Graph-based approach applied to large radiomics feature dataset

Alessandro Ceresi

March 11, 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 (MM), 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]. However, due to the complex nature of the lesions brought by MM, clinicians still make it difficult to assess disease progression with certainty. To make the identification of individuals most susceptible to possible relapse occur automatically, while trying to reduce misclassification errors, this study aims to exploit the field of radiomics.

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 characteristic can limit 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, when executing an algorithm that assesses variables independently, it may show less than optimal performance. Nevertheless, 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. In this regard, graph-based approaches for feature pruning is an interesting and promising process. The objective of this study, embedded within a radiomics pipeline, is to assess the performance of graph-based approaches in classification problems.

The graph-based algorithm employed in this work is the DNetPRO algorithm [4], showcasing its efficacy in generating signatures based on lower-dimensional features. Graph theory, in this

context, serves as a powerful 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 and radiomics derived numerical properties, 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. 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 with the corresponding spine segmentation. The features were extracted from 93 patients diagnosed with multiple myeloma (MM), some of which were subjected to a relapse during the period of the study. As it can be seen in the image shown in Fig. 1, there is no clear visual difference between those who are going to experience that event and those who do not. The classification model has to assess these lesions and verify if the patient is prone to relapse.

These features, totaling 806, 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 MM 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:

- dividing of the dataset in two subsets;
- evaluation of the model with the total set of features;

((a)) ((b))

Figure 1: Two CT images representing the spine region of two MM patients, obtained from a sagittal projection. The image on the left (a) represents a patient labeled as True for the PFS, meaning it is going to experience a relapse, while on the right (b) image is represented a MM patient labeled as False.

- selection of the feature couple;
- evaluation of the classification model with Cross-Validation on the couple of features;
- creation of 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;
- conserving a specified number of the edges with the highest weights;
- remove the pendant nodes;
- reduce the dataset features with the selection performed;
- evaluation of the Logistic Regression (LR) model with the selected features.

Firstly, the dataset is normalized and divided into two sets. The first set, accounting for the 66% of the total dataset, is employed for the selection of the most performative feature couples. The second set, accounting for the resting 34%, is employed as hold-out for the validation of the selection, in order for it to be independent data for the test.

The LR model is evaluated with the complete set of features. Given the high-dimensionality feature dataset, the training needs to be performed with a penalization which regularize features which are not contributing on the model. The selected penalization is LASSO penalization, which allows to avoid the contribution of those non-related features. Afterward, the feature selection is performed in order to verify whether it is able to improve the model results.
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

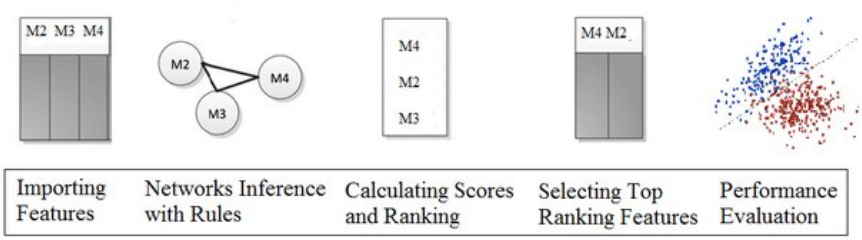

Figure 2: Diagram of the pipeline of this work. Example provided with a dataset of three features. [6]

feature. For each of the feature pairs, that is part of all possible feature combinations, performance is evaluated using the Logistic Regression (LR) model. LR is a is a fundamental statistical method used for binary classification tasks, popular due to its simplicity, interpretability, and efficiency. The classification is performed on the Progression-Free Survival (PFS) label, which states if the patient had a disease relapse during the study period. Performance is evaluated 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 as the one we have available.

A fully connected symmetric weighted network is created with the results of all the models training, each performed with a combinations of two features. In this network, the nodes represent the features, and the edges represent 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 steps involves selecting the most important features obtained. It is performed by specifying a number of edges that one wants to obtain at the end of the selection. The algorithm selects the chosen number of edges from the ones with the highest weight, and the resulting features selected are the ones connected by them. The chosen number of selected edges in this pipeline is 10. This allowed to obtain an abundant number of features in order to verify the performance of the LR model with a progressive number of features selected, until finding the one with the best score.

Afterward, the edges that do not fall into the selection are removed. 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.

To demonstrate the competitiveness of the graph-based approach, an alternative method is employed. This entails utilizing a feature selection approach that relies on the individual performance of features. The process involves training and testing a LR model on one-dimensional datasets, each containing a single feature from the complete set

3 Results and analysis

When evaluating a LR model on the complete set of 806 features, as illustrated in the previous paragraph, with a LASSO penalization and a standard penalizer value of 1.5. The model score, when performed with 5 fold cross validation on the hold-out set, results with a mean value of 0.51 ± 0.09, showing poor classification results. It is clear that even with LASSO penalization the model struggles when featuring such a high-dimensionality dataset, and therefore it needs to be reduced.

Feature selection with graph-based method was performed on the subset of the data accounting for the 66% of the dataset, and then validated on the same hold-out set as the complete model. 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 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. 7. Nodes in the graph represents a feature, and 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.

The validations of the LR model with the subset of feature selected, evaluated on the hold-out set, are performed with 5 folds cross validation on sets with a progressive number of feature between the selected ones. The combination with the highest result gives a mean score of 0.64 ± 0.11. The obtained score is not representing a model with a high accuracy. The low accuracy values are partly due to the complexity of discriminating between the two labels of the classification, given the information enclosed in the employed dataset. However, it is possible to note that the trained model results are higher than the ones obtained employing the complete feature dataset, showing a clear improvement of the model strength.

The graph-based method is compared with a traditional method for selecting features, evaluating the performance of the single features instead of employing a two-dimensional space evaluation. The training of the LR model is performed with the same procedures as the previous one, just employing a selection of features based on single-variable performance. The model mean score results 0.60 ± 0.12. The result shows an higher result than the model evaluated with the complete

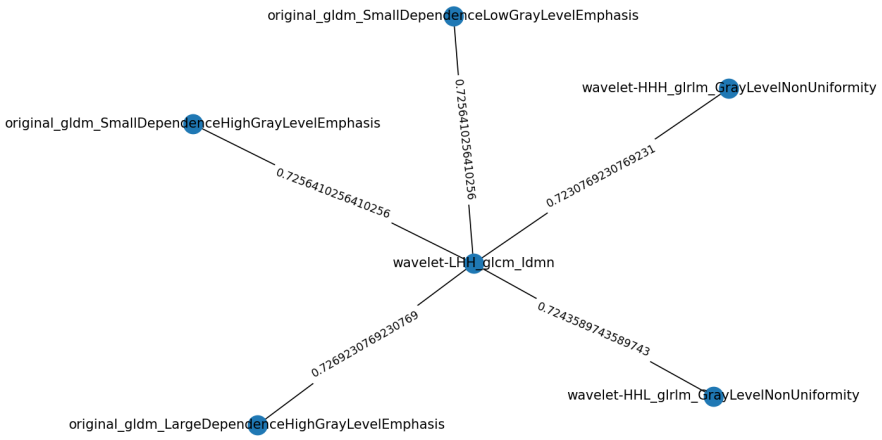

Figure 3: One of the connected graphs resulting from the complete feature reduction. Each node (blue dots) represents a feature and the edges have the score obtained from the two connected features written on them.

set of features, however, it does not show an high accuracy value. As for the previous case, the low accuracy result is due to the difficulties in obtaining an effective pattern recognition with the available data.

Comparing the models results it is possible to notice that feature selection is crucial for this kind of classification analysis. Both the methods involving feature selection obtained a higher score model with respect to the model trained on the complete feature set. Moreover, it is possible to notice that graph-based approach for the feature selection results a more suited method in this case, obtaining slightly higher results for the classification LR model.

Therefore, employing a graph-based approach for feature selection on a high-dimensionality dataset, it is possible to obtain a model with higher performances without the risk of overfitting. Furthermore, it appears that the process of selecting variables for multi-dimensional datasets relying on individual feature performance, does not take into consideration relationships between features which are important for the model learning process. Hence, the selection of feature performed taking into account two-dimensional combinations, in the present case, gives better results than a method relying on individual feature performance.

4 Conclusions

The pipeline employed in this study has proven to be effective in identifying informative features for the classification of the PFS label, employing a graph-based approach for a dataset of 93 patients suffering from multiple myeloma. The results of the LR model on the dataset in which graph-based feature selection was performed gave an accuracy score of 0.64 ± 0.11. The accuracy is not par-

ticularly high, however this is due to the lack of highly relevant information in the dataset for the classification.

Firstly, it is compared with the model evaluated on the complete set of features, with a LASSO regularization. The results of this model are quite poor, and the graph-based feature selection perform better avoiding overfitting with a feature set of reduced dimensions.

The Graph-based approach is compared also with an approach relying on individual feature performance. It turns out that the results are comparable, with the graph-based approach which shows a slightly higher score. Demonstrating the competitiveness of the method and the utility of graphs inside feature reduction approaches, for the discovery of pattern between different features in the same dataset.

The process has still room for improvements. The input dataset has a limited amount of samples, and with a higher number of patients results could be more reliable and generalizable. Another improvement that could be implemented is the way of testing the set of feature selected. Instead of evaluating the LR model with a progressively higher number of features from the ones with the best score to the least ones, all the combinations could be tested in order to find the ultimate best feature set.

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

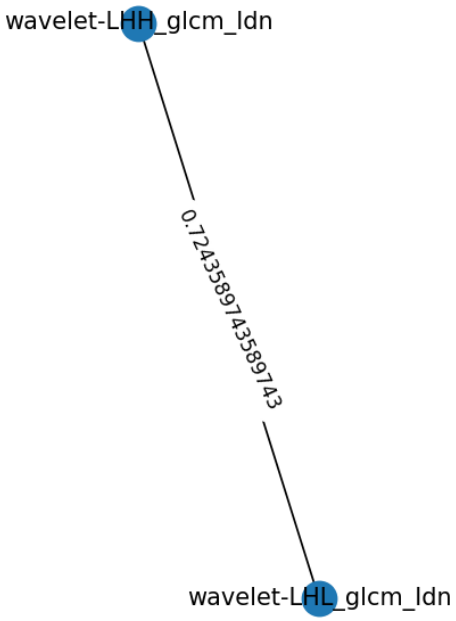

Figure 4: One of the connected graphs resulting from the complete feature reduction. Each node (blue dots) represents a feature and the edges have the score obtained from the two connected features written on them.

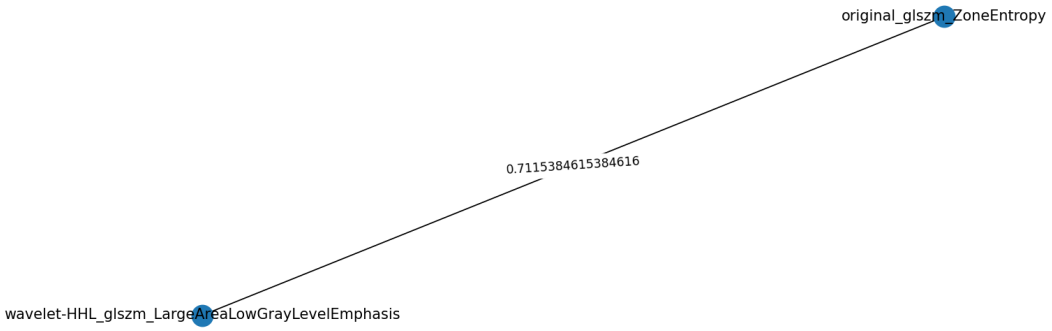

Figure 5: One of the connected graphs resulting from the complete feature reduction. Each node (blue dots) represents a feature and the edges have the score obtained from the two connected features written on them.

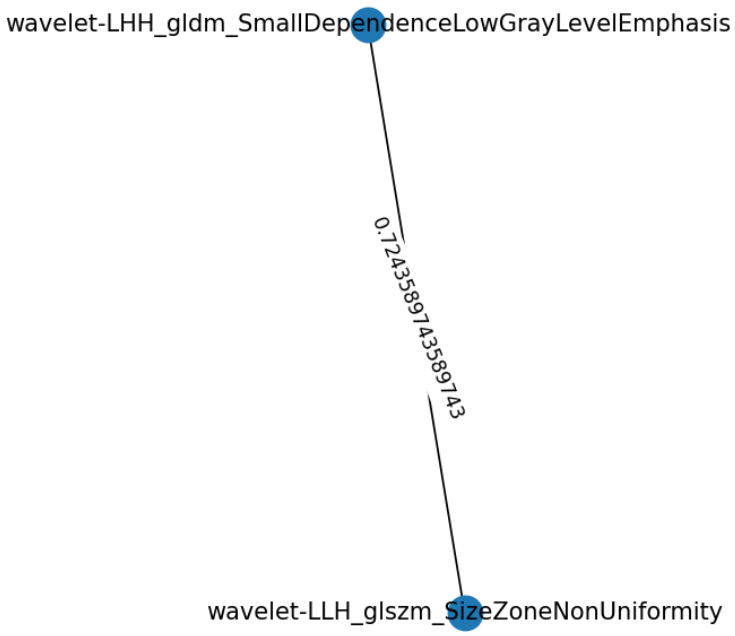

Figure 6: One of the connected graphs resulting from the complete feature reduction. Each node (blue dots) represents a feature and the edges have the score obtained from the two connected features written on them.

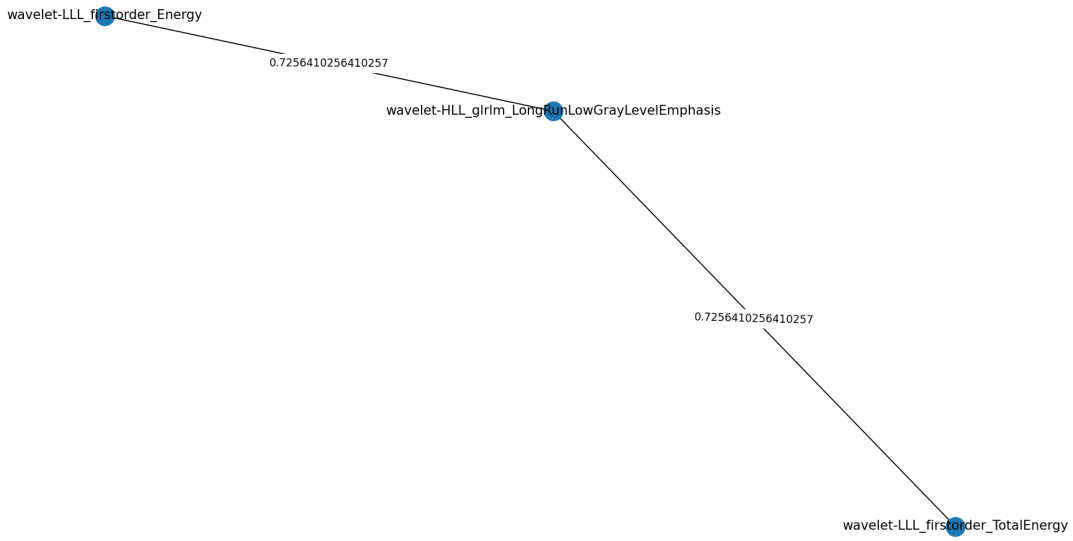

Figure 7: One of the connected graphs resulting from the complete feature reduction. Each node (blue dots) represents a feature and the edges have the score obtained from the two connected features written on them.