MOX-Report No. 17/2023

Distant supervision for imaging-based cancer sub-typing in Intrahepatic Cholangiocarcinoma

Savin, M.S.; Cavinato, L.; Costa, G.; Fiz, F.; Torzilli, G.; Vigano', L.; Ieva, F.

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/EMBC48229.2022.9871262

Distant supervision for imaging-based cancer sub-typing in Intrahepatic Cholangiocarcinoma

M.S. Savino[1], L. Cavinato[1], G. Costa[2], F. Fiz[3], G. Torzilli[2], L. Viganò[2], F. Ieva[4]

Abstract— **Finding effective ways to perform cancer sub-typing is currently a trending research topic for therapy optimization and personalized medicine. Stemming from genomic field, several algorithms have been proposed. In the context of texture analysis, limited efforts have been attempted, yet imaging information is known to entail useful knowledge for clinical practice. We propose a distant supervision model for imaging-based cancer sub-typing in Intrahepatic Cholangiocarcinoma patients. A clinically informed stratification of patients is built and homogeneous groups of patients are characterized in terms of survival probabilities, qualitative cancer variables and radiomic feature description. Moreover, the contributions of the information derived from the ICC area and from the peritumoral area are evaluated. The findings suggest the reliability of the proposed model in the context of cancer research and testify the importance of accounting for data coming from both the tumour and the tumour-tissue interface.**

Clinical relevance—**In order to accurately predict cancer prognosis for patients affected by ICC, radiomic variables of both core cancer and surrounding area should be exploited and employed in a model able to manage complex information.**

I. Introduction

Intrahepatic Cholangiocarcinoma (ICC) is an aggressive disease of the family of cholangiocarcinomas, which are tumors that stem from cholangiocytes of the biliary tree [1]. Because of its increasing incidence and mortality over the past three decades, it is now more than ever arising the urgency to further characterize the disease at early stages, as to modulate therapies and clinical decisions. In fact, detecting at baseline information that might inform the therapeutic pathway would allow to design more efficient lines of treatments. Such perspective has recently grown and developed in a research field, called personalized medicine, that hinges its root in efficiently extracting insights from multi-source patient data to shape clinical practice.

The promise made by personalized medicine in cancer research calls for special efforts for fully exploiting the information of data generated from different sources. A pivotal role in this sense has been played by imaging texture analysis, i.e., radiomics. It has become more and more important thanks to its advantage to non-invasively give access to tumor characterization [2]. Pertinently, radiomics consists in high-throughput quantitative features extracted from regions of interest in medical images such as CT or MRI scans. These features, also known as radiomic or texture features, can be many and are agnostic with respect to the clinical application. They represent a way to describe the information entailed in medical images and transform such information into matrix-shaped data, easier to handle and study [3]. However, radiomics is known to intrinsically possess some limitations, among all instability with respect to segmentation procedures and complexity in exhaustively shape the imaging representation of the lesions. In the context of ICC research, it has been recently proposed to explore a wider area of liver tumor for analysis, including both the very core of the lesion and the margin surrounding it, as to capture also the information of the tumor-tissue interface [4].

On the other hand, for as high-dimensional as they can be, radiomic data need to be properly analyzed to stratify patients basing on their cancer imaging texture characteristics. Ultimately, such analysis would devise subpopulations with different prognosis on which different therapeutic actions could be implemented. Many different techniques exist in literature to perform cancer sub-tying, mainly related to genomics. Recently, a promising distant-supervised approach has been borrowed from genomics and proposed for radiomic data, with the scope of carrying out clinically insightful patient clustering [5]. Such approach was proven to outperform other cancer subtyping methods proposed for genomic-based stratification purposes [6]. The concept of distant supervision comes from the Natural Language Processing field, where it is used to do relation extraction and sentiment analysis [7]. It consists on the training of a model for a task different from the final scope, using labels that are not completely pertinent with the problem to be tackled. It thus brings the possibility to solve tasks with non-retrievable labels in a supervised way. Here, the aim is to cluster patients in groups with different prognosis exploiting their imaging characteristics to predict survival estimates.

In this work, we exploit the Survival Supervised Graph Clustering (S2GC) model [6] as a distant supervision approach for imaging-based cancer sub-typing in ICC. Our aims and contributions are intended to be two-fold: (1) to provide radiomic characterization of groups of patients at different risk of death from ICC, in a risk stratification fashion, and (2) to study the contributions of the core cancer

[1] M.S. Savino (matteostefano.savino@mail.polimi.it) and L. Cavinato (lara.cavinato@polimi.it) are with the Department of Mathematics, Politecnico di Milano, Milan, Italy

[3] G. Costa (guido.costa@humanitas.it)), G. Torzilli (guido.torzilli@hunimed.eu) and L. Viganò (luca.vigano@hunimed.eu) are with the Division of Hepatobiliary and General Surgery, IRCCS Humanitas Research Hospital, Rozzano, Milan, Italy and the Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Italy

[4] F. Fiz (francesco.fiz@galliera.it) is with the Nuclear Medicine Unit, E. O. Ospedali Galliera, Genova

[6] F. Ieva (francesca.ieva@polimi.it) is with the Department of Mathematics, Politecnico di Milano, Milan, Italy and the Center for Health Data Science, Human Technopole, Milan, Italy

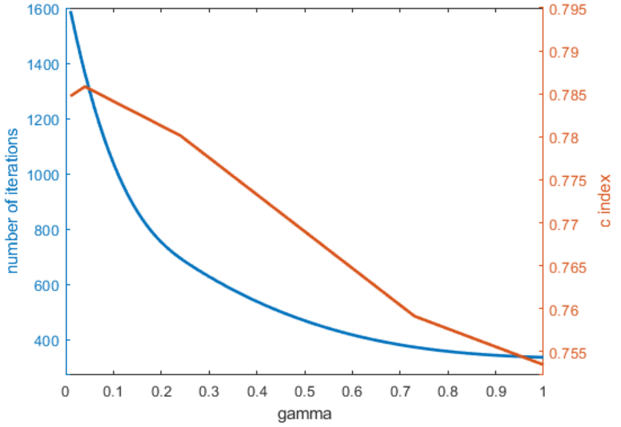

Fig. 1. In blue the number of iterations needed by the model to converge expressed as a function of γ, while in red the value of c-index as function of γ

information and of the peritumoral tissue information as to discuss potentialities and limitations of such approach.

II. Data Collection

Our study included two hundred and fifty-nine patients diagnosed with ICC from six different centers. Per every patient radiomic features, clinical variables and qualitative disease information were collected. Both the segmentation of regions of interest and the feature extraction phases were carried out from the portal phase of the CT scans by experienced radiologists using the LIFEx software (www.lifexsoft.org, [8]). The extracted radiomic data consisted of 50 variables for the core cancer segmentation and 50 variables for the margin segmentation. Pertinently, the margin was computed as the 5-mm region that was semi-automatically generated around the tumor by the software and then manually corrected to ensure that only peritumoral liver tissue had been included. Personal, i.e., sex and age, and tumor characteristics, i.e., size, number of nodules, ICC pattern [9] and grading, were included along with comorbidities and treatment information. The characteristics of the entire study population were coherent with those of the patients treated in the coordinating center [10]. This study was performed according to the Declaration of Helsinki. The local review board approved the study and informed consent was waived given the observational retrospective design of the study.

III. Methods

The analyses were developed as follows. First, a patient representation has been built from radiomic vectors as extracted from CT regions of interest, i.e., the lesions. Every patient vector carried the information extracted from both the core and the margin. In this sense, two different views of the tumor were assessed and analyzed for stratification, describing the core-margin relationship of the lesion. Second, the distant-supervised cancer sub-typing has been performed by (1) estimating a patient-to-patient graph basing on their imaging characteristics and survival probabilities and (2) clustering such graph in homogeneous subpopulations of nodes with similar properties. The algorithm's hyperparameters have been optimized. Finally, subpopulations of patients have been clinically characterized with clinical variables, exogenous to the model building, in order to validate the stratification procedure.

A. Patient-to-patient graph estimation

According to Supervised Survival Graph Clustering model [6] and its application to radiomic data [5], we performed the abovementioned two steps to perform cancer sub-typing and find clinically relevant clusters in ICC patients. The distant-supervised patient-to-patient similarity graph estimation was optimized basing on the following objective function:

$$
\begin{aligned}
&\min_{w;S} \sum_{k=1}^{m} \left(-\sum_{i=1}^{n} \delta_i \left(X_i^k w^k - log \sum_{j\in R_i} exp(X_j^k w^k) \right) \right) \\
&+ \lambda \sum_{k\neq j} \|X^k w^k – X^j w^j\|_2^2 + \eta \sum_{k=1}^{m} \|w^k\|_1 \\
&+ \min_{S} \gamma \sum_{i=1}^{n} \sum_{j=1}^{n} (\|X_i - X_j\|^2 + \|X_i w - X_j w\|^2) S_{i,j} + \mu S_{i,j}^2 \\
&s.t. \sum_{j}^{n} S_{I,j} = 1, S_i \succeq 0; i = 1,2,\ldots,n.
\end{aligned}
\tag{1}
$$

The loss function in (1) is composed by four terms, each with a specific methodological meaning and a clinical counterpart. The first one represents the estimate of the overall survival risks w^k for each radiomic feature of the $k-th$ view. Estimates were computed by solving the negative partial log-likelihood of the Cox model, where X_i^k is the radiomic vector in $k-th$ view of $i-th$ patient and R_i the set of patients observed alive almost at time T_i. In addition, δ_i is the censoring variable, n is the number of patients and m the number of radiomic views. Co-regularization between views' contributions on prediction and penalization of covariates are performed by a L2 regularization (second term) and L1 regularization (third term) respectively. Specifically, λ drives the regularization between radiomic views which, in this particular case, refer to the tumor core texture and the margin texture. By analyzing the control parameter λ we want to investigate the relationship of the core-margin interface with respect to prognostic risks. On the other hand, the sparsity control parameter η addresses the problem of high-dimensional data, in a feature selection fashion. In addition, importance ranking of features may be deduced according to the penalization coupled with each variable. These terms embody the core of distant supervision. In fact, we predict survival-related risks using a Cox proportional model and intend to exploit such risks, along with the imaging itself, in the definition of similarity between patients. Accordingly, the final term of (1) performs the learning of the graph S structure, i.e., its affinity matrix. It considers both the distance between observations in terms of radiomic views and the survival information of patients estimated in the first term. S is the $\mathbb{R}^{n\times n}$ affinity matrix of the patient-to-patient similarity graph where $S_{i,j}$ represents the similarity between

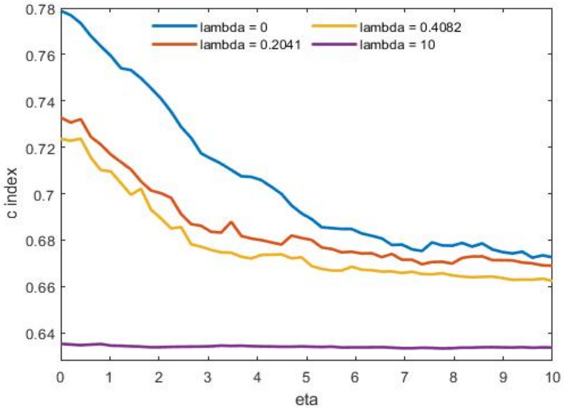

Fig. 2. Values of c-index as function of η for four different values of λ

patients i and j. γ is the learning rate and μ a trade-off parameter. In this way, two tasks are performed: the survival analysis with the computation of w given S and the similarity graph S estimation given the risks w.

B. Hyperparameters optmization

Grid search has been implemented for parameter optimization: optimal values were found for λ (the co-regularization parameter), η (the l_1 penalization parameter) and γ (the learning rate) by maximizing the Harrell's concordance index (c-index) of the estimated survival risks. Values returning the higher c-index were selected as optimal values. As displayed in Figure 1, the optimal choice was 0.04 for γ, meaning that convergence is almost guaranteed but requires several iterations, whereas regularization was found to be negligible. As shown in Figure 2, $\lambda = 0$ and $\eta = 0$ were the values that lead to the higher c-index performance.

C. Spectral clustering

A spectral clustering algorithm has been implemented for clustering the graph nodes, i.e., the patients, as it is suitable for medical application involving graphs [11]. The number of clusters nc has been chosen by following the eigengap heuristic, which can be applied to the graph Laplacians, either normalized or non-normalized [12]. This consists in choosing nc such that all the eigenvalues up to the $nc-th$ one are small whereas the $(nc+1)-th$ one is relatively large. Accordingly, the value of $nc = 4$ was selected. Clusters have been further characterized with exogenous clinical variables, testing differences on survival times and tumor qualitative scores. Radiomic contributions to risk of death from ICC have also been analyzed and discussed. P-values lower than 0.05 were considered significant and Bonferroni correction for multiple testing has been used.

IV. Results

Four cluster of patients have been obtained from the proposed pipeline, leading to four different risk classes. Of note, exploratory analysis has detected no confounders among clinical and personal variables. In Figure 3, the Kaplan-Meier overall survival probability curves for such groups are displayed. Group 3 (yellow line) and group 4 (grey line), however containing only few patients, were associated to better and worse prognosis, with not-achieved and 42 days median survival time respectively; group 1 (red line) featured patients with slightly poorer yet promising prognosis with median survival time of 3324 days; while group 2 (blue line) exhibited patients at higher risk of death from ICC with median survival time of 779 days.

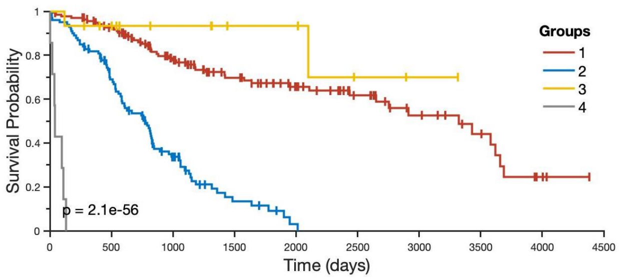

Fig. 3. Kaplan-Meier curves estimating clusters' survival probability.

Beside life expectancy, the four groups were different in terms of qualitative tumor assessment. Dimensions of tumors were found higher in groups at worse prognosis (p-value = 0.0050). Tumors with pattern 3 were associated to bad-prognosis groups (p-value = 0.0129) while tumors with pattern 1 were significantly more present in the better-prognosis groups. Groups with higher median survival time contained patients with single nodule and at-risk groups had patients with a higher number of nodules (p-value = 0.008). The number of comorbidities was significantly different among subpopulations as well, with higher values in at-risk groups (p-value = 0.0166). Additionally, the majority of patients who underwent Neoadjuvant Chemotherapy and Minor Hepatectomy were found in the better-prognosis groups (p-value = 2.072e-05) while patients mainly undergoing Major Hepatectomy without perioperative chemotherapy in the worse-prognosis groups (p-value = 0.0025). This shows that patients undergoing minor surgery without perioperative chemotherapy and those undergoing major surgery with chemotherapy have intermediate prognosis.

For what radiomic features are concerned, a ranking has been made according to their associated risk w and correspondent parameter of penalization. In Table I, the ten most relevant features have been reported, each with its own counterpart in the other view. In most of the cases, variables provided opposite contributions to the cumulative risk of death, supporting both the importance and the difference in the two regions of interest.

V. Discussion

The optimal values of the regularization terms, η for L1 and λ for L2, have been set to zero. On one hand, the null L1 sparsity penalization implies the importance of all the radiomic features in the prognosis estimating process. On the other hand, the null L2 consistency radiomic view regularization suggests the hypothesis of independence between core and margin texture analysis, as claimed in previous works [4]. Indeed, in the prediction of clinically relevant cancer

TABLE I

Weights of the ten most relevant radiomic features

Variable	Core Risk	Margin Risk
HUQ3	-0.3862	2.3952
HUmin	-0.7687	-0.2666
GLZLM_ZP	0.6417	-0.9747
GLCM_Contrast	-3.5331	4.0841
GLZLM_LZLGE	5.9590	-3.4327
NGLDM_Contrast	-0.4712	-0.3298
HUExcessKurtosis	-0.9529	0.1930
NGLDM_Coarseness	0.6010	-0.2838
NGLDM_Busyness	-0.1775	-0.1023

sub-typing, the prognostic information carried by the two views is both mandatory to consider and valuable to access. As variable-dependent risk coefficients w can be studied according to the penalization factor η, features that are more likely to survive at different levels of η are to be considered robust and important with respect to the task. Among these, we have noticed how several radiomic variables provided negative, i.e., subtractive, quantities to the patients' cumulative hazard. Interestingly, the very same variables provided a different contribution when coming from the margin of the tumor. For instance, *HUQ3*, which represents the third quartile of the CT Hounsfield values, diminishes the survival rates (risk of death) when high in the core area, since highly calcified lesions present a marked contrast enhancement and usually a decelerating growth pattern. It however enforces the survival rates when high in the margin. This means that the more accentuated this difference in the tumor-tissue interface, the more aggressive the disease, thus the poorer the prognosis of the outcome. Similar considerations can be made for *GLCM contrast*, which is the variability of the grey level co-occurrence matrix, and for *HU Excess Kurtosis*, describing how widespread are the Hounsfield values around the median. Opposite yet analogue conclusions can be drawn for *GLZLM ZP*, that measures homogeneity of the homogeneous zone, *GLZLM LZLGE*, which is the distribution of the long homogeneous zones with low grey-levels and *NGLDM coarseness*, i.e., the level of spatial change rate in intensity. These parameters could be proxies of intra-tumoral necrosis, detecting regions where Neoadjuvant Chemotherapy was effective. Additionally, also when the risk coefficient w brings the same sign in the two views, the absolute value is never equal, leading to a milder yet similar discussion. According to these findings, the two views provided different information that have proven their importance to achieve a good performance in both cancer sub-typing and survival analysis. Pertinently, a new frontier of texture analysis is currently rising, that is the delta-texture analysis (DTA) [13]. In fact, evaluating the difference between two region of interest (spatial DTA) or the same region of interest in two separate clinical time instant (temporal DTA) has been shown to be more robust in oncological predictive task. Moreover, the most undiscovered underpinnings of tumor evolution would be explained with models encompassing both delta-radiomic and genomic tumor information.

VI. Conclusions

In this work we proposed a distant supervision application for radiomics in Intrahepatic Cholangiocarcinoma patients. We performed cancer sub-typing for stratification of patients into clinically relevant subpopulations. We provided radiomic characterization of groups of patients at different risk and we assessed the different contributions of the core cancer information with respect to the margin information. Such application could pave the way to spatial delta-texture analysis in cancer research.

Acknowledgements

We thank the surgical and radiological units of the participating centers (Policlinico Rossi, Verona; Mauriziano Hospital, Torino; Gemelli Hospital, Roma; S. Orsola Hospital, Bologna; Morgagni-Pierantoni Hospital, Forlì).

References

[1] S. Rizvi and G. J. Gores, "Pathogenesis, diagnosis, and management of cholangiocarcinoma," *Gastroenterology*, vol. 145, no. 6, pp. 1215–1229, 2013.

[2] E. Scalco and G. Rizzo, "Texture analysis of medical images for radiotherapy applications," *The British journal of radiology*, vol. 90, no. 1070, p. 20160642, 2017.

[3] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[4] F. Fiz, G. Costa, N. Gennaro, L. la Bella, A. Boichuk, M. Sollini, L. S. Politi, L. Balzarini, G. Torzilli, A. Chiti *et al.*, "Contrast administration impacts ct-based radiomics of colorectal liver metastases and non-tumoral liver parenchyma revealing the "radiological" tumour microenvironment," *Diagnostics*, vol. 11, no. 7, p. 1162, 2021.

[5] L. Cavinato, N. Gozzi, M. Sollini, C. Carlo-Stella, A. Chiti, and F. Ieva, "Recurrence-specific supervised graph clustering for subtyping hodgkin lymphoma radiomic phenotypes," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 2155–2158.

[6] C. Liu, C. Wenming, S. Wu, W. Shen, D. Jiang, Z. Yu, and H. San Wong, "Supervised graph clustering for cancer subtyping based on survival analysis and integration of multi-omic tumor data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[7] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[8] C. Nioche, F. Orlhac, S. Boughdad, S. Reuzé, J. Goya-Outi, C. Robert, C. Pellot-Barakat, M. Soussan, F. Frouin, and I. Buvat, "Lifex: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity," *Cancer research*, vol. 78, no. 16, pp. 4786–4789, 2018.

[9] A. Nanashima, Y. Sumida, T. Abo, M. Oikawa, G. Murakami, H. Takeshita, H. Fukuoka, S. Hidaka, T. Nagayasu, I. Sakamoto *et al.*, "Relationship between pattern of tumor enhancement and clinicopathologic characteristics in intrahepatic cholangiocarcinoma," *Journal of surgical oncology*, vol. 98, no. 7, pp. 535–539, 2008.

[10] F. Fiz, C. Masci, G. Costa, M. Sollini, A. Chiti, F. Ieva, G. Torzilli, and L. Viganò, "Pet/ct-based radiomics of mass-forming intrahepatic cholangiocarcinoma improves prediction of pathology data and survival," *European Journal of Nuclear Medicine and Molecular Imaging*, pp. 1–14, 2022.

[11] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[12] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[13] V. Nardone, A. Reginelli, C. Guida, M. P. Belfiore, M. Biondi, M. Mormile, F. B. Buonamici, E. Di Giorgio, M. Spadafora, P. Tini *et al.*, "Delta-radiomics increases multicentre reproducibility: a phantom study," *Medical Oncology*, vol. 37, no. 5, pp. 1–7, 2020.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

16/2023 Savin, M.S.; Cavinato, L.; Costa, G.; Fiz, F.; Torzilli, G.; Vigano', L.; Ieva, F.
Distant supervision for imaging-based cancer sub-typing in Intrahepatic Cholangiocarcinoma

15/2023 Ragni, A.; Masci, C.; Ieva, F.; Paganoni, A. M.
Clustering Hierarchies via a Semi-Parametric Generalized Linear Mixed Model: a statistical significance-based approach

Bertoletti, A.; Cannistrà, M.; Diaz Lema, M.; Masci, C.; Mergoni, A.;Rossi, L.; Soncin, M.
The Determinants of Mathematics Achievement: A Gender Perspective Using Multilevel Random Forest

13/2023 Masci, C.; Cannistrà, M.; Mussida, P.
Modelling time-to-dropout via Shared Frailty Cox Models. A trade-off between accurate and early predictions

11/2023 Gatti, F.; Perotto, S.; de Falco, C.; Formaggia, L.
A positivity-preserving well-balanced numerical scheme for the simulation of fast landslides with efficient time stepping

10/2023 Corti, M.; Antonietti, P.F.; Bonizzoni, F.; Dede', L., Quarteroni, A.
Discontinuous Galerkin Methods for Fisher-Kolmogorov Equation with Application to Alpha-Synuclein Spreading in Parkinson’s Disease

09/2023 Buchwald, S.; Ciaramella, G.; Salomon, J.
Gauss-Newton oriented greedy algorithms for the reconstruction of operators in nonlinear dynamics

08/2023 Bonizzoni, F.; Hu, K.; Kanschat, G.; Sap, D.
Discrete tensor product BGG sequences: splines and finite elements

06/2023 Artoni, A.; Antonietti, P. F.; Mazzieri, I.; Parolini, N.; Rocchi, D.
A segregated finite volume - spectral element method for aeroacoustic problems

07/2023 Garcia-Contreras, G.; Còrcoles, J.; Ruiz-Cruz, J.A.; Oldoni, M; Gentili, G.G.; Micheletti, S.; Perotto, S.
Advanced Modeling of Rectangular Waveguide Devices with Smooth Profiles by Hierarchical Model Reduction