Artificial Intelligence-assisted Raman Spectroscopy for Liver cancer diagnosis

Concetta Esposito[1,2], *Mohammed* Janneh[1,2], *Sara* Spaziani[1,2], *Vincenzo* Calcagno[1,2], *Mario Luca* Bernardi[2,3], *Martina* Iammarino[2,3], *Chiara* Verdone[2,3], *Maria* Tagliamonte[2,4], *Luigi* Buonaguro[2,4], *Marco* Pisco[1,2,*], *Lerina* Aversano[2,3], and *Andrea* Cusano[1,2]

[1]Optoelectronic Division-Engineering Department, University of Sannio, 82100 Benevento, Italy
[2]Centro Regionale Information Communication Technology (CeRICT Scrl), 82100 Benevento, Italy
[3]Informatics Group, Engineering Department, University of Sannio, 82100 Benevento, Italy
[4]National Cancer Institute-IRCCS "Pascale", Via Mariano Semmola, 52, 80131 Napoli, Italy

Abstract. Hepatocellular carcinoma (HCC), the most common form of primary liver cancer, represents a global health challenge due to its complexity and the limitations of current diagnostic techniques. By combining Raman spectroscopy and Artificial Intelligence (AI), we have succeeded in classifying tumor cells. In fact, we have performed a first Raman spectral analysis based on the characterization and differentiation between uncultured primary human liver cells derived from resected HCC tumor tissue and the adjacent non-tumor counterpart. Biochemical analysis of the collected Raman spectra revealed that there is more DNA in the nuclei of the tumor cells than in non-tumor cells. We then develop three machine learning approaches, including multivariate models and neural networks, to rapidly automate the recognition and classification of the Raman spectra of both cells. To evaluate the performance of the developed AI models, we prepared and analyzed two additional cell samples with a ratio of 4:1 and 3:1 between tumor and non-tumor cells and compared the obtained results with the nominal percentages (accuracy of 80 and 60%, respectively). These results confirm that the models are able to make classifications at the level of a single spectrum, indicating the possibility of rapidly analysing and classifying a primary HCC cell.

1 Introduction

Hepatocellular carcinoma (HCC), the most common form of primary liver cancer, is a global health issue with a high mortality rate [1]. The limitations of current clinical imaging and histopathological methods play a crucial role in diagnosis of HCC. Over the last decade, Raman spectroscopy (RS) has emerged as a powerful vibrational technique in the field of cancer diagnosis, demonstrating its ability to characterize and distinguish healthy from cancer cells/tissues based on their different biochemical composition. RS is a label-free, non-invasive and non-destructive approach which, combined with its chemical specificity, provides an important advantage to obtain the spectral information associated with the biochemical fingerprint of the biological samples. In our study, RS was applied to investigate the biochemical composition of uncultured primary human non-tumor and tumor cells from a patient with HCC. However, the complexity and weakness of the Raman signal pose a great challenge for its analysis and interpretation and require the use of Artificial Intelligence (AI) algorithms [2]. Several studies reported the powerful combination of RS and machine learning (ML) to identify and discriminate between non-tumor and tumor spectral patterns for cell classification [3]. To this end, we have developed a combined approach based on RS and supported by different AI models, namely two different LDA-based methods and a neural network model. In our work, we have shown that RS, supported by AI, is an important tool for rapid and automated identification and classification of the Raman spectra of tumor cells to improve the diagnosis of HCC.

2 Materials and Methods

Non-tumor and tumor liver tissue was obtained from a patient with HCC. After dissociation into single cells, uncultured non-tumor and tumor cells and two additional samples with different ratios of non-tumor and tumor cells, namely Mix1 (20/80%) and Mix2 (40/60%), were fixed and plated on CaF_2 slides.
Raman measurements were performed using the LabRAM HR Nano from HORIBA with a laser wavelength of 532 nm and a 100× air objective (N.A. 0.90). The Raman spectra were pre-processed by subtracting the background and baseline, vector normalization and removing outliers by partial least square (PLS) regression. In addition, three supervised ML models were developed to recognize and classify Raman spectra. Two ML approaches are based on Linear Discriminant Analysis (LDA), namely Hyper-parameters tuned-LDA and principal component analysis (PCA)-

* Corresponding author: pisco@unisannio.it

LDA. In addition, a model based on a convolutional neural network (CNN) and a recurrent network based on Long-Short Term Memory (LSTM) cells, the CNN-LSTM classification model, has been developed.

3 Results

Raman measurements were performed on the nucleus of uncultured primary liver cells and the molecular composition of non-tumor and tumor samples was determined (**Fig.1a**). In particular, we analyzed the two mean normalized Raman spectra and assigned the Raman spectral peaks to the corresponding cellular components such as nucleic acids, proteins and lipids. Differential analysis revealed that the negative peaks at 785, 1094, 1335 and 1578 cm^{-1} were assigned to nucleic acids, especially DNA (**Fig.1b**). The Raman result confirmed that there is more DNA in the nuclei of cancer cells than in non-tumor cells, which is consistent with the altered nuclear ploidy of HCC cells [4]. In addition, RS was able to show the main biochemical changes between cancer and healthy cells and emphasisesthe link between HCC and high nuclear DNA content [5].

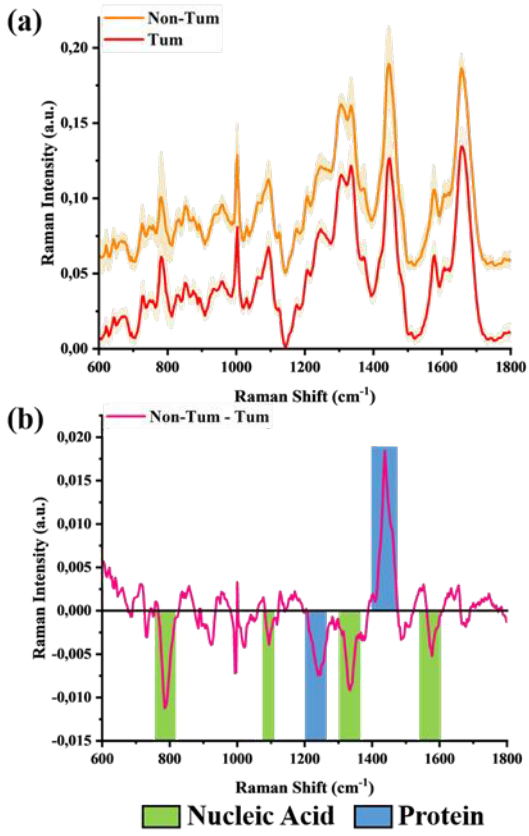

Fig. 1. Averaged Raman spectra of non-tumor and tumor cells (a). Difference between the averaged Raman spectra (b). Raman peaks associated with nucleic acids and proteins are highlighted in green and blue, respectively.

Subsequently, the development of different AI models was necessary to accurately classify non-tumor and tumor cells. First, an LDA model was created using the processed Raman spectra (80% training and 20% test set) and then optimized using Hyper-parameter tuning based on a grid search method. For confirmation, a PCA-LDA model was created with the first 30 selected principal components (PCs). Neural models based on the CNN-LSTM network were also developed, with the training phase applied directly to the Raman spectra and including the Hyper-parameter optimization phase. Accurate identification and classification of tumor cells is a crucial aspect, especially for real samples with a different ratio of tumor and non-tumor cells. To demonstrate the robustness of our integrated approach, we perform blind prediction with the three ML-based Raman classification models in the case of mixed samples. The investigated samples have the following ratios of tumor to non-tumor cells : 5 to 0, 4 to 1 (Mix1) and 3 to 2 (Mix2). As can be seen in **Table 1**, the best CNN-LSTM model achieved a classification accuracy of almost 93% for the tumor cell spectra. Finally, the results show that the AI models assisted RS have high predictive power, with a classification accuracy of almost 80% and 60% for the tumor cell spectra for Mix1 and 2, respectively, which are given as nominal values in **Table 1**.

Table 1. Classification of tumor and Mixes spectra samples.

	Tum (% Tum)	Mix1 (% Tum)	Mix2 (% Tum)
Nominal value	100.00%	80.00%	60.00%
Hyper-parameter tuned LDA	87.00%	**80.10%**	62.30%
PCA-LDA	89.00%	82.40%	58.00%
CNN-LSTM-22	91.60%	82.76%	58.33%
CNN-LSTM-56	**92.70%**	81.67%	**61.54%**

4 Conclusion

AI methods in combination with RS enabled rapid and specific classification of liver cancer cell spectra and provided valid support for HCC diagnosis.

References

1. A. Villanueva, A. N. Engl. J. Med. 380, 1450–1462 (2019)
2. R. Luo, J. Popp, T. Bocklitz. Analytica. **3**, 287–301 (2022)
3. S. Elumalai, S. Managó, A.C. De Luca. Sensors (Switzerland). **20**, 1–19 (2020)
4. M. Bou-Nader, S. Caruso, R. Donne, S. Celton-Morizur, J. Calderaro, G. Gentric, M. Cadoux, A. L'Hermitte, C. Klein, T. Guilbert, T. et al. Gut. **69**, 355–364 (2020)
5. C. Esposito, M. Janneh, S. Spaziani, V. Calcagno, M. L. Bernardi, M. Iammarino, C. Verdone, M. Tagliamonte, L. Buonaguro, M. Pisco, L. Aversano and A. Cusano. Cells. **12**, 2645 (2023)