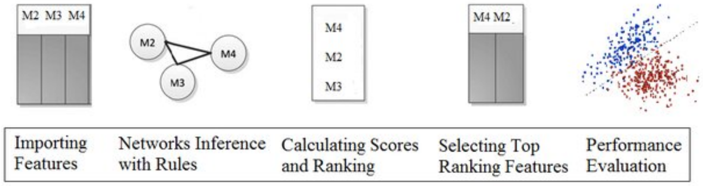
M2 M3 M4
M2
M4
M3
M4
M2
M3
M4 M2
Importing Features
Networks Inference with Rules
Calculating Scores and Ranking
Selecting Top Ranking Features
Performance Evaluation

Figure 1: Diagram of the pipeline of this work. Example provided with a dataset of three features. [6]

The dataset is high-dimensional dataset, consisting of 982 features extracted using PyRadiomics from the medical images of 93 patients with multiple myeloma. The selection of feature pairs is a crucial step as it allows exploring potential interactions and dependencies between different features that might not be apparent when considering each single feature.

Once the feature pairs are selected, their performance is evaluated using a Logistic Regression (LR) model. LR is a machine learning algorithm known for its effectiveness in binary classification processes also in high-dimensional spaces. The performance of the LR model is assessed using 5-fold cross-validation. In this process, the dataset is divided into five subsets or 'folds'. The model is trained on four of these folds and tested on the remaining fold. This process is repeated five times, with each fold serving as the test set once. This method provides a robust estimate of the model's predictive performance and helps prevent overfitting, which is particularly important when dealing with a limited sample size like 93 patients.

Following the model evaluation, a fully connected symmetric weighted network is created. In this network, each node represents a feature, and each edge represents the performance of the LR model when using the two features it connects. The weight providing a measure of the 'importance' or 'relevance' of the feature pair it connects. This network serves as a visual and mathematical representation of the complex relationships between different features and their collective impact on the model's performance.

The next step involves removing edges with a weight lower than a selected threshold, which is 0.70 in this case. This step is designed to simplify the network and focus on the most predictive feature pairs. By removing edges with low weights, we eliminate weak or potentially spurious relationships between features, thereby reducing noise and improving the interpretability of the network.

After edge removal, pendant nodes, which are nodes connected by a single edge, are also removed. These nodes are typically less informative because they are only connected to one other feature. By removing these nodes, we further simplify the network and reduce the dimensionality of the data.

The final step involves reducing the dataset features based on the selection performed in the previous steps. This step is crucial as it helps to eliminate irrelevant or redundant features, thereby improving the efficiency and performance of the model. This methodology ensures a thorough and systematic approach to feature selection and model evaluation, providing a solid foundation for the analysis of the multiple myeloma dataset.

3 Results and analysis

The analysis performed yielded significant results that highlighted a set of feature combination that are known for their informativeness for Progression-Free Survival (PFS) information. One of the combination of the features selected is represented in Fig. 2. It is possible to note that the combination is made of a feature from the original image and a feature from the wavelet filter. It is a combination known for its informativeness, which results frequently also in survival analysis employing Cox model. It appears to be a powerful combination in terms of accuracy score also for other selection performed. In fact, it is possible to see from the image in Fig. 3 that even with a more complex graph the high-score combinations are brought from original-wavelet combinations. The feature combinations selected, have been identified as key predictors in the LR model, as evidenced by their high weights in the fully connected symmetric weighted network.The resulting graph (showed in Fig. 4), which is a visual representation of the network, clearly illustrates the relationships between these informative features. Each node in the graph represents a feature, and each edge represents the performance of the LR model when using the two features it connects. The weight of an edge corresponds to the model's performance, providing a measure of the 'relevance' of the feature pair it connects. This visual representation allows for an intuitive understanding of the complex relationships between different features and their collective impact on the model's performance.

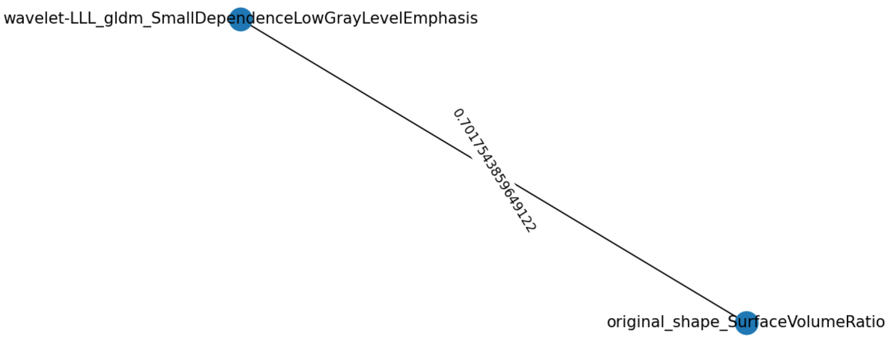
wavelet-LLL_gldm_SmallDependenceLowGrayLevelEmphasis
0.7017543859649122
original_shape_SurfaceVolumeRatio

Figure 2: One of the connected graphs resulting from the complete feature reduction. Each node (blue dots) represents a feature and the link has the score obtained from the two connected features written on it.

The total list of features selected through this process is comprehensive and includes all the features that have been identified as informative for PFS information. This list, presented in Tab. 1 serves as a valuable resource for further analysis and can be used to guide future research in the field of multiple myeloma.

Employing all the features listed in the table, that have been proved significant, it is possible

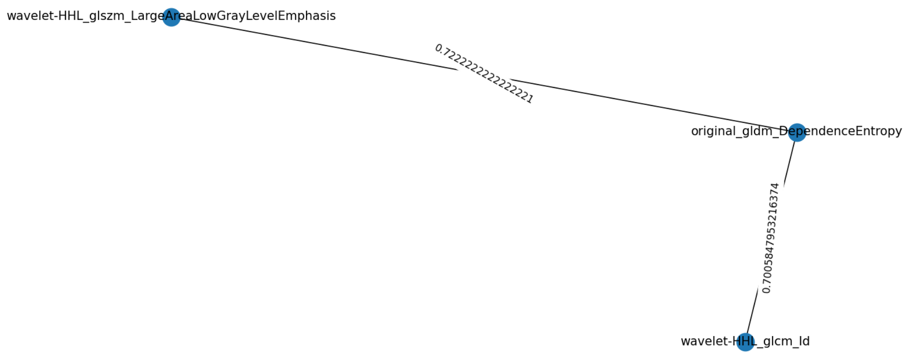
wavelet-HHL_glszm_LargeAreaLowGrayLevelEmphasis
0.7222222222222221
original_gldm_DependenceEntropy
0.7005847953216374
wavelet-HHL_glcm_Id

Figure 3: One of the connected graphs resulting from the complete feature reduction. Each node (blue dots) represents a feature and the links have the score obtained from the two connected features written on it.

Prefix	**Features**
`original`	gldm_HighGrayLevelEmphasis, shape_SurfaceVolumeRatio, glrlm_HighGrayLevelRunEmphasis, glcm_Autocorrelation, gldm_DependenceEntropy, glrlm_ShortRunHighGrayLevelEmphasis, glszm_HighGrayLevelZoneEmphasis, glrlm_RunEntropy, glszm_SmallAreaHighGrayLevelEmphasis, firstorder_Entropy
`wavelet-HHL`	glszm_LargeAreaLowGrayLevelEmphasis, glcm_Id, glcm_MaximumProbability
`wavelet-HLL`	glcm_Autocorrelation
`wavelet-LLL`	firstorder_InterquartileRange, glcm_Imc2, gldm_SmallDependenceLowGrayLevelEmphasis
`wavelet-LLH`	gldm_LargeDependenceHighGrayLevelEmphasis, firstorder_Kurtosis

Table 1: Table with all the features selected by the algorithm.

to obtain a model with high performances without the risk of overfitting that one encounters when uses such a high-dimensionality features dataset.

4 Conclusions

The pipeline employed in this study has proven to be effective in identifying informative features for PFS information in a dataset of 93 patients with multiple myeloma. The results obtained provide valuable insights into the complex relationships between different features and their impact on the model's performance. In particular combinations between original filter and wavelet filter is revealed as a pattern that seems to provide particularly suited classification results.

These insights can be used to guide future research and improve the prediction of PFS in patients with multiple myeloma.

References

[1] Ankit Agarwal, Alin Chirindel, Bhartesh A. Shah, and Rathan M. Subramaniam. Evolving role of FDG PET/CT in multiple myeloma imaging and management. *AJR. American journal of roentgenology*, 200(4):884–890, April 2013.

[2] Cristina Nanni, Annibale Versari, Stephane Chauvie, Elisa Bertone, Andrea Bianchi, Marco Rensi, Marilena Bellò, Andrea Gallamini, Francesca Patriarca, Francesca Gay, Barbara Gamberi, Pietro Ghedini, Michele Cavo, Stefano Fanti, and Elena Zamagni. Interpretation criteria for FDG PET/CT in multiple myeloma (IMPeTUs): final results. IMPeTUs (Italian myeloma criteria for PET USe). *European Journal of Nuclear Medicine and Molecular Imaging*, 45(5):712–719, May 2018.

[3] Sathish Gopalakrishnan, Anita D'Souza, Emma Scott, Raphael Fraser, Omar Davila, Nina Shah, Robert Peter Gale, Rammurti Kamble, Miguel Angel Diaz, Hillard M. Lazarus, Bipin N. Savani, Gerhard C. Hildebrandt, Melhem Solh, Cesar O. Freytes, Cindy Lee, Robert A Kyle, Saad Z. Usmani, Siddhartha Ganguly, Amer Assal, Jesus Berdeja, Abraham S. Kanate, Binod Dhakal, Kenneth Meehan, Tamila Kindwall-Keller, Ayman Saad, Frederick Locke, Sachiko Seo, Taiga Nishihori, Usama Gergis, Cristina Gasparetto, Tomer Mark, Yago Nieto, Shaji Kumar, and Parameswaran Hari. Revised-International Staging System (R-ISS) is Predictive and Prognostic for Early Relapse (<24 months) after Autologous Transplantation for Newly Diagnosed Multiple Myeloma (MM). *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation*, 25(4):683–688, April 2019.

[4] Nico Curti. Nico-Curti/DNetPRO, September 2023. original-date: 2019-09-14T13:08:33Z.

[5] Computational Radiomics System to Decode the Radiographic Phenotype - PubMed.

[6] Salih Tutun, Sina Khanmohammadi, and Chun-An Chou. *A Network-based Approach for Understanding Suicide Attack Behavior.* May 2016.

Annexes

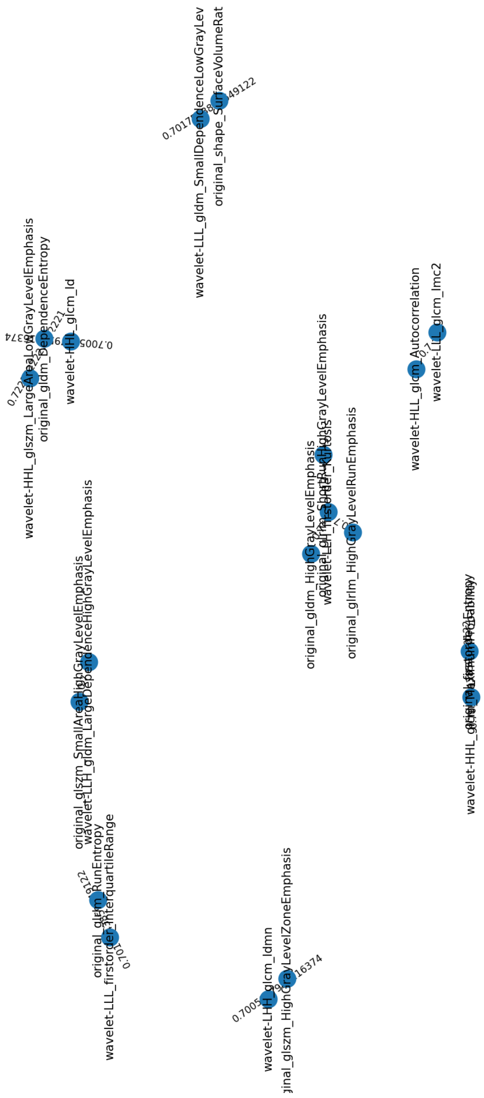
wavelet-LLL_gldm_SmallDependenceLowGrayLev
original_shape_SurfaceVolumeRat
0.7017
49122
wavelet-HHL_glszm_LargeAreaLowGrayLevelEmphasis
original_gldm_DependenceEntropy
wavelet-HHL_glcm_Id
0.7005
wavelet-HLL_gldm_Autocorrelation
wavelet-LLL_glcm_Imc2
original_gldm_HighGrayLevelEmphasis
wavelet-LLH_firstorder_Kurtosis
original_glrlm_HighGrayLevelRunEmphasis
original_glszm_SmallAreaHighGrayLevelEmphasis
wavelet-LLH_gldm_LargeDependenceHighGrayLevelEmphasis
original_glrlm_RunEntropy
wavelet-LLL_firstorder_InterquartileRange
wavelet-LHH_glcm_Idmn
original_glszm_HighGrayLevelZoneEmphasis
0.7005
16374

Figure 4: Total network representation resulting from the complete feature reduction. Each node (blue dots) represents a feature and the link has the score obtained from the two connected features written on it.