IL NUOVO CIMENTO **47 C** (2024) 267
DOI 10.1393/ncc/i2024-24267-8

COMMUNICATIONS: SIF Congress 2023

Artificial Intelligence-assisted thyroid cancer diagnosis from Raman spectra of histological samples

L. BELLANTUONO(1)(2)(*)

(1) *Università degli Studi di Bari Aldo Moro, Dipartimento di Biomedicina Traslazionale e Neuroscienze (DiBraiN) - Bari, I-70124, Italy*
(2) *Istituto Nazionale di Fisica Nucleare, Sezione di Bari - Bari, I-70125, Italy*

received 31 January 2024

Summary. — Raman spectroscopy emerges as a highly promising diagnostic tool for thyroid cancer due to its capacity to discern biochemical alterations during cancer progression. This non-invasive and label/dye-free technique exhibits superior efficacy in discriminating malignant features compared to traditional molecular tests, thereby minimizing unnecessary surgeries. Nevertheless, a key challenge in adopting Raman spectroscopy lies in identifying significant patterns and peaks. This study proposes an artificial intelligence approach for distinguishing healthy/benign from malignant nodules, ensuring interpretable outcomes. Raman spectra from histological samples are collected, and a set of peaks is selected using a data-driven, label-independent approach. Machine Learning algorithms are trained based on the relative prominence of these peaks, achieving performance metrics with an area under the receiver operating characteristic curve exceeding 0.9. To enhance interpretability, eXplainable Artificial Intelligence (XAI) is employed to compute each feature's contribution to sample prediction.

1. – Introduction

Thyroid cancer, marked by the malignant growth of thyroid gland cells, stands as the most prevalent malignancy within the endocrine system. Its incidence rates, among the top ten cancers globally, have risen due to enhanced diagnostic capabilities. The most common types are represented by papillary thyroid carcinoma (PTC), follicular carcinoma (FC), and the follicular variant of papillary thyroid carcinoma (FV-PTC). PTC, predominant in younger individuals, constitutes about 80% of cases, while FC, affecting older individuals, represents 10–15%. FV-PTC poses diagnostic challenges due to its follicular growth pattern. Early-stage detection thyroid cancer entails a high 5-year relative survival rate, emphasizing the need for accurate diagnostic tools [1].

(*) E-mail: loredana.bellantuono@uniba.it

Clinical challenges encompass the surge in thyroid nodule detection, managing cytologically indeterminate nodules, and inter-observer diagnosis variability [2-7]. Molecular analyses aim to minimize unnecessary surgeries, yet their predictive value remains limited [8-12]. In this context, Raman spectroscopy, a technique that recognizes vibrational fingerprints of molecules through Raman scattering, emerges as a promising tool to identify thyroid neoplastic lesions [13], considering recent evidence in support of its ability to detect biochemical changes during oncogenesis, along with its non-invasiveness [14-19].

Despite its potential, Raman spectroscopy faces both practical and conceptual challenges: correctly interpreting spectral characteristics and overcoming diagnostic variability require a standardized workflow. Artificial intelligence, particularly machine learning, emerges as a strategic solution to this kind of problems, since it automates classification workflows, ensures uniform diagnostic criteria, and adapts to evolving datasets. Application of machine learning to analyze a dataset of Raman spectra from histological samples for thyroid cancer diagnosis has been proposed in ref. [17], while analogous methods have shown promising in various other types of cancer. The study presented in this paper advances prior research by developing supervised predictive models capable of classifying new spectra, not included in the training datasets, additionally exploring the fingerprints of thyroid cancer through quantitative procedures, thus shedding light on potential biomarkers [20]. The combination of global and local approaches, employing the Boruta method and eXplainable Artificial Intelligence (XAI), enhances model informativeness, generalization and transparency, aiming to make the decision process as intelligible as possible [21, 22], which is especially relevant in realistic scenarios [23-30]. In the following sections, we provide an overview of the main findings of this work and of the methodology employed therein.

2. – Results

The comprehensive study presented herein unfolds through three fundamental blocks outlined in fig. 1: the clinical step, involving patient enrollment, thyroid gland surgery, and pathological evaluation; the spectroscopy step, where Raman spectra associated with each histological sample are obtained; the artificial intelligence step, where first machine learning is used to distinguish between spectra associated to healthy/benign and cancerous tissue, and then the most influential features are identified through XAI.

2˙1. *Data and feature engineering.* – The analyzed dataset includes 59 Raman spectra from histological samples, excised from thyroids with suspected cancer, which provide the basis for training machine learning algorithms. Study population included 54 subjects (34 females, 20 males) affected by thyroid nodular pathology, with age distribution centered at 46.3 years, with a 11.2 years standard deviation. Raman spectra were collected by utilizing the extended scan mode across the 100–3600 cm^{-1} Raman-shift range, with a spectral resolution of approximately 1 cm^{-1}. Further details of the enrolling procedure and the Raman spectroscopy stage are reported in ref. [20].

Feature engineering begins with a preprocessing stage, where spectra are initially interpolated, normalized and smoothed through cubic spline. Then, a univariate Gaussian mixture model is used to identify for peak detection, leading to the identification of 32 peak values in the spectra, each characterized by a central value and a width. The procedure is refined by merging partially overlapping peaks, which results in 29 non-overlapping selected intervals [20]. To compare our results with established knowledge on the topic, we pay specific attention to the interval including the peaks correspond-

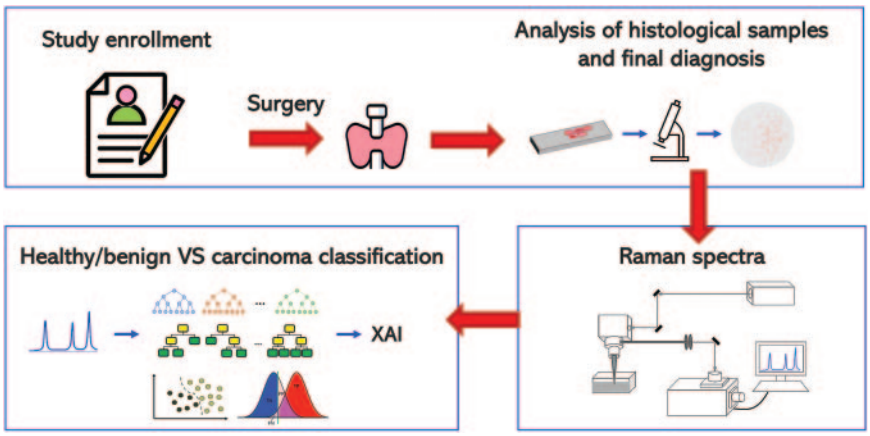

Fig. 1. – General workflow of the analysis [20].

ing to the following Raman shifts, expressed in cm^{-1}: 747, 1125, 1302, 1584, associated to reduced cytochrome c; 1003, 1155, 1516, associated to carotenoids; 1376, associated to oxidised cytochrome b; 1638, associated to oxidised cytochrome c. Raman spectra characteristics are distilled into features based on the highest intensity value (prominence) in each of the 29 intervals. In particular, the obtained feature set consists of 406 not redundant ratios between prominences, allowing for a detailed representation of the spectral structure. The choice of these features is made by evaluating the distributions of a ratio and its reciprocal on the entire dataset and keeping the quantity characterized by the largest mean value compared to the standard deviation.

Finally, a wrapper method for feature selection based on the Boruta framework [31] is used to mitigate noise and data redundancy effects. The procedure selects the features that are uncorrelated with each other and significantly improve the performance of a supervised learning Random Forest algorithm, exploiting the idea that training sample randomization and system perturbation help to soften the negative impact on the algorithm determined by random fluctuations and correlations.

2˙2. *Machine learning workflow.* – The artificial intelligence workflow implemented in our study, displayed in fig. 2, consists of two nested loops, one for Synthetic Minority Over-sampling TEchnique (SMOTE), which generates new instances by synthesizing them from the samples already existing in the minority class [32] and another for a leave-one-out classification procedure. In this study, the oversampling of the minority class is performed within each leave-one-out iteration.

The described process ensures robust model training despite the limited dataset size and class imbalance. Various machine learning algorithms, including Random Forest [33], XGBoost [34], Support Vector Machine [35], and Gaussian Naïve Bayes [36], are explored in search of the best performance in distinguishing healthy/benign from carcinoma spectra. Specifically, we focus on the Receiver Operating Characteristic (ROC) curve, which lies in a plane whose axis are the false positive rate and true positive rate. Then, we quantify the performance of each classifier in terms of the area under the curve (AUC). The analysis reveals that the best-performing algorithm, with median AUC 0.9441 and interquartile range 0.0049 over 100 SMOTE runs, is the Random Forest classifier with n_estimators = 50, max_depth = 5, criterion = “entropy”. Even by optimizing their

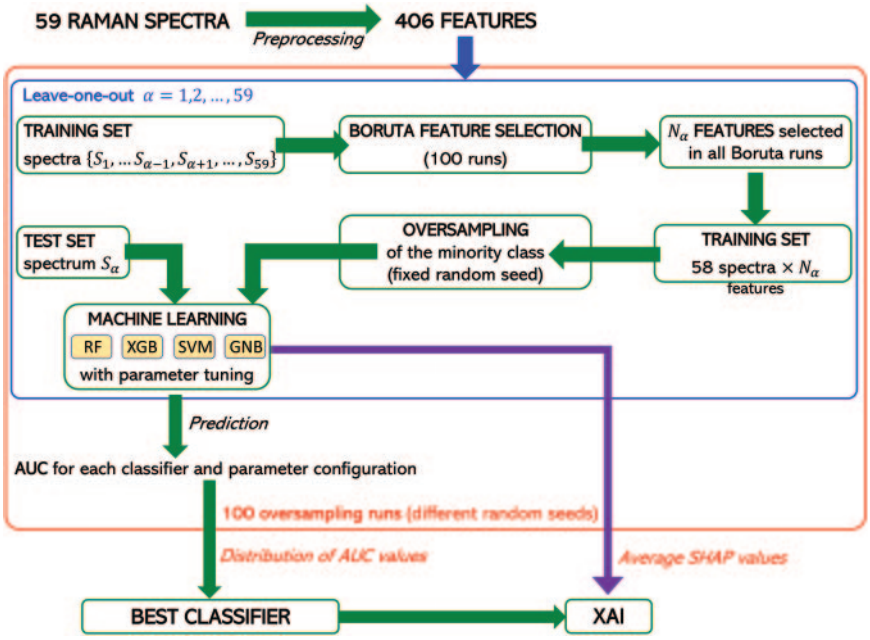

Fig. 2. – Detailed workflow of the machine learning and eXplainable Artificial Intelligence (XAI) analysis [20]. After preprocessing, 100 runs of the Synthetic Minority Over-sampling TEchnique (SMOTE) with different random seeds are executed. In each SMOTE run, a leave-one-out classification is implemented, and in the i-th leave-one-out iteration (where i ranges from 1 to 59) the Boruta algorithm selects N_i relevant features, that are used to construct the training set; then, before implementing different machine learning algorithms, SMOTE is applied to oversample the minority class. The classification algorithms employed in this study are Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM) and Gaussian Naïve Bayes (GNB). Their performances are quantified by the AUC metrics, which is the area under the Receiver Operating Characteristic (ROC) curve. The impact of features on the prediction provided by the best performing classifier for each instance is evaluated through the Shapley (SHAP) values, averaged over all SMOTE runs.

internal parameters, the performances of other algorithms are worse, with median AUC strictly smaller than 0.094 (see ref. [20] for details). The optimal classifier, a Random Forest algorithm with specific internal parameters (n_estimators, max_depth, criterion), demonstrated superior performance in distinguishing between healthy/benign and carcinoma spectra. The ROC curves shown in fig. 3 elucidate the classifier's effectiveness, with Random Forest outperforming other algorithms.

Once we have selected the optimal classifier, we use the related ROC curve to optimize the classification threshold in order to maximize the geometric mean of sensitivity and specificity. The normalized confusion matrix related to the classification of 59 samples with the optimal threshold of 0.5 underlines the model capacity to discern between diagnostic categories, with the true and the predicted label coinciding in 91.0% cases for healthy/benign tissue, and 91.1% cases for cancerous tissue. To expand the model applicability, we explore the retroactive use of prediction probabilities for discerning diagnostic subcategories. Median prediction probabilities indicated a hierarchy among diagnostic labels, showing consistency with previous literature [17]: 0 (with IQR 0.12) for Healthy,

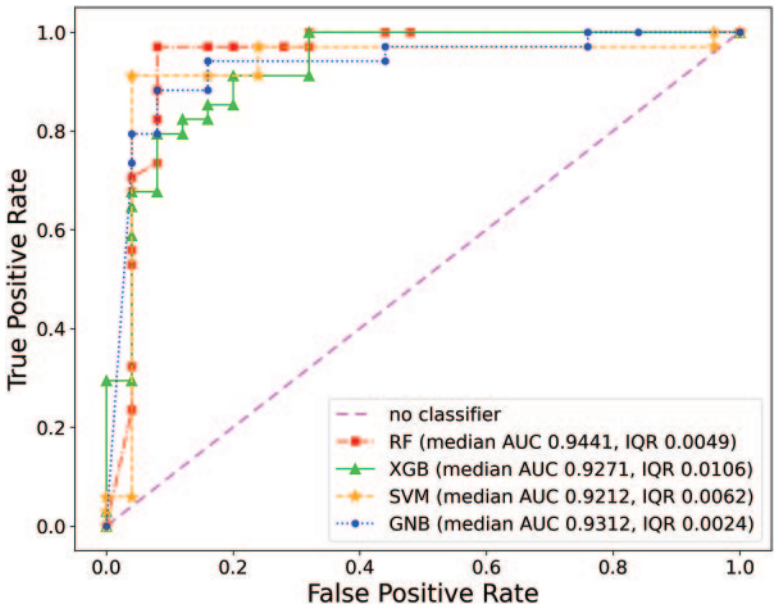

Fig. 3. – Receiver Operating Characteristic (ROC) curves for one of the Random Forest (RF) classifiers that maximize median AUC (n_estimators = 50, max_depth = 5, criterion = "entropy"), for XGBoost (XGB) and Support Vector Machine (SVM) algorithms with arbitrary internal parameters and for the Gaussian Naïve Bayes (GNB) algorithm. Plots referred to XGB and SVM have been obtained in the configurations num_parallel_tree = 100, max_depth = 3, n_jobs = 1, and $c = 1$, kernel = "entropy", respectively. The true positive rate and false positive rate coordinates of points in the ROC curves are median values computed over 100 SMOTE runs [20].

0.04 (with IQR 0.36) for Benign, 0.80 (with IQR 0.45) for FC, 0.92 (with IQR 0.16) for FV-PTC, 0.96 (with IQR 0.22) for PTC.

2˙3. *eXplainable Artificial Intelligence analysis.* – To enhance the interpretability of the model, an XAI analysis was conducted, focusing on the best-performing Random Forest classifier. SHapley Additive exPlanations (SHAP) values were computed, providing insights into the impact of different features on prediction outcomes. Specifically, for each given feature, the SHAP is computed by evaluating the difference between the model output's prediction with and without the considered feature, averaging over all possible subsets of features [37, 38]. From the XAI analysis, it emerges that the 5 most influential features among the ratios P_j/P_k of peak prominences in the intervals j and k are the following ones:

- P_{24}/P_{11}, with the interval #24 containing the line 1376 cm^{-1} (oxidised cytochrome b) and the interval #11 not containing lines associated with known categories of molecules; higher values drive the classifier towards the healthy/benign prediction.
- P_{29}/P_{17}, with the interval #29 containing the line 1638 cm^{-1} (oxidised cytochrome c) and the interval #17 containing the line 1155 cm^{-1} (carotenoids); higher values drive the classifier towards the healthy/benign prediction.
- P_8/P_{20}, with the intervals #8 and #20 not containing lines associated with known

categories of molecules; higher values drive the classifier mostly towards the cancer prediction.

- P_{29}/P_{18}, with the interval #29 containing the line $1638\,\mathrm{cm}^{-1}$ (oxidised cytochrome c) and the interval #18 not containing lines associated with known categories of molecules; higher values drive the classifier towards the healthy/benign prediction.
- P_2/P_{17}, with the interval #2 not containing lines associated with known categories of molecules and the interval #17 containing the line $1155\,\mathrm{cm}^{-1}$ (carotenoids); higher values drive the classifier towards the healthy/benign prediction.

An extended list is reported in ref. [20]. We discuss the relevance of the above findings in the next section.

3. – Discussion

The developed AI model exhibited satisfactory accuracy in distinguishing between thyroid spectra associated to healthy/benign and cancerous tissue. The interpretability of the model was further enhanced through XAI analysis, shedding light on the features that are pivotal to the classification process. Notably, the dominance of carotenoid lines in carcinoma spectra and the prevalence of oxidized cytochrome b and c lines in healthy/benign spectra emerged as influential patterns. These findings align with existing knowledge, underlining the model capability to discern molecular signatures associated with thyroid cancer.

The clinical implications of our study are profound. The ability to accurately classify thyroid lesions through Raman spectroscopy and AI introduces a non-invasive, label-free diagnostic approach. By leveraging characteristic molecular fingerprints, this methodology can potentially mitigate unnecessary surgeries and enhance diagnostic precision. The transparency in model predictions, facilitated by XAI, addresses a crucial aspect in the clinical adoption of AI tools. Clinicians can gain insights into the features driving predictions, fostering trust and facilitating informed decision-making.

While our study presents promising outcomes, certain limitations merit consideration. The dataset, though meticulously curated, is relatively small and exhibits an imbalance between healthy/benign and carcinoma spectra. This calls for cautious generalization of the model performance to diverse populations. Additionally, the current model primarily focuses on distinguishing between healthy/benign and cancerous spectra, overlooking finer categorizations within carcinoma types. Future research endeavors should prioritize expanding the dataset, ensuring diverse representation, and refining the model to discern between specific carcinoma subtypes. In particular, new data are expected to come from cytological spectra of samples acquired through fine needle aspiration, representing a less invasive technique that does not require surgery.

The application of artificial intelligence in the clinical setting necessitates a seamless integration into existing diagnostic workflows. Overcoming the practical and conceptual barriers associated with Raman spectroscopy requires collaborative efforts between computer scientists, spectroscopists, and clinicians. The development of user-friendly interfaces and standardized workflows can pave the way for the seamless adoption of machine learning and XAI tools in thyroid cancer diagnostics. The ambition of this research line is the implementation of an apparatus designed for clinical environment, which can generate spectra from samples and recognize the fingerprints related to oncogenesis. To develop such a support system in a clinical context, we believe that it would be beneficial

integrating the procedure with a further classifier, that would examine clinical data of a different type, such as the dimensional parameters that are determined from thyroid ultrasound scans.

The potential clinical utility of our artificial intelligence model extends beyond the confines of thyroid cancer diagnostics. The successful application of similar methodologies in diverse cancer types highlights the versatility of Raman spectroscopy coupled with artificial intelligence, and the role of the latter as an indispensable ally in the realm of precision medicine. The development of specialized diagnostic tools for each cancer type holds the promise of revolutionizing early detection and personalized treatment strategies.

* * *

The author acknowledges Raffaele Tommasi, Ester Pantaleo, Martina Verri, Nicola Amoroso, Pierfilippo Crucitti, Michael Di Gioacchino, Filippo Longo, Alfonso Monaco, Anda Mihaela Naciu, Andrea Palermo, Chiara Taffon, Sabina Tangaro, Anna Crescenzi, Armida Sodo and Roberto Bellotti for collaboration. The study protocol adhered to the Declaration of Helsinki and to the International Conference on Harmonization Good Clinical Practice and received approval by the Ethical Committee of the "Fondazione Policlinico Universitario Campus Bio-Medico" (UCBM) (prot. 33.15 TS ComEt CBM and 31/19 PAR ComEt CBM from 26th July 2019). All participants granted informed consent. Enrolled patients were recorded in a codified file with an anonymous ID code, which was registered in the software database of the Endocrine Organs and Neuromuscolar Pathology Unit of the UCBM.

REFERENCES

[1] NIH National Cancer Institute, *Thyroid Cancer – Cancer Stat Facts*, `https://seer.cancer.gov/statfacts/html/thyro.html`, accessed: 22 June 2023 (2023).
[2] Vaccarella S. *et al.*, *N. Engl. J. Med.*, **375** (2016) 614.
[3] Rusinek D. *et al.*, *Int. J. Mol. Sci.*, **18** (2017) 1817.
[4] Patel K. N. *et al.*, *Ann. Surg.*, **271** (2020) e21.
[5] Alyami J. *et al.*, *Medicine*, **101** (2022) e31106.
[6] Elsheikh T. M. *et al.*, *Am. J. Clin. Pathol.*, **130** (2008) 736.
[7] Trimboli P. *et al.*, *Endocr. Pathol.* (2022) 457.
[8] McMurtry V. *et al.*, *Diagn. Cytopathol.*, **51** (2023) 36.
[9] Livhits M. J. *et al.*, *JAMA Oncol.*, **7** (2021) 70.
[10] Agarwal S. *et al.*, *Cancers*, **14** (2022) 204.
[11] Valderrabano P. *et al.*, *JAMA Otolaryngol. Head Neck Surg.*, **145** (2019) 783.
[12] DiGennaro C. *et al.*, *Thyroid*, **32** (2022) 1144.
[13] Krafft C. and Popp J., *Anal. Bioanal. Chem.*, **407** (2015) 8263.
[14] Teixeira C. S. B. *et al.*, *Analyst*, **134** (2009) 2361.
[15] Li Z. *et al.*, *Laser Phys. Lett.*, **11** (2014) 045602.
[16] Rau J. V. *et al.*, *Sci. Rep.*, **7** (2017) 1.
[17] Sbroscia M. *et al.*, *Sci. Rep.*, **10** (2020) 1.
[18] Sodo A. *et al.*, *Diagnostics (Basel)*, **11** (2020) 43.
[19] Palermo A. *et al.*, *J. Clin. Endocrinol. Metab.*, **107** (2022) 3309.
[20] Bellantuono L. *et al.*, *Sci. Rep.*, **13** (2023) 16590.
[21] Flach P., *Proc. AAAI Conf. Artif. Intell.*, **33** (2019) 9808.
[22] Vollmer S. *et al.*, *BMJ*, **368** (2020) l6927.
[23] Lombardi A. *et al.*, *Brain Inform.*, **9** (2022) 1.
[24] Lombardi A. *et al.*, *App. Sci.*, **12** (2022) 7227.

[25] Bellantuono L. *et al.*, *Front. Big Data*, **5** (2022) 1027783.
[26] Jiménez-Luna J. *et al.*, *Nat. Mach. Intell.*, **2** (2020) 573.
[27] Miller T., *Artif. Intell.*, **267** (2019) 1.
[28] Bussmann N. *et al.*, *Front. Artif. Intell.*, **3** (2020) 26.
[29] Bellantuono L. *et al.*, *Sci. Rep.*, **13** (2023) 839.
[30] Lacalamita A. *et al.*, *Int. J. Mol. Sci.*, **24** (2023) 15286.
[31] Kursa M. and Rudnicki W., *J. Stat. Software*, **36** (2010) 1.
[32] Chawla N. V. *et al.*, *J. Artif. Intell. Res.*, **16** (2002) 321.
[33] Breiman L., *Mach. Learn.*, **45** (2001) 5.
[34] Chen T. and Guestrin C., *XGBoost: A Scalable Tree Boosting Sytem*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY) 2016, pp. 785–794.
[35] Cortes C. and Vapnik V., *Mach. Learn.*, **20** (1995) 273.
[36] Paul A. *et al.*, *IEEE Trans. Image Process.*, **27** (2018) 4012.
[37] Lundberg S. and Lee S., *A unified approach to interpreting model predictions*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates, Inc.) 2017, pp. 44768–4777.
[38] Lundberg S. *et al.*, *Nat. Mach. Intell*, **2** (2020) 56.