Vilnius University

Machine Learning-Based Diagnosis of Cancer and Fibrosis with Second Harmonic Generation Microscopy

Yaraslau Padrez

DOCTORAL DISSERTATION
2025

Natural Sciences
Physics **N 002**

https://doi.org/10.15388/vu.thesis.816
https://orcid.org/0000-0003-0852-6579

VILNIUS UNIVERSITY
CENTER FOR PHYSICAL SCIENCES AND TECHNOLOGY

Yaraslau Padrez

Machine Learning-Based Diagnosis of Cancer and Fibrosis with Second Harmonic Generation Microscopy

DOCTORAL DISSERTATION

Natural sciences,
Physics (N 002)

VILNIUS 2025

This dissertation was prepared between 2021 and 2025 (State research institute Center for Physical Sciences and Technology). The research was supported by the Research Council of Lithuania with scholarships that were granted for academic accomplishments twice: for the years 2022 (reg. Nr. P-DAP-23-60) and 2025 (reg. Nr. P-DAP-25-153), and financial support for research visit to the University of Eastern Finland (Joensuu, Finland) for four weeks during the first semester in 2022 (reg. Nr. P-DAK-22-45).

Academic Supervisor – Dr. Renata Karpicz (State Research Institute Center for Physical Sciences and Technology, Natural Sciences, Physics – N 002).
Academic Consultant – Dr. Danielis Rutkauskas (State Research Institute Center for Physical Sciences and Technology, Natural Sciences, Physics – N 002).

This doctoral dissertation will be defended in a public meeting of the Dissertation Defense Panel:
Chairman: Prof. Dr. Vitalijus Karabanovas (National Cancer Institute, Natural Sciences, Physics – N 002).
Members:
Prof. Dr. Raquel Cuevas-Diaz Duran (Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Natural sciences, Biology – N 010),
Prof. Dr. Justinas Čeponkus (Vilnius University, Natural Sciences, Physics – N 002),
Doc. Dr. Andrius Gelžinis (Vilnius University, Natural Sciences, Physics – N 002),
Dr. Ilze Irbe (University of Latvia, Natural Sciences, Biophysics – N 011).

The dissertation shall be defended at a public meeting of the Dissertation Defense Panel at 11:00 on the 17th of September 2025 in Room D401 of the State research institute Center for Physical Sciences and Technology. Address: Sauletekio av. 3, Vilnius, Lithuania
Tel. +370 5 264 8884; e-mail: office@ftmc.lt

The text of this dissertation can be accessed at the libraries of the Vilnius University, as well as on the website of Vilnius University: www.vu.lt/naujienos.ivykiu-kalendorius

https://doi.org/10.15388/vu.thesis.816
https://orcid.org/0000-0003-0852-6579

VILNIAUS UNIVERSITETAS
FIZINIŲ IR TECHNOLOGIJOS MOKSLŲ CENTRAS

Yaraslau Padrez

Mašininiu mokymusi grindžiama vėžio ir fibrozės diagnostika taikant antrosios harmonikos generavimo mikroskopiją

DAKTARO DISERTACIJA

Gamtos mokslai,
Fizika (N 002)

VILNIUS 2025

Disertacija rengta 2021–2025 metais Fizinių ir technologijos mokslų centre.
Mokslinius tyrimus rėmė Lietuvos mokslo taryba skirdama stipendijas 2022 (reg. Nr. P-DAP-23-60) ir 2024 (reg. Nr. P-DAP-25-153) metais už akademinius pasiekimus, bei 2022 paskirdamas finansavimą stažuotei Rytų Suomijos universitete (Suomijoje) (reg. Nr. P-DAK-22-45).

Mokslinė vadovė – dr. Renata Karpicz (Fizinių ir technologijos mokslų centras, gamtos mokslai, fizika – N 002).
Mokslinis konsultantas – dr. Danielis Rutkauskas (Fizinių ir technologijos mokslų centras, gamtos mokslai, fizika – N 002).

Gynimo taryba:
Pirmininkas: prof. dr. Vitalijus Karabanovas (Nacionalinis vėžio institutas, gamtos mokslai, fizika – N 002).
Nariai:
Prof. dr. Raquel Cuevas-Diaz Duran (Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Meksika, gamtos mokslai, biologija – N 010),
Prof. dr. Justinas Čeponkus (Vilniaus universitetas, gamtos mokslai, fizika – N 002),
Doc. dr. Andrius Gelžinis (Vilniaus universitetas, gamtos mokslai, fizika – N 002),
Dr. Ilze Irbe (Latvijos universitetas, gamtos mokslai, biofizika – N 011).

Disertacija ginama viešame Gynimo tarybos posėdyje 2025 m. rugsėjo mėn. 17 d. 11:00 val. Valstybinio mokslinių tyrimų instituto Fizinių ir technologijos mokslų centro D401 konferencijų salėje. Adresas: Saulėtekio al. 3, Vilnius, Lietuva), tel. +370 5 264 8884; el. paštas office@ftmc.lt

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir Vilniaus universiteto interneto svetainėje adresu: https://www.vu.lt/naujienos/ivykiu-kalendorius

Acknowledgments

First and foremost, I would like to express my deepest gratitude to all those people who have supported, inspired and guided me on this challenging but rewarding journey.

My sincere gratitude goes to Elena Guseinova, my first physics teacher, who laid the foundation for my scientific path with her belief in me and her early encouragement.

I am grateful to my first academic supervisor, Dr. Dzmitry Bychanok, for introducing me to the fascinating world of research. His guidance and enthusiasm showed me that science can be both engaging and inspiring.

I would also like to thank Dr. Polina Kuzhir, whose recommendation led me to Vilnius. Her management of the laboratory in Minsk, where I started my research, played an important role in shaping my early scientific endeavors.

Special thanks go to the research group from Romania led by Dr. Radu Hristu. Their active participation in discussions, joint publications and general support greatly enriched a large part of this dissertation.

My growth as a *Python* developer has been accompanied by useful tutorials from Dr. Igor Timoshchenko, ranging from the simplest to the extremely complex.

I thank my academic supervisor, Dr. Renata Karpicz, and my academic advisor, Dr. Danielis Rutkauskas, for their academic supervision and guidance throughout this process.

I owe immeasurable gratitude to Dr. Lena Golubewa. Her tireless support from the beginning to the completion of this dissertation was simply exceptional. Without her expertise, dedication and countless contributions, this work — and the numerous articles that followed — would not have been possible.

ABBREVIATIONS AND NOTATIONS

AI – Artificial intelligence
ATC – Anaplastic thyroid carcinoma
AUC – Area under the receiver operating characteristic curve
aC – Correlation based on gray level co-occurrence matrix
CAF – Cancer-associated fibroblasts
CARS – Coherent anti-Stokes Raman scattering
CHI – Calinski-Harabasz index
CMOS – Complementary metal–oxide–semiconductor
CNN – Convolutional neural network
C-SVC – C-Support vector classification
DAB – Diaminobenzidine
DBI – Davies-Bouldin index
DFT – Discrete Fourier transform
DL – Deep learning
E – Energy based on gray level co-occurrence matrix
ECM – Extracellular matrix
ERR – Experimental error
FFT – Fast Fourier transform
FN – False negatives
FOS – First-order statistics
FP – False positives
FTC – Follicular thyroid carcinoma
GLCM – Gray level co-occurrence matrix
GLN – Gray-level non-uniformity based on gray level run length matrix
GLRLM – Gray level run length matrix
H – Entropy based on gray level co-occurrence matrix
H&E – Hematoxylin and eosin
HOS – Higher-order statistics
I – Inertia based on gray level co-occurrence matrix
IARC – International agency for research on cancer
IE – Index of expression
IHC – Immunohistochemistry
IQR – Interquartile range
L – Local homogeneity based on gray level co-occurrence matrix
LightGBM – Light gradient-boosting machine
LR – Logistic regression
LRE – Long run emphasis based on gray level run length matrix
MCT – Monocrotaline
MI – Mutual information
MIFS – Mutual information feature selector
ML – Machine learning
MLP – Multi-layer perceptron
MMP – Matrix metalloproteinase
MRI – Magnetic resonance imaging

MTC – Medullary thyroid carcinoma
OI – Orientation index
PAH – Pulmonary arterial hypertension
PC – Principal component
PCA – Principal component analysis
PET – Positron emission tomography
PIA – Permutation importance analysis
PSG – Polarization state generator
PSHG – Polarization-resolved second harmonic generation
PTC – Papillary thyroid carcinoma
R2 – Coefficient of determination
RF – Random forest
RFECV-LinearSVC – Recursive feature elimination with cross-validation with linear support vector machines estimator
RLN – Run length non-uniformity based on gray level run length matrix
ROC – Receiver operating characteristic
ROI – Regions of interest
RP – Run percentage based on gray level run length matrix
SC – Silhouette coefficient
SHG – Second harmonic generation
SNR – Signal-to-noise ratio
SOS – Second-order statistics
SRE – Short run emphasis based on gray level run length matrix
SVM – Support vector machines
TACS – Tumor-associated collagen signatures
THG – Third harmonic generation
TIMP – Tissue inhibitors of metalloproteinases
TME – Tumor microenvironment
TN – True negatives
TP – True positives
TPEF – Two-photon excitation microscopy
XGBoost – eXtreme gradient boosting
2D – Two-dimensional
μ_1 – Mean
σ – Standard deviation
g_1 – Skewness
g_2 – Kurtosis

CONTENTS

1 INTRODUCTION

Cancer is one of the leading causes of human mortality worldwide [1]. According to the International Agency for Research on Cancer (IARC) report published in 2024 [2], about 20 million new cancer cases and about 9.7 million cancer deaths were registered worldwide in 2022. In addition to the current statistics, the IARC has also made a demographic prediction that the number of new cancer cases will reach 35 million by 2050, which will lead to an increase in mortality rates. Early diagnosis and the discovery of new markers and characteristic features of cancer progression cannot prevent extreme growth in cancer cases, but they could significantly reduce mortality rates, prevent cancer recurrence and metastasis, and improve patients' recovery and lives after treatment. With the technological advancement in instrumentation and the incorporation of artificial intelligence (AI) in cancer data analysis, the advancement in computerized image analysis of cancer tissue is becoming a powerful tool that can efficiently complement the conventional analysis of cancer tissue samples, which is traditionally performed by direct visual inspection by experienced pathologists, and provide new insights for cancer diagnosis [3].

Thyroid cancer is one of the most common malignant diseases of the endocrine system and is characterized by the uncontrolled proliferation of cells within the thyroid gland. Its incidence has steadily increased in recent decades, with around 586,000 new cases reported worldwide in 2020 [4]. Papillary thyroid carcinoma (PTC) and follicular thyroid carcinoma (FTC) are the most common well-differentiated carcinomas and together account for approximately 88% of all thyroid tumors [5].

The encapsulation status of thyroid nodules, i.e. whether they are surrounded by a fibrous collagen capsule, is a key histopathologic feature. The presence and integrity of the capsule influences the assessment of tumor invasiveness, malignant potential and prognosis. In PTC, the capsule may be incomplete or infiltrative, whereas in FTC it is usually better developed. Capsular invasion, in which tumor cells penetrate and traverse the entire thickness of the capsule, is a key criterion for distinguishing between benign and malignant follicular tumors [6]. However, classification of tumors based on capsular invasion is challenging due to observer variability and complex histologic criteria. In addition, standard histopathologic methods, which often examine only limited areas of tissue sections, run the risk of overlooking microinvasions.

While the prognosis for FTC is often worse than for PTC [9,10] and complete thyroidectomy is required [11], low-risk PTC cases are often overdiagnosed and overtreated [12]. Therefore, accurate differentiation between FTC and PTC is essential to avoid unnecessary aggressive treatment and minimize postoperative complications [13].

Arterial hypertension is one of the most common concomitant diseases in cancer patients [14] and is a frequent adverse effect of cancer treatment. Cancer-related pulmonary arterial hypertension (PAH) has been observed in patients undergoing treatment with, e.g., dasatinib or other tyrosine kinase inhibitors [15]. PAH is a severe vascular disease characterized by increased pulmonary arterial pressure leading to vascular remodeling and excessive deposition of fibrillar collagen in the pulmonary arteries. These changes contribute to vascular stiffening and disease progression [16].

Structural changes in collagen content and its distribution are the main features of extracellular matrix (ECM) remodeling associated with pathological conditions such as PAH. In thyroid cancer, collagen remodeling also plays a critical role in tumor progression, with capsular changes correlating with malignancy and invasiveness. Understanding these common pathological mechanisms emphasizes the importance of collagen determination for the diagnosis and monitoring of disease progression. Due to the complexity and high overlap of the main pathological changes, PAH is considered a cancer-like disease [17]. Understanding the relationship between ECM remodeling in cancer and PAH with the progression, severity and degree of these diseases, and using this data for accurate diagnosis, could enable detailed pathological condition interpretation and more effective treatment of both cancer and PAH.

Second harmonic generation (SHG) microscopy provides a label-free method for visualizing fibrillar collagen, the primary structural component of thyroid nodule capsules. SHG bioimaging is particularly effective for assessing changes in collagen-rich tissue [18], as collagen produces a strong SHG signal due to its non-centrosymmetric structure [19]. Previous studies [20,21], have shown that SHG microscopy in conjunction with quantitative image analysis can differentiate between benign and malignant thyroid nodules. While conventional SHG imaging is based on scanning laser beams, SHG wide-field microscopy allows visualization of whole histological slides [22,23], enabling comprehensive analysis of collagen architecture.

SHG microscopy is particularly valuable when it comes to assessing the structural anisotropy of tissue by using the polarization of light. Polarization-resolved SHG microscopy (PSHG) has been used to quantitatively analyze collagen microstructure in tissues [24–26], including the thyroid gland [21,27,28]. While variants of SHG microscopy with laser beam scanning have proven successful in biomedical imaging, wide-field SHG microscopy is gaining recognition [23] for imaging whole histologic slides. Similar to the related imaging technique, two-photon excited fluorescence microscopy, for which wide-field variants also exist [31], wide-field SHG microscopy has evolved from intensity-based applications [32] to quantitative analysis [33] and even live imaging applications [34].

In addition to SHG imaging, texture analysis methods, including first-order statistics (FOS), second-order statistics (SOS), and higher-order statistics (HOS), provide quantitative descriptions of the properties of

collagen networks [35,36]. These methods are widely used in medical imaging such as computer tomography [37] and magnetic resonance imaging (MRI) [38].

The interpretation of image data is limited when analyzed manually with the traditional image processing pipeline. Taking advantage of large, scanned areas and high spatial resolution of wide-field SHG microscopy, the possibility of combining it with two-photon excited fluorescence microscopy and adding AI to the diverse features extracted from the imaging data paves the way for automated computerized image analysis of cancer tissue samples with high accuracy and high throughput.

Machine learning (ML) approaches are increasingly being used to automate and improve cancer tissue analysis and classification. Unsupervised ML techniques are mainly aimed at image segmentation and detection of specific patterns in image features based on the inherent relationships [39,40]. Supervised ML algorithms are mainly applied for solving classification problems and have recommended themselves in disease diagnosing. The ML classifiers, including deep learning (DL) models, have shown promising results in discriminating different cancer types based on MRI, computer tomography and SHG image analysis [41–44]. However, effective ML-based classification requires high-quality data, as label noise (mislabeling of samples) and feature noise (irrelevant or redundant parameters) can significantly affect model performance [45]. Real-world image data rarely fulfills this criterion, leaving room for further data-specific adaptations of ML algorithms, architectures and strategies.

Quantitative insights into the capsular structures of thyroid nodules and remodeling of the ECM are necessary to enable advances and accuracy in the diagnosis of thyroid cancer and to understand the evolution of pathological conditions. Large-scale wide field SHG microscopy as a label-free, collagen-specific imaging technique can simplify the preparation of tissue samples by eliminating the steps of tissue fixation and staining and allowing samples to be measured immediately after surgical removal. In turn, the development of ML-based methods for analyzing large-scale SHG images of cancer tissue suggests an automated diagnostic approach that can complement traditional visual inspection of cancer tissue samples, reducing the likelihood of misdiagnosis and supporting optimal clinical decision making. Therefore, detailed studies in this direction are crucial, as the simplicity, speed and specificity of SHG wide-field microscopy combined with effective ML methods of cancer analysis can enable accurate and timely diagnosis, curbing the rise in cancer deaths, even though the number of cases is expected to increase rapidly over the next decades.

1.1 Goal of the research work

The aim of the Thesis is to develop machine learning models that use SHG large-scale scans of the tissue sections for the interpretation of the pathological conditions and the diagnosis of diseases.

1.2 Tasks for the research work

To achieve the goal set above, the following tasks were formulated and solved within the framework of the Thesis:

1. To perform a statistical analysis of SHG intensity characteristics and texture features of wide-field SHG scans of lung tissue samples from rats with monocrotaline-induced PAH at different disease stages and samples from the control group of rats, to reveal patterns in the features of SHG images and to compare them with the results of immunohistochemical analysis.

2. To apply the unsupervised machine learning algorithm *k*-means clustering for the analysis of 2D maps of parameters extracted from wide-field PSHG images of whole thyroid nodule sections based on a cylindrical model of collagen fiber hyperpolarizability, and to reveal patterns in the ultrastructure of collagen fibers in the intact capsule and in regions of invasion.

3. To perform the unsupervised machine learning-based analysis (principal component analysis (PCA) and *k*-means clustering) of intensity and texture features of wide-field SHG scans of collagen distribution in papillary thyroid carcinoma sections to identify regions of capsular invasion and propose a quantitative description of the intact capsule and areas of invasion.

4. To develop supervised ML models for the automatic differential diagnosis of papillary and follicular thyroid carcinomas using wide-field SHG imaging considering the effects of label noise and feature noise on the predictive performance of these models.

1.3 Statements of the Thesis

1. Statistical analysis of wide-field SHG images of lung tissue sections reveals and qualitatively and quantitatively describes characteristic changes in collagen organization, morphology and collagen content associated with the different stages of pulmonary arterial hypertension.

2. *k*-Means clustering of cylindrical model parameters extracted from wide-field polarization-resolved SHG images of whole thyroid nodule sections allows differentiation between areas of capsular invasion and unaffected regions of the capsule surrounding cancer cells by revealing patterns in the ultrastructure of collagen.

3. Unsupervised machine learning improves SHG image analysis, reveals the textural heterogeneity of papillary thyroid carcinoma capsule, and enables identification of capsular invasion, poorly distinguishable microinvasions and regions requiring additional examination based on the specific sets of image parameters.

4. A supervised machine learning model C-SVC enables differential diagnosis of papillary and follicular thyroid carcinomas based on SHG imaging, with an accuracy of 84.73%.

1.4 Novelty and relevance

1. Based on the statistical parameters extracted from the SHG images of the collagen network organization, in combination with the results of immunohistochemical analysis, the specific phases of PAH progression were revealed, and their quantitative description was proposed. These phases could potentially be evaluated as checkpoints of PAH pathogenesis.

2. Unsupervised ML analysis of collagen ultrastructure and orientation-related parameters extracted from the PSHG image sets of an entire thyroid nodule section, enabled the creation of characteristic maps of the thyroid nodules and their quantitative description, facilitating comparison between different regions of the nodule capsule and highlighting areas of invasion within the capsule.

3. Unsupervised ML analysis of the sets of wide-field SHG images of entire thyroid nodules enabled to reveal the textural heterogeneity of the collagen capsule surrounding PTC. The quantitative description of this heterogeneity reflected in texture features and specific spatial distribution across the collagen capsule sheds light on the changes in the collagen network associated with cancer spread.

4. The developed supervised ML-based approach using large datasets of SHG images of thyroid slices enables efficient discrimination between FTC and PTC. The proposed data noise management strategy improves diagnostic accuracy and demonstrates the feasibility of automated diagnosis of PTC and FTC based on SHG microscopy.

The results of the Thesis pave the way for an automated and quantitative SHG imaging-based diagnosis of PAH and have the potential to become an additional objective technique of choice for the treatment of PAH-associated fibrosis that is free from error-prone human decision-making. Moreover, quantitative analysis enabled by both PSHG and wide-field SHG microscopy may prove beneficial for the automated assessment of capsular invasion sites in thyroid pathology, helping to reveal poorly distinguishable invasions and highlighting areas of the PTC capsule that require closer and more careful examination. Collagen ultrastructure data can provide insights into the molecular basis of cancer progression, spread and metastasis. All this provides a reliable basis for considering ML-assisted SHG microscopy as a new efficient method for the diagnosis of thyroid cancer.

1.5 List of publications

This thesis is based on the results that were published in the following peer reviewed scientific papers:

Paper A **Y. Padrez,** L. Golubewa, T. Kulahava, T. Vladimirskaja, G. Semenkova, I. Adzerikho, O. Yatsevich, N. Amaegberi, R. Karpicz, Y. Svirko, P Kuzhir, D. Rutkauskas; Quantitative and Qualitative Analysis of Pulmonary Arterial Hypertension Fibrosis Using Wide-Field Second Harmonic Generation Microscopy. *Sci. Rep.* 2022, *12*, 7330, doi: 10.1038/s41598-022-11473-5.

Paper B L.G. Eftimie, **Y. Padrez, L.** Golubewa, D. Rutkauskas, R. Hristu; Widefield Polarization-Resolved Second Harmonic Generation Imaging of Entire Thyroid Nodule Sections for the Detection of Capsular Invasion. *Biomed. Opt. Express* 2024, *15*, 4705, doi: 10.1364/BOE.523052.

Paper C **Y. Padrez,** L. Golubewa, I. Timoshchenko, A. Enache, L.G. Eftimie, R. Hristu, D. Rutkauskas; Machine Learning-Based Diagnostics of Capsular Invasion in Thyroid Nodules with Wide-Field Second Harmonic Generation Microscopy. *Comput. Med. Imaging Graph.* 2024, *117*, 102440, doi: 10.1016/j.compmedimag.2024.102440.

Paper D **Y. Padrez,** R. Hristu, I. Timoshchenko, L.G. Eftimie, D. Rutkauskas, L. Golubewa; Supervised Machine Learning Thyroid Carcinoma Diagnosis Using Wide-Field SHG Microscopy; *IEEE Access*, vol. 13, pp. 112021-112038, 2025, doi: 10.1109/ACCESS.2025.3583435

The results of the thesis were presented at the following conferences:

1. D. Rutkauskas, L. Golubewa, **Y. Padrez**, R. Karpicz, G. Semenkova, I. Adzerikho, O. Yatsevich, T. Vladimirskaja, T. Kulahava, P. Kuzhir. Wide-field second-harmonic generation microscopy for analysis of fibrosis progression during pulmonary arterial hypertension. 44th the Lithuanian National Physics Conference, Vilnius, Lithuania, 6-8 October 2021.
2. **Y. Padrez**, L. Golubewa, R. Karpicz, D. Rutkauskas. Collagen orientation index determination in wide-field SHG microscopic images of lung tissue, Advanced Properties and Processes in Optoelectronic Materials and Systems (APROPOS 18), Vilnius, Lithuania, 5-7 October, 2022.
3. **Y. Padrez**, L. Golubewa, I. Timoshchenko, D. Rutkauskas. Clustering of second harmonic generation microscopy images of collagen capsules of thyroid nodules, 13-oji FTMC doktorantų ir jaunųjų mokslininkų konferencija (FizTech2023), Vilnius, Lithuania, 18-19 October 2023.

4. **Y. Padrez**, L. Golubewa, I. Timoshchenko, D. Rutkauskas. Evaluation of key statistical parameters of second harmonic generation microscopy images of thyroid nodules. 45th the Lithuanian National Physics Conference, Vilnius, Lithuania, 25-27 October 2023.

5. **Y. Padrez**, L. Golubewa, D. Rutkauskas. Wide-field second harmonic generation microscopy with mashine learning for thyroid cancer detection. 45th the Lithuanian National Physics Conference, Vilnius, Lithuania, 25-27 October 2023.

6. **Y. Padrez**, L. Golubewa, L. G. Eftimie, R. Hristu, D. Rutkauskas. Determination of collagen ultrastructure in cancer tissue via unsupervised ML analysis of P-SHG image parameters, Advanced Properties and Processes in Optoelectronic Materials and Systems (APROPOS 19), Vilnius, Lithuania, 1-4 October 2024.

7. **Y. Padrez**, L. Golubewa, I. Timoshchenko, L. G. Eftimie, R. Hristu, D. Rutkauskas. Machine learning algorithms for thyroid cancer diagnosis based on second harmonic generation microscopy. 14-oji FTMC doktorantų ir jaunųjų mokslininkų konferencija (FizTech2024), Vilnius, Lithuania, 15-17 October 2024.

8. **Y. Padrez**, L. Golubewa, I. Timoshchenko, A. Enache, L. G. Eftimie, R. Hristu, D. Rutkauskas, Thyroid cancer assessment with machine learning-assisted wide-field second harmonic generation microscopy, Proc. SPIE 13324, Multiphoton Microscopy in the Biomedical Sciences XXV, 133240L (19 March 2025), doi: 10.1117/12.3051011.

9. **Y. Padrez**, L. Golubewa, A. Enache, L. G. Eftimie, R. Hristu, D. Rutkauskas. Towards an automated diagnosis of well-differentiated thyroid carcinomas based on wide field second harmonic generation microscopy imaging. 3RD EuroCC Vilnius Workshop on Using HPC, Vilnius, Lithuania, January 20–21, 2025.

10. **Y. Padrez**, L. Golubewa, I. Timoshchenko, A. Enache, L. G. Eftimie, R. Hristu, D. Rutkauskas, Thyroid cancer assessment with wide-field second-harmonic generation microscopy. Carpathian Biophotonics Meeting, Sinaia, Romania, 8-12 September 2025.

Scientific publications that were not included in the thesis:

1. D. Bychanok, **Y. Padrez**, N. Liubetski, A. Arlouski, U.Kushniarou, I. Korobov, I. Halimski, T. Kulahava, M. Demidenko, A. Urbanowicz, J. Macutkevic, P. Kuzhir. Window tinting films for microwave absorption and terahertz applications, Journal of Applied Physics. **2022**, 131, 025110, doi: 10.1063/5.0075497.

2. L. Golubewa, **Y. Padrez,** S Malykhin, T. Kulahava, E. Shamova, I. Timoshchenko, M. Franckevicius, A. Selskis, R. Karpicz, A. Obraztsov, Y. Svirko, P. Kuzhir. All-Optical Thermometry with NV and SiV Color

Centers in Biocompatible Diamond Microneedles. *Adv. Opt. Mater.* **2022**, *10*, 2200631, doi: 10.1002/adom.202200631.

3. **Y. Padrez**, L. Golubewa, A. Bahdanava, M. Jankunec, I. Matulaitiene, D. Semenov, R. Karpicz, T. Kulahava, Y. Svirko, P. Kuzhir. Nanodiamond Surface as a Photoluminescent PH Sensor. *Nanotechnology* **2023**, *34*, 195702, doi: 10.1088/1361-6528/acb94b.

4. L. Golubewa, A. Klimovich, I. Timoshchenko, **Y. Padrez**, M. Fetisova, H. Rehman, P. Karvinen, A. Selskis, S. Adomavičiūtė-Grabusovė, I. Matulaitienė, A. Ramanavicius, R. Karpicz, T. Kulahava, Y. Svirko, P. Kuzhir. Stable and Reusable Lace-like Black Silicon Nanostructures Coated with Nanometer-Thick Gold Films for SERS-Based Sensing. *ACS Appl. Nano Mater.* **2023**, *6*, 4770–4781, doi: 10.1021/acsanm.3c00281.

5. L. Golubewa, H. Rehman, **Y. Padrez,** A. Basharin, S. Sumit, I. Timoshchenko, R. Karpicz, Y. Svirko, P. Kuzhir. Black Silicon: Breaking through the Everlasting Cost vs. Effectivity Trade-Off for SERS Substrates. *Materials (Basel).* **2023**, *16*, 1948, doi: 10.3390/ma16051948.

6. **Y. Padrez**, L. Golubewa. Black Silicon Surface-Enhanced Raman Spectroscopy Biosensors: Current Advances and Prospects. *Biosensors* **2024**, *14*, 453, doi: 10.3390/bios14100453.

7. **Y. Padrez**, V. Boiko, J. Kajan, T. Gregor, V. Tinkovae, R. Karpicza, M. Chaika. Temperature dependence of Charge Transfer Luminescence in Yb3+:YAG single crystal, Optical Materials: X, 22, May 2024, 100315, doi: 10.1016/j.omx.2024.100315.

8. A. Klimovich, L. Golubewa**, Y. Padrez,** I. Matulaitiene. Characterization of Human Urotensin II Peptide Adsorbed on Silver Electrode by Surface-Enhanced Raman Scattering Spectroscopy, Spectrochim Acta A Mol Biomol Spectrosc. Vol 329, pp.125565, (2025), doi: 10.1016/j.saa.2024.125565.

The conferences that were not included in the thesis:

1. **Y. Padrez**, N. Liubetski, A. Arlouski, U. Kushniarou, D. Bychanok. Characterization of electromagnetic properties of thin resistive films in microwave and terahertz ranges. 44th the Lithuanian National Physics Conference, Vilnius, Lithuania, 6-8 October 2021

2. **Y. Padrez**, L. Golubewa, T. Kulahava, B. Zousman, O. Levinson, R. Karpicz, Y. Svirko, P. Kuzhir. Steady-state spectroscopic properties of magnetic nanodiamonds. Open Readings 2022: Proceedings of the 65th International Conference for students of physics and natural sciences, Vilnius, Lithuania, Mart 15 – Mart 18, 2022. – 49 p.

3. L. Golubewa, **Y. Padrez**, H. Rehman, R. Karpicz, T. Kulahava, O. Levinson, P. Kuzhir. Structure analysis of carbon nanomaterials using black silicon based SERS substrate. 8th Workshop on Nanocarbon Photonics and Optoelectronics (NPO2022). Polvijarvi, Finland, 31 July – 5 August 2022 – 9 p.
4. **Y. Padrez**, L. Golubewa, T. Kulahava, B. Zousman, O. Levinson, R. Karpicz, P. Kuzhir. Spectroscopic characterization of nanodiamonds with sp2-sp3 graphene-like shell. 8th Workshop on Nanocarbon Photonics and Optoelectronics (NPO2022). Polvijarvi, Finland, 31 July – 5 August 2022 – 31 p.
5. L. Golubewa, **Y. Padrez**, R. Karpicz, T. Kulahava, O. Levinson, P. Kuzhir. All-optical pH sensing with hybrid sp2-sp3 carbon nanostructures 8th Workshop on Nanocarbon Photonics and Optoelectronics (NPO2022). Polvijarvi, Finland, 31 July – 5 August 2022 – 53 p.
6. **Y. Padrez**, L. Golubewa, S. Malykhin, T. Kulahava, R. Karpicz, A. Obraztsov, Y. Svirko, P. Kuzhir. Temperature-dependent fluorescence of SiV and NV color centers in micron-sized single crystal diamond needles. 8th Workshop on Nanocarbon Photonics and Optoelectronics (NPO2022). Polvijarvi, Finland, 31 July – 5 August 2022 – 54 p.
7. **Y. Padrez**, S. Malykhin L. Golubewa, R. Karpicz, P. Kuzhir. Dynamic spectroscopic properties of single-crystal diamond needles synthesized by different methods. Hanseatic Workshop on Exciton Dynamics and Spectroscopy, Vilnius, Lithuania, 24-26 August 2022 – 47 p.
8. M. Quarshie, **Y. Padrez**, L. Golubewa, S. Malykhin, Synthesis and Characterization of Core-Shell Graphene-Diamond Structures for Dual Functioning in Photothermal Cancer Therapy, Graphene Week, Munich, Germany, 5-9 September 2022
9. **Y. Padrez**, L. Golubewa, R. Karpicz. Features of fluorescence properties of nanodiamonds, 66th International science conference 'Open Readings 2023', Vilnius, Lithuania,18-21 April 2023, p. 261.
10. Malykhin, M. Quarshie, R. Ismagilov, **Y. Padrez**, P. Balasubramanian, F. Jelezko, A. Obraztsov, P. Kuzhir. Diamond needles for quantum applications, PREIN Photonics Symposium, Helsinki, Finland, 22-24 March 2023
11. M. Quarshie, S. Malykhin, L. Golubewa, **Y. Padrez,** P. Kuzhir. Passive targeting theragnostics with diamond needles, PREIN Photonics Symposium, On ship from Helsinki to Stockholm, Sweden., 22-24 March 2023.
12. S. Malykhin, M. Quarshie, R. Ismagilov, **Y. Padrez**, P. Balasubramanian, F. Jelezko, A. Obraztsov, P. Kuzhir. Quantum

applications with diamond needles. Nordic Optics and Photonics Days (NOPD), March 19-21 2023, Riga, Latvia.

13. **Y. Padrez**, L. Golubewa, A. Bahdanava, R. Karpicz, T. Kulahava. Nanodiamonds for sensing applications, *2023 IEEE Nanotechnology Materials and Devices Conference (NMDC)*, Paestum, Italy, 2023, pp. 753-754, doi: 10.1109/NMDC57951.2023.10343682.
14. M. Zalieckas, V. Čirgelis, **Y. Padrez**, L. Golubewa, R. Karpicz. Stability of Graphene Quantum Dots and Doxorubicin aggregates evaluated by optical methods, *2023 IEEE Nanotechnology Materials and Devices Conference (NMDC)*, Paestum, Italy, 2023, pp. 278-278, doi: 10.1109/NMDC57951.2023.10343608.
15. M. Quarshie, S. Malykhin, P. Kuzhir, **Y. Padrez** Diamond needles as solid-state single photon emitters, Optics & Photonics Days 2023, Joensuu, Finland, 30 May – 1 June 2023.
16. **Y. Padrez**, R. Karpicz, M. Chaika, J. Kajan, T. Gregor, G. Gamazyan. F+→Yb3+ charge transfer process in Yb:YAG single crystals, Conference: International Conference on Excited States of Transition Elements (ESTE) Świeradów Zdrój, Poland, 3-8 September 2023.
17. I. Halimski, D. Pashnev, M. Talaikis, A. Dementjev, R. Karpicz, J. Chmeliov, I. Kašalynas, A. Urbanovič, **Y. Padrez**, M. Tutkus, P. Lamberti, M. Shundalau. Interaction of organic molecules and nanomaterials: Spectroscopic and theoretical insights into a weak-bound complex of hexagonal boron nitride and trans-stilbene. Advanced Properties and Processes in Optoelectronic Materials and Systems (APROPOS 19), Vilnius, Lithuania, 1-4 October, 2024.
18. K. Chernyakova, M. Zaleckas, **Y. Padrez**, L. Golubewa, G. Grincienė, R. Karpicz. Stability of Graphene/Carbon/Boron nitride Quantum Dots and Doxorubicin aggregates evaluated by optical methods, 9th Workshop on Nanocarbon Photonics and Optoelectronics (NPO2024). Polvijarvi, Finland, 4-9 August 2024 – 53 p.
19. **Y. Padrez**, L. Golubewa, A. Zelioli, A. Špokas, B. Čechavičius, A. Vaitkevičius, E. Dudutiene, R. Butkute. Machine Learning-Based Hyperspectral Image Analysis of Emission Homogeneity of InGaAs Multiple Quantum Wells. 30th OptoElectronics and Communications Conference/International Conference on Photonics in Switching and Computing 2025 (OECC/PSC 2025), June 29-July 3, 2025, Sapporo, Japan.
20. L. Golubewa, **Y. Padrez**, A. Špokas, A.Zelioli, A. Štaupienė, B. Čechavičius, E. Dudutienė, A. Vaitkevičius, R. Butkutė. Unsupervised Machine Learning Study of GaAsBi Quantum Well Evolution After Annealing Based on Spatially Resolved micro-Photoluminescence Imaging. 30th OptoElectronics and Communications

Conference/International Conference on Photonics in Switching and Computing 2025 (OECC/PSC 2025), June 29-July 3, 2025, Sapporo, Japan.

1.6 Author's contribution

The doctoral student was responsible for the majority of the experimental work, including wide-field SHG and TPEF image data acquisition, formal analysis and writing of scientific publications. A significant portion of this work involved the development of custom code for data processing, computational modeling, and analysis, which was fully implemented by the author. All calculations, computations, visualizations and subsequent analysis and interpretation were performed by the doctoral student.

Wide-field SHG setup modification for PSHG imaging of PTC samples and PSHG image dataset acquisition was carried out by Dr. Danielis Rutkauskas (Center for Physical Sciences and Technology, Vilnius, Lithuania) and Dr. Radu Hristu (Center for Microscopy-Microanalysis and Information Processing, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania).

The preparation of the lung tissue sections and immunohistochemical staining were performed by Dr. Nadezda Amaegberi (Belarusian State University, Minsk, Belarus), Dr. Tatyana Vladimirskaja and Olga Yatsevich (Belarusian Medical Academy of Postgraduate Education, Minsk, Belarus). Tissue sections of PTC and FTC nodules were prepared by Dr. Lucian George Eftimie (Central University Emergency Military Hospital, Bucharest, Romania; The National University of Physical Education and Sports, Bucharest, Romania).

1.7 Author's biography

Yaraslau Padrez was born in Borisov, Belarus. In 2014, he graduated from the state educational institution "Lyceum of Borisov" and entered the Faculty of Physics of Belarusian State University (Minsk, Belarus). In 2020, he obtained a specialist degree (5.5 years of study) in nuclear physics and technology. In 2019-2020 he received the scholarship of the World Federation of Scientists. In 2021, he began his PhD studies at the State Research Institute Center for Physical Sciences and Technology in the Department of Molecular Compounds Physics.

2 LITERATURE OVERVIEW

2.1 SHG Microscopy

In recent years, the SHG has become an increasingly important nonlinear optical contrast mechanism in biological and biophysical imaging applications. SHG is a coherent process in which two photons of lower energy are combined to produce a photon with exactly twice the incident frequency (or half the wavelength) *Figure 2.1*. The conversion of radiation frequency occurs through virtual states of the system without complete transfer of energy into the system. This process was first demonstrated in biological systems by Freund and colleagues in 1986 [46]. They used SHG to study the polarity of collagen fibers in rat tail tendons. Although their work was performed at a relatively low resolution (~50 micrometers), it served as an important proof of concept and laid the foundation for modern SHG imaging techniques [47].

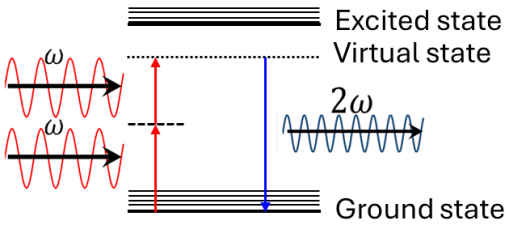


Figure 2.1 SHG energy diagram.

Since then, SHG imaging has evolved considerably, mainly due to advances in technology. The development of commercially available ultrafast lasers, combined with improvements in laser scanning and data acquisition systems, has made SHG a widely used tool in both material science and biomedical imaging. These innovations also reflect advances in other nonlinear microscopy techniques, such as multiphoton fluorescence microscopy, which have become routine in research laboratories.

Despite the advantages of classical optical microscopes, they have difficulties in imaging large (opaque) or complex specimens due to the limited penetration depth and scattering of the radiation [48]. SHG microscopy overcomes these limitations and offers high-resolution imaging with deeper tissue penetration. This renders SHG microscopy an exceptional imaging instrument for biomedical applications that necessitate high-resolution, non-destructive imaging.

The combination of reliable, robust ultrafast laser technologies and turnkey microscope systems has made SHG microscopy more accessible and further solidified its role in improving our understanding of biological and material systems [49,50]. Continued improvements in hardware and software

are expected to further enhance the utility of SHG microscopy for a wide range of scientific and clinical applications.

SHG is an instantaneous, scattering-based process that preserves photon energy and phase coherence. SHG microscopy typically uses a pulsed femtosecond laser source or a picosecond laser source focused through a high numerical aperture objective that generates a strong electric field in the focal volume. The standard method for SHG imaging is to tightly focus the laser light onto the sample with an objective lens, raster the focal volume over an area in the sample, and reconstruct an image by linking the intensity of the generated SHG signal detected by a detector to the position of the illumination volume in the sample [51].

The signal intensity scales quadratically with the intensity of the incident light and is very sensitive to molecular organization, making it a powerful tool for quantifying structural anisotropy and spatial arrangement in biological tissue [52].

2.1.1 Advantages/disadvantages of SHG microscopy

The following principal advantages of SHG microscopy can be identified:

- *High resolution and specificity*: SHG microscopy offers high spatial resolution (~300 nm – 500 nm lateral resolution and ~800 nm – 1200 nm axial resolution [19,53,54]) and is therefore ideal for imaging structures in the micrometer range. It can capture fine details of biological tissue, such as collagen fibers, with excellent specificity due to its unique contrast mechanism [53].
- *Based on endogenous contrast:* Unlike many fluorescence microscopy techniques, SHG does not require exogenous fluorescent dyes or contrast agents. SHG arises from the sample itself, but no additional dyes [55]. This is particularly advantageous in biological studies where the introduction of additional substances (reagents) can alter cellular processes and lead to erroneous conclusions [46].
- *Label-free imaging*: SHG microscopy is a label-free imaging technique, meaning it can image naturally occurring structures such as collagen, myosin and other biomolecules without the need for chemical markers or labels. This reduces the risk of interference with the natural behavior of the sample [53].
- *Free from photobleaching and instantaneous*: in contrast to fluorescence microscopy, SHG is not limited in time and is not limited by repetition rate [56].
- *Improved imaging depth*: SHG microscopy can penetrate deeper into tissue compared to conventional fluorescence microscopy, since NIR-I (700-1000 nm) and NIR-II (1000-1300 nm) spectral ranges are traditionally employed for excitation [48,52,57]. Such excitation minimizes absorption by water and biomacromolecules, limits scattering, and makes SHG

microscopy particularly valuable for imaging thick samples or tissue *in vivo* [48].

- *Optical slicing*: SHG provides intrinsic optical slicing capabilities that enable visualization of thick tissue sections without the need for physical slicing. This is particularly useful for 3D imaging and reduces the complexity of sample preparation [58].
- *Non-destructive:* SHG microscopy uses near-infrared excitation light, which causes significantly less photodamage and phototoxicity compared to visible or ultraviolet light used in other imaging techniques. Therefore, the SHG method is ideal for imaging living tissue over an extended period of time or performing repeated measurements without compromising the integrity of the samples [47,59].

On the other hand, the following disadvantages of SHG can be highlighted:

- *Limited to non-centrosymmetric structures*: SHG can only be generated in non-centrosymmetric structures, i.e. it is not suitable for imaging all types of biological macromolecules or tissue structures. This limits its applicability to certain biomolecules such as collagen, myosin and several other asymmetric structures [52].
- *Low-intensity signal*: SHG microscopy generates signals with relatively low intensity compared to fluorescence techniques, which require the use of powerful pulsed laser sources and highly sensitive detection systems. These hardware requirements can increase the overall cost and complexity of SHG setups, potentially limiting wider application in standard laboratory environments [60].
- *Technically demanding and resource-intensive:* SHG microscopy requires sophisticated instrumentation, including ultrafast mode-locked lasers, advanced optical components and sensitive detection systems. It also requires precise experimental conditions — such as the operation of a high-power laser and careful optimization of the wavelength — which can be difficult to maintain. These factors make SHG microscopy more complex and costly compared to conventional imaging techniques, potentially limiting its accessibility in resource-constrained laboratories [51,61].
- *Limited chemical information*: While SHG provides excellent structural information, it does not provide detailed chemical information about the sample, unlike techniques such as Raman spectroscopy/microscopy, FTIR, etc. This means that SHG alone cannot distinguish between different types of biomolecules or provide insights into molecular composition [62].
- *Limited by scattering in dense tissues:* Although SHG microscopy benefits from lower scattering compared to linear fluorescence microscopy - due to the use of near-infrared excitation and coherent signal generation - it is still susceptible to scattering effects in very heterogeneous or optically dense tissues. At greater imaging depths, scattering can affect signal strength and

resolution, requiring additional optical corrections or signal processing strategies [63].

- *Low signal-to-noise ratio*: The inherently weak SHG signal can lead to a low signal-to-noise ratio, especially when imaging thick biological tissue or structures with low SHG efficiency. In such cases, careful optimization of sample preparation, imaging depth and optical alignment is crucial to obtain high quality images [18].

It is also worth noting that the most important and unique feature of SHG microscopy is the sensitivity of SHG signatures to physical structure [18] (described in detail below). This can be considered both a disadvantage and an advantage.

2.1.2 Physical principles of SHG microscopy

The fundamental physical principles governing SHG microscopy stem from nonlinear optics, particularly second-order nonlinear light-matter interactions.

The response of the material to the applied electric field $\vec{E}$ is described by its total non-linear polarization $\vec{P}$ (neglecting permanent polarization) according to the following equation [18]:

Equation 1: $$\vec{P} = \varepsilon_0 \hat{\chi}^{(1)} \cdot \vec{E} + \varepsilon_0 \hat{\chi}^{(2)} : \vec{E}\vec{E} + \varepsilon_0 \hat{\chi}^{(3)} \vdots \vec{E}\vec{E}\vec{E} + \cdots,$$

or alternatively:

Equation 2: $$P_i = \varepsilon_0 (\chi_{ij}^{(1)} E_j + \chi_{ijk}^{(2)} E_j E_k + \chi_{ijkl}^{(3)} E_j E_k E_l + \cdots),$$

where $\vec{P}$ is the induced polarization, P_i (E_i) is the *i*-th Cartesian coordinate of the polarization (electric field), $\hat{\chi}^{(n)}$ is the *n*-th order non-linear susceptibility tensor of rank (n+1), and $\vec{E}$ is the applied electric field, repeating indices imply summation. For convenience, it will be assumed that $1/n!$ term from Taylor series expansion is included within the susceptibility tensor [64]. The $\hat{\chi}^{(n)}$ corresponds to the next optical effects:

- 1st-order processes, $\hat{\chi}^{(1)}$: absorption and reflection;
- 2nd-order processes, $\hat{\chi}^{(2)}$: SHG, sum and difference frequency generation, hyper-Rayleigh scattering;
- 3rd-order processes, $\hat{\chi}^{(3)}$: multiphoton absorption, third harmonic generation, coherent anti-Stokes Raman scattering, Kerr effect, self-phase modulation, cross-phase modulation [55].

The nature of the second-order SHG imposes strict symmetry constraints on the mappable harmonophores and their structure. The nonlinear susceptibility tensor, $\hat{\chi}^{(2)}$, is a bulk property and the quantity that can be measured in an experiment. It represents the macroscopic nonlinear response of the medium, which is composed of elementary harmonophores (scatterers)

on a smaller scale. In collagen fibers, the peptide bonds along the collagen fibers serve as elementary nonlinear scatterers and form the basis of the contrast mechanism. These elementary harmonophores at the molecular level are described by the first hyperpolarizability tensor, $\hat{\beta}$ (β_{ijk}). This parameter is defined in terms of the second-order non-linear dipole moment [53]:

Equation 3: $\vec{d}^{(2)} = \hat{\beta}\vec{E}\vec{E}$

or alternatively:

Equation 4: $d_i^{(2)} = \beta_{ijk}E_jE_k$,

where d_i (E_i) is the *i*-th Cartesian coordinate of the induced second-order non-linear dipole moment (electric field strength).

The molecular and the macroscopic properties of such molecules as collagens are linked by the following formula:

Equation 5: $\hat{\chi}^{(2)} = \sum_s N_s < \hat{\beta}^{(2)}{}_s >$,

where N_s is the density of the groups of molecules denoted as *s* and the brackets denote an orientational average. The hyperpolarizability tensor $\hat{\beta}$ is related to the dipole moment. If these dipole moments are non-randomly aligned within the focal volume of the microscope stage, the bulk property constraints are satisfied and the macroscopic characteristic $\hat{\chi}^{(2)}$ is non-zero and can be measured (*Equation 5*). Therefore, the maximum SHG contrast is observed for well-aligned molecules that assemble into fibrils.

The intensities of the second harmonic in such media scale as follows [65]:

Equation 6: $SHG_{sig} \propto p^2\tau\left(\chi_{eff}^{(2)}\right)^2$,

where $\chi_{eff}^{(2)} = \sum_{ijk}\chi_{ijk}^{(2)} \cdot E_i(\omega)E_j(\omega)E_k(2\omega)$, p and τ are the laser pulse energy and pulse width, respectively.

The magnitude of the SHG intensity can be greatly enhanced when the energy of the SHG signal ($2\hbar\omega$) is in resonance with an electronic absorption band ($\hbar\omega_{ge}$). Within the two-level system model [66], the first hyperpolarizability $\beta_{two-level}$ (dominant tensor component aligned with the transition dipole moment direction, which enables simplified scalar representation) and thus the SHG efficiency is defined as [67]:

Equation 7: $\beta_{two-level} \approx \frac{3e^2}{2\hbar^3}\frac{\omega_{ge}f_{ge}\Delta\mu_{ge}}{[\omega_{ge}^2-\omega^2][\omega_{ge}^2-4\omega^2]}$,

where e is the electron charge and ω_{ge}, f_{ge} and $\Delta\mu_{ge}$ are the energy difference, the oscillator strength or the integral of the absorption spectrum and the change in dipole moment between the ground state and the excited state, respectively (for a two-level system with μ_{ge} and $\Delta\mu_{ge}$ along the z-axis). Other components (e.g., β_{xxy}) could be zero or smaller depending on symmetry. Due to the denominator in this equation, the resonance-enhanced SHG has a dependence on the wavelength of the incident light, which is similar to the two-photon excitation spectrum.

2.1.3 Biomacromolecules suited for SHG microscopy

A crucial prerequisite for SHG is the absence of centrosymmetry in the medium (*Equation* 6). In contrast to fluorescence-based imaging, SHG does not occur in isotropic or centrosymmetric materials due to symmetry constraints in the nonlinear susceptibility tensor. Common biological structures that support SHG include collagen, myosin, and microtubules, as these biological assemblies exhibit non-centrosymmetric organization.

Important biomacromolecules that exhibit SHG:

- *Collagen*: collagen fibers are the most commonly studied biomolecule in SHG microscopy. They have a highly ordered, non-centrosymmetric triple-helical structure, which makes them ideal sources of SHG in connective tissues.
- *Myosin*: Found in muscle tissue, myosin filaments exhibit a non-centrosymmetric organization that enables SHG imaging of sarcomere structures in skeletal and cardiac muscles.
- *Microtubules*: As key components of the cytoskeleton, microtubules consist of aligned tubulin dimers arranged in a non-centrosymmetric lattice and contribute to SHG contrast in cellular imaging.
- *Fibrillar proteins*: In some cases, when highly ordered, actin filaments can exhibit SHG properties, especially in structured tissues such as neurons or contractile cells.
- *Crystalline lipid structures*: Certain lipid structures, such as myelin sheaths surrounding nerve fibers, can generate SHG signals due to their anisotropic molecular arrangement.

These biomacromolecules often form highly ordered structures that enhance nonlinear susceptibility, and their molecular symmetry breaks at the supramolecular level, enabling the generation of coherent SHG signals. Furthermore, their biological significance makes SHG a valuable tool for the study of tissue morphology, disease progression and dynamic cellular processes.

It is also worth noting that the use of exogenous dyes (with high quadratic hyperpolarizability) can produce SHG contrast in tissues that do not have sufficiently high intrinsic hyperpolarizabilities for SHG. As shown in *Figure 2.2*, a portion of the incident light at frequency (ω) is converted to a SHG signal at frequency (2ω) when it interacts with molecules that do not

exhibit inversion symmetry. The SHG reaction is further enhanced when the virtual state matches the electronic energy levels of the molecules. In solutions where the molecules are randomly aligned, there is usually no coherent SHG signal because the electric fields of the emitted second-order light waves cancel each other out by destructive interference. In contrast, when these molecules are adsorbed on a surface, they tend to be more ordered, allowing constructive interference and the detection of a measurable SHG signal [68].

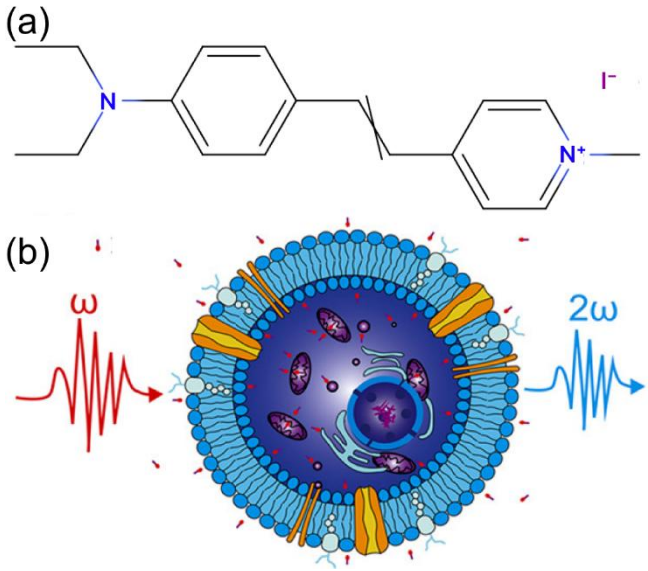


Figure 2.2 (a) Molecular structure of D289. (b) Schematic representation of the interaction between D289 molecules and a living tumor cell. Adapted with permission from [68] *Copyright © 2021, American Chemical Society.*

2.2 SHG Microscopy of collagen fibers

SHG microscopy enables detailed visualization of the architecture of collagen fibers in various tissues such as skin, cornea, tendons and bone. Its optical sectioning capability reveals the orientation, density and bundling of collagen fibers in three dimensions, which is crucial for understanding tissue biomechanics and pathology.

Collagen is the most prominent endogenous SHG emitter in biological tissues due to its abundant occurrence and unique molecular organization. Its triple-helical structure forms fibrils that are highly ordered and non-centrosymmetric, thus fulfilling the requirements for the generation of strong SHG signals. SHG microscopy of collagen is often used to assess tissue architecture and remodeling as well as pathological changes.

2.2.1 Theoretical description of SHG microscopy of collagen fibers

The general case of three-wave mixing, from two fields with frequencies ω_1 and ω_2 to a third field with $\omega_1 + \omega_2$ can be defined at the bulk level as follows [54]:

Equation 8:
$$P_i^{(2)}(\omega_1 + \omega_2) = \\ = \varepsilon_0 \sum_{jk} \chi_{ijk}^{(2)}(-(\omega_1 + \ +\omega_2); \omega_1, \omega_2) E_j(\omega_1) E_k(\omega_2),$$

or at the molecular level as:

Equation 9:
$$d_i^{(2)}(\omega_1 + \omega_2) = \\ = \sum_{jk} \beta_{ijk}(-(\omega_1 + \omega_2); \omega_1, \omega_2) E_j(\omega_1) E_k(\omega_2).$$

Here and further in the text, all the frequency arguments of susceptibility and hyperpolarizability tensors will be represented according to [69].

The dipole moment $d_i^{(2)}$ and the hyperpolarizability β_{ijk} are the molecular counterparts of the polarization $P_i^{(2)}$ and the susceptibility $\chi_{ijk}^{(2)}$, respectively. The generation of the second harmonic is a degenerate case of three-wave mixing such that $\omega_1 = \omega_2 = \omega$. Therefore, from *Equation 1*,

Equation 10:
$$P_i^{(2)}(2\omega) = \varepsilon_0 \sum_{jk} \chi_{ijk}^{(2)}(-2\omega; \omega, \omega) E_j(\omega) E_k(\omega) = \\ = \varepsilon_0 \sum_{kj} \chi_{ikj}^{(2)}(-2\omega; \omega, \omega) E_k(\omega) E_j(\omega) = \\ = \varepsilon_0 \sum_{jk} \chi_{ikj}^{(2)}(-2\omega; \omega, \omega) E_j(\omega) E_k(\omega).$$

Hence,

Equation 11:
$$\chi_{ikj}^{(2)}(-2\omega; \omega, \omega) = \chi_{ijk}^{(2)}(-2\omega; \omega, \omega).$$

Given this symmetry, a contracted notation can be used [54],

Equation 12:
$$\chi_{is}^{(2)}(\omega) = \chi_{ijk}^{(2)}(-2\omega; \omega, \omega),$$

where *i* remains a spatial index with values from 1 to 3 and *s* from 1 to 6, with the following relationships between *s*, *j* and *k:*

Equation 13:

s	1	2	3	4	5	6
j	1	2	3	2	3	1.
k	1	2	3	3	1	2

A further simplification can be introduced if ω_1 and ω_2 are far away from any natural frequencies (outside resonance). In this case, susceptibility can be assumed to be independent of frequency:

Equation 14: $\chi_{ijk}^{(2)}(-(\omega_1+\omega_2);\omega_1,\omega_2) \cong \chi_{ijk}^{(2)}$.

As a result, it could also freely interchange the indices, a condition known as Kleinman symmetry. This symmetry follows from an assumption that there is no dispersion, which is applicable to the middle of the visible – NIR range:

Equation 15: $\chi_{ijk}^{(2)} = \chi_{ikj}^{(2)} = \chi_{jki}^{(2)} = \chi_{jik}^{(2)} = \chi_{kij}^{(2)} = \chi_{kji}^{(2)}$.

The general case of three-wave mixing has 27 independent elements in the susceptibility tensor *Equation 8*. Similarly, the SHG has 18 independent elements in the contracted matrix, *Equation 14*. Finally, if Kleinman symmetry holds, there are only 10 independent elements *Equation 15*.

Given the cylindrical symmetry of collagen fibers (see *Figure 2.3a*), consider the properties of the susceptibility tensor under this simplification. For a general rotation from the axes (x', y', z') to (x, y, z), the components of the susceptibility tensor *Equation 14* are modified to:

Equation 16: $\chi_{ijk}^{(2)} = \sum_{i'j'k'} cos(\theta_{ii'})\, cos(\theta_{jj'})\, cos(\theta_{kk'})\ \chi_{i'j'k'}^{(2)}$,

where $\theta_{ii'}$ is the angle between the axes i' and i. For the invariance under xy-rotations around the longitudinal axis of the cylinder, it follows that:

Equation 17:
$$\begin{cases} \chi_{zzz}^{(2)} = \chi_{33} \\ \chi_{zxx}^{(2)} = \chi_{zyy}^{(2)} = \chi_{13} \\ \chi_{xxz}^{(2)} = \chi_{xzx}^{(2)} = \chi_{yyz}^{(2)} = \chi_{yzy}^{(2)} = \chi_{15} \\ \chi_{xyz}^{(2)} = \chi_{xzy}^{(2)} = -\chi_{yxz}^{(2)} = -\chi_{yzx}^{(2)} = \chi_{14} \end{cases}.$$

Angle θ represent mean harmonophore orientation (*Figure 2.3a*). According to the collagen model, θ ranges from 49° to 57° with a maximum disorder width $\delta = 67°$ when $\theta = 57°$. $P = 9.5$ Å and $R = 1.5$ Å are helix pitch and radius respectively [70].

At the same time, all the other components vanish. In the one-letter notation of Mazely and Hetherington [71], there are therefore only four independent elements for the cylindrical symmetry: χ_{33}, χ_{13}, χ_{15} and χ_{14}.

For molecules with a certain distribution of molecular orientation in the axes (x', y', z') relative to the bulk axes (x, y, z), the bulk susceptibility is derived from the molecular hyperpolarizability as follows:

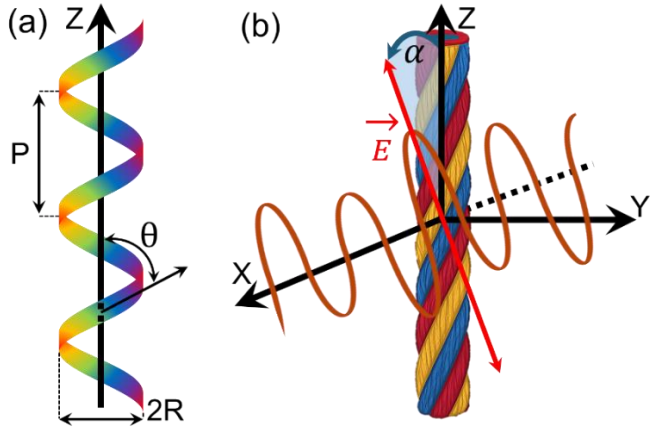


Figure 2.3 (a) Model of single helix collagen. (b) A beam of linear polarized light propagating along the X axis of the laboratory coordinates (X, Y, Z), focusing on a collagen fiber aligned in the XY plane. The red arrow corresponds to the direction of light polarization.

Equation 18: $\chi^{(2)}_{ijk} = \sum_{i'j'k'} < cos(\theta_{ii'})\, cos(\theta_{jj'})\, cos(\theta_{kk'}) > \beta_{i'j'k'}$,

where the angle brackets correspond to the averaging over the molecules. Let us consider a molecule with a single preferred axis of hyperpolarizability, then:

Equation 19: $\beta_{z'z'z'} = \beta$,

for which all other components vanish. Let us assume that the molecules are distributed with a constant polar angle, $\theta_{zz'} = \theta$, and a random azimuthal angle, φ. From *Equation 17*, it follows that

Equation 20:

$$\begin{cases} \chi^{(2)}_{zzz} = \chi_{33} = N\, cos^3(\theta)\, \beta \\ \chi^{(2)}_{zxx} = \chi_{13} = \frac{1}{2} N\, cos(\theta)\, sin^2(\theta)\, \beta \\ \chi^{(2)}_{xxz} = \chi_{15} = \frac{1}{2} N\, cos(\theta)\, sin^2(\theta)\, \beta = \chi_{13} \\ \chi^{(2)}_{xyz} = \chi_{14} = N\, cos(\theta)\, sin^2(\theta) < cos(\varphi)\, sin(\varphi) > \beta = 0 \end{cases}.$$

There are only two independent elements for the distribution of uniaxial molecules in a cylindrically symmetric approximation (χ_{33} and χ_{13}). The characteristic polar angle θ can be calculated from their ratio. Such an arrangement can be achieved by a random distribution of molecules in a monolayer [71] or by the formation of ordered structures such as the triple-helical structure of collagen, which form cylindrical arrangements of polypeptide spirals (*Figure 2.3a*).

Assuming that the light propagates along the direction X and collagen is aligned along the axis Z (*Figure 2.3b*), *Equation 10* can be represented in the transverse wave approximation ($E_X(\omega) = 0$):

Equation 21:
$$\begin{cases} P_X(2\omega) = 0 \\ P_Y(2\omega) = 2\chi_{15}E_Y(\omega)E_Z(\omega) \\ P_Z(2\omega) = \chi_{31}\left(E_Y(\omega)\right)^2 + \chi_{33}\left(E_Z(\omega)\right)^2 \end{cases}.$$

Using $E_Y(\omega) = E(w)\sin\alpha$ and $E_Z(\omega) = E(w)\cos\alpha$, where $E(w)$ is the electric field strength of the applied field (*Figure 2.3b*) [46], the SHG intensity from *Equation 6* can be changed according to *Equation 21* in the following way:

Equation 22: $$I(2\omega) \sim \left[sin^2\,2\alpha + \left(\frac{\chi_{31}}{\chi_{15}}sin^2\,\alpha + \frac{\chi_{33}}{\chi_{15}}cos^2\,\alpha\right)^2\right].$$

Typical dependence of the intensity (*Equation 22*) of the SHG signal of collagen in rat tail tendon and their SHG images are presented in *Figure 2.4*.

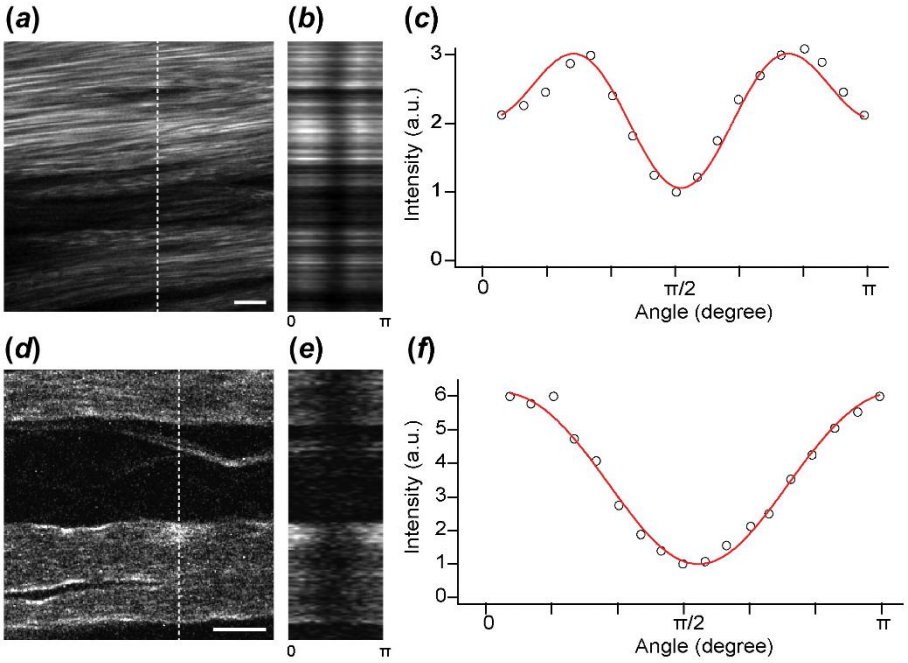


Figure 2.4 Polarization-resolved SHG of collagen in rat tail tendon and MT in the retinal nerve fibers. Scale bars, 20 μm. Reprinted from [72] *© copyright 2015, by permission of Informa UK Limited, trading as Taylor & Francis Group*

There are at least 28 types of collagen [73]. However, not all of them can be visualized with SHG. Basically, collagen I and collagen type III give a

signal, while collagen type IV, which is the major component of basement membranes, cannot be visualized with SHG because its reticular, non-fibrillar structure lacks the highly ordered, non-centrosymmetric arrangement required for efficient SHG [19,74]. SHG microscopy is particularly sensitive to fibrillar collagens due to their triple-helical, highly aligned structure, which facilitates constructive interference of the emitted SHG signal [75]. In contrast, the more amorphous and reticular architecture of collagen IV results in weak or undetectable SHG signals. Therefore, SHG imaging is best suited for studying the organization and remodeling of interstitial collagens such as types I and III, which predominate in the extracellular matrix of connective tissue and play a critical role in processes such as fibrosis, tumor progression, and wound healing [76,77].

2.2.2 Collagen modification in fibrosis and cancer

Changes in collagen organization are a hallmark of various diseases, including fibrosis and cancer. Fibrotic tissue typically exhibits increased collagen deposition and altered fiber alignment, both of which are detectable by SHG. Similarly, in the tumor microenvironment, collagen fibers can become straighter and more directional, promoting tumor progression and metastasis [78,79].

During carcinogenesis and cancer development, tumor cells overcome the physical barrier formed by the basement membrane and the interstitial matrix. This crucial step of invasion is accompanied by a profound remodeling of the ECM, especially collagen fibrils [80]. Cancer-associated fibroblasts (CAFs) are activated in response to tumor-derived signals and play a central role in this process. These fibroblasts increase collagen deposition and actively reorganize the fibrillar network, resulting in a denser, better aligned collagen structure and contributing to increased matrix stiffness [81]. This stiffening of the tumor microenvironment (TME) promotes mechanotransduction signals in the tumor cells, which increases their invasive and proliferative capabilities [82].

In parallel, tumor-associated macrophages (TAMs) and CAFs secrete matrix metalloproteinases (MMPs) and other proteolytic enzymes that degrade native collagen structures [83]. This controlled degradation, in conjunction with the deposition of new collagen, leads to a dynamic remodeling of the collagen matrix characterized by the alignment of collagen fibers perpendicular to the tumor border. Such changes are observed in several types of tumors and nowadays are classified as characteristic tumor-associated collagen signatures (TACS) [79].

The TACS concept, originally introduced in 2006, describes a spatially resolved classification of the collagen fiber organization around invasive tumors, particularly in breast cancer, into three structurally distinct zones: TACS-1, TACS-2 and TACS-3 [79]. TACS-1 is characterized by a dense, tangled collagen matrix directly adjacent to the tumor mass, which represents the earliest structural response of ECM to tumor development. TACS-2

exhibits collagen fibers arranged in concentric, spherical shells around the tumor, reflecting a transitional organization that may facilitate cell migration. However, the most clinically significant form is TACS-3 — defined by straight, linear collagen fibers-oriented perpendicular to the tumor border. These fibers extend radially into the surrounding tissue, forming "highways" that facilitate the directional migration of cancer cells into the stroma and ultimately promote local invasion and metastasis [79,84]. The localization of the TACS layers by linear protrusions of collagen emanating from the tumor can be seen in *Figure 2.5*. In detail, TACS1 describes an increased collagen deposition around the tumor, TACS2 – spherical alignment of layers around the tumor, and TACS3 – the development of vertical collagen tracts away from the tumor.

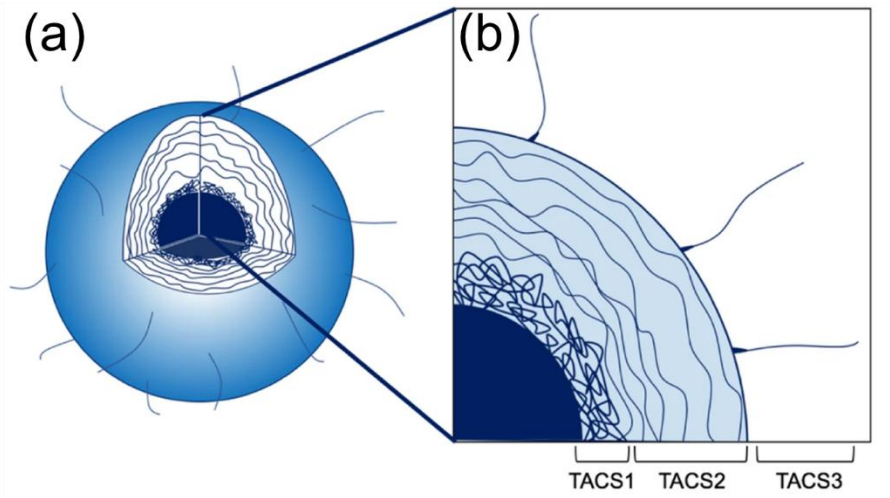


Figure 2.5 Illustration of TACS1-3: (a) schematic representation of a 3-dimensional tumor, (b) magnified breakdown of the TACS layers from the inside to the outside. Adapted from [85].

In a recent study based on multiphoton microscopic imaging in a broad cohort of breast cancer tissue samples, a refined classification was proposed that captures the dynamic evolution of collagen structures during tumor progression [86]. The broader spectrum of 8 different collagen morphologies does not replace the original TACS model but rather extends it by adding complexity and clinical value by recognizing the gradual and multifactorial nature of ECM remodeling during cancer progression [86]. This suggests that a more detailed separation of collagen changes in different diseases is needed.

The altered organization of collagen fibers reduces the mechanical constraints on migrating tumor cells, creating "highways" that facilitate directional invasion into the surrounding tissue [87]. In addition, the reorganized collagen network enhances integrin-mediated adhesion and the tensile forces required for cell motility [88]. Cross-linking enzymes such as

lysyl oxidase, which are upregulated in the TME, further stabilize and stiffen the collagen matrix and reinforce these invasive pathways [89].

The evolution of collagen fibril organization during tumor progression is a highly coordinated process that involves increased collagen deposition, proteolytic remodeling, enzymatic cross-linking, and fiber realignment *Figure 2.6*. These changes not only support tumor cell invasion, but also create a tumor-friendly microenvironment that promotes immune evasion, angiogenesis, and metastatic spread [82].

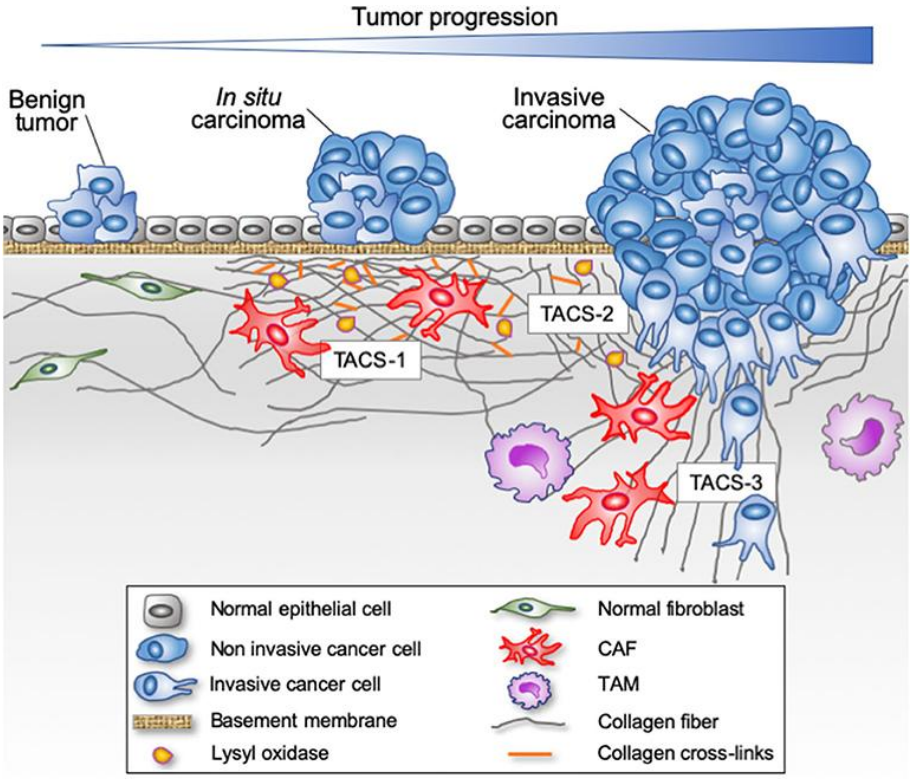


Figure 2.6 Evolution of fibrillar collagen organization during tumor progression. Reprinted from [90].

2.2.3 Collagen changes in PAH

Similar patterns of collagen remodeling and ECM dysregulation are observed in non-malignant diseases characterized by abnormal tissue stiffness and fibrosis, such as PAH. PAH is a rare but often fatal disease that contributes significantly to high morbidity and mortality in both adult and pediatric patients with lung disease [91]. A key pathological feature of all forms of pulmonary hypertension is pulmonary arterial remodeling. This process involves the deposition of ECM components such as fibronectin and collagen as well as the proliferation, migration and hypertrophy of vascular smooth

muscle cells. These changes lead to thickening and muscularization of the pulmonary arteries and ultimately increase pulmonary vascular resistance in PAH [92]. Although analysis of lung tissue of patients is complicated and mainly impossible since tissue removal is a highly invasive approach, the stages of the PAH progression can be successfully analyzed based on animal models. To model pulmonary hypertension in animals, chemical reagent monocrotaline (MCT) is usually administered by subcutaneous or intraperitoneal injections. This drug selectively damages the endothelial cells of the pulmonary arteries and thus triggers the onset of pulmonary hypertension [93], creating the conditions which mimic PAH.

2.2.4 Collagen changes in thyroid cancer

Thyroid cancer is the most common endocrine malignancy, accounting for 3.4% of all cancers diagnosed annually [94]. It has a wide spectrum of clinical behavior and histological subtypes. The most common forms include PTC, FTC, medullary thyroid carcinoma (MTC) and anaplastic thyroid carcinoma (ATC). Among these, PTC and FTC, collectively referred to as differentiated thyroid carcinomas, account for more than 90% of cases and generally have a favorable prognosis [95]. However, a subset of these tumors exhibits aggressive behavior, including local invasion and distant metastasis, which is strongly influenced by changes in the TME – particularly in the ECM.

Similar to other cancers, the progression of thyroid cancer is accompanied by remodeling of the ECM. One of the major components of the ECM in thyroid tumors is collagen, particularly types I and III. ECM remodeling is accompanied by changes in collagen that include increased deposition, fiber realignment, and cross-linking. Different collagen subtypes are strongly associated with the development and progression of thyroid cancer. In view of this, collagen subtypes are potential therapeutic targets and biomarkers for cancer diagnosis [96].

In PTC, collagen remodeling is evident early in tumor development, with increased deposition of types I and III collagen observed in both the peritumoral and intratumoral regions [97]. A hallmark of PTC is the formation of a collagen-rich capsule that initially serves as a physical barrier, separating the tumor from the adjacent thyroid tissue. This fibrous capsule, composed primarily of type I collagen, is often associated with a desmoplastic reaction [97]. As the tumor progresses, the integrity of this capsule can be compromised by enzymatic degradation and remodeling by MMPs and other collagenases secreted by cancer cells and tumor-associated stromal cells. Destruction of the capsule facilitates the spread of tumor cells into the surrounding parenchyma and lymphatic vessels, contributing to the high propensity for lymphatic spread observed in PTC [98].

In FTC, the collagen changes are more diffuse and the tumor nodules are surrounded by a pronounced fibrotic encapsulation. Studies suggest that the tumor capsule in FTC often contains a dense collagenous stroma that may initially serve as a barrier to invasion. However, once this barrier is breached,

aligned collagen pathways may serve as conduits for vascular infiltration and hematogenous spread. Increased expression of collagen-modifying enzymes such as lysyl oxidase has also been associated with the promotion of metastatic potential in FTC [99].

2.3 SHG Microscopy for disease diagnostics

Alterations of collagen structure that could potentially serve as an early diagnostic marker have been studied with SHG microscopy in a number of types of cancer, such as breast [100], ovarian [101], skin [102], lung [103], colonic [104], prostate [105], thyroid [21,106] and other kinds of tumors.

SHG imaging is label-free and non-destructive. It enables high-contrast imaging without the need for external dyes or markers. Since no staining is required, collagen architecture remains intact, which is crucial for subsequent quantitative analyzes. These properties make SHG imaging a potential tool for in vivo imaging, with the results obtained on tissue sections serving as proof-of-concept for future experiments. In fact, there are already clinical variants of SHG imaging systems (e.g. JenLab's MPTFlex) already exist with applications in skin pathology [107]. SHG imaging belongs to a family of complementary nonlinear optical methods such as two-photon excitation microscopy (TPEF), third harmonic generation (THG) and coherent anti-Stokes Raman scattering (CARS), which generally require little or no modification to the microscope and can be used in parallel to provide different contrast sources on the same sample surface.

2.4 Image analysis for medical purposes

Medicine is one of the most dynamic and rapidly evolving fields, driven by the constant search for tools to improve diagnostic precision and therapeutic monitoring. Medical imaging has evolved from simple mechanical tools to sophisticated, technology-driven systems capable of visualizing the intricate structures and functions of the human body. Modern imaging encompasses a wide range of techniques, each based on different physical principles and tailored to specific clinical applications. The general medical image analysis workflow includes the following crucial steps: (i) extraction and selection of meaningful features of images, (ii) image segmentation and selection of regions of interest (ROIs), (iii) image classification based on the specific patterns of the images and relation of the images to the some medical condition (normal state, disease, type of the disease, stage of the disease, etc.).

2.4.1 Types of medical images

Common methods include ultrasound, which relies on the reflection of sound waves to image soft tissue and blood flow; radiography and computer tomography, which use ionizing radiation to capture internal anatomical details; MRI, which uses magnetic fields and radio waves to image soft tissue with high resolution, taking into account both functional and structural aspects. Positron emission tomography (PET) allows imaging of metabolism

by tracking radiotracers, while endoscopy provides direct visual access to internal organs. Thermography is an external imaging modality for dermatologic assessments. Histopathologic techniques such as immunohistochemistry (IHC) and hematoxylin and eosin (H&E) staining remain essential for microscopic assessment of cell architecture and biomolecular markers. Moreover, other forms of microscopy, such as SHG microscopy for visualization of collagen fibers and tissue microarchitecture, expand the toolkit of medical imaging [108].

In addition, all measurement and recording techniques such as electroencephalography and magnetoencephalography, which are not primarily used to generate images but produce data that can be displayed as maps, can be considered forms of medical imaging [109].

2.4.2 Major medical image features applicable for diagnosis

To extract clinically relevant information from medical images, robust analytical methods are required. Traditional quantitative methods start with FOS and focus on the intensity values of individual pixels to derive features (detailed descriptions in subsection FOS Analysis of PTC/FTC SHG images). While these measures are simple, they often lack spatial context.

To capture more complicated spatial patterns, second-order statistics such as the gray level co-occurrence matrix (GLCM) evaluate the frequency of pixel intensity pairs in specific spatial relationships, providing insights into texture properties (detailed descriptions in subsection SOS Analysis of PTC/FTC SHG images). More advanced higher-order statistics such as the gray level run length matrix (GLRLM) examine the length and distribution of consecutive pixels with identical intensity and provide a detailed characterization of texture (detailed descriptions in subsection HOS Analysis of PTC/FTC SHG images), which is particularly valuable in the analysis of pathological tissue [110].

In addition to spatial statistical methods, frequency domain analysis using the fast Fourier transform (FFT) is used to assess image anisotropy by calculating orientation indices (detailed descriptions in subsection FFT Analysis of SHG images of rat lung tissue samples). This is particularly effective in the analysis of fibrous tissue networks, such as collagen, where orientation and organization play a crucial role in disease progression.

Texture analysis in particular has emerged as a crucial method due to its balance of interpretability and diagnostic relevance. In contrast to purely abstract features, texture descriptors often match histopathologic patterns observed by clinicians, making them a valuable bridge between automated analysis and expert interpretation [111].

2.4.3 ML for image segmentation

With the advent of ML, the automation of image segmentation tasks has reached an unprecedented level of precision. Unsupervised learning techniques such as *k*-means clustering allow images to be divided into

meaningful regions without the need for extensive annotated datasets. These methods group pixels based on feature similarity and effectively highlight structures such as tumors, organs, and vascular networks in complex anatomical landscapes [112].

The integration of DL has significantly improved segmentation capabilities. Convolutional Neural Networks (CNNs), especially architectures such as U-Net and its variants, have shown exceptional performance in describing fine anatomical boundaries and complex tissue morphologies. In contrast to conventional ML methods, DL models automatically learn hierarchical feature representations directly from raw image data. This eliminates the need to create features by hand and significantly improves segmentation accuracy across different imaging modalities [113]. However, despite these considerable advantages, DL approaches also have their limitations. One major problem is the high computational cost, as training deep neural networks requires powerful hardware and extensive memory resources.

2.4.4 ML for image classification

In medical image analysis, classification tasks play a decisive role in diagnostics and decision-making processes, e.g. in differentiating between benign and malignant tumors or in identifying stages of disease progression. Importantly, each classification task is unique as it is highly dependent on the type of input data, the specific medical condition under investigation and the goals of the clinical application. Therefore, there is no universal model that is suitable for all tasks. Selecting the appropriate classification algorithm requires a careful understanding of the data set and the problem at hand.

There are different types of classification tasks. Two very different types of classification should be emphasized: Binary classification, when there are only two possible outcomes, and multi-label classification, when there are more than two possible outcomes.

When faced with a new classification problem, it is advisable to start with well-established, classical machine learning algorithms such as Random Forest (RF) [114], Logistic Regression (LR), eXtreme Gradient Boosting (XGBoost) [115], Light Gradient-Boosting Machine (LightGBM) [116], C-Support Vector Classification (C-SVC) [117] and Multi-Layer Perceptron (MLP) [118]. Not only are these models relatively quick to train and easy to tune, but they also offer a significant advantage in terms of interpretability. In particular, many of these algorithms can provide insight into the importance of features, allowing researchers and clinicians to understand which factors most influence classification results.

DL models, especially CNNs, have shown top performance in many image classification tasks. These models excel at learning complex, non-linear patterns directly from raw data and often outperform classical approaches in terms of accuracy, robustness to noise, and their ability to generalize to new, unseen datasets [119]. However, DL models largely act as "black boxes".

Unlike classical models, they do not readily reveal which features were most influential in the predictions. While post-hoc interpretation tools such as Grad-CAM [120] or integrated gradients exist, they do not yet offer the same level of clarity or reliability as the feature importance measures of classical algorithms. While DL is powerful, its opacity remains a limitation, especially in clinical settings where explainability is critical.

2.4.5 Medical image quality

The quality of medical images is a critical factor influencing the accuracy of diagnostic interpretations and the performance of ML models. Medical imaging data is often affected by various sources of noise. These include label noise (which stems from diagnostic ambiguities or inter-observer variability) and feature noise (which stems from acquisition artifacts, patient motion, or suboptimal imaging conditions).

Real-world clinical data also suffers from heterogeneity in imaging protocols, equipment, and operator skills, which complicates model training and deployment. To mitigate these challenges, advanced pre-processing techniques such as denoising algorithms, artifact correction and data augmentation are routinely used. In addition, harmonization strategies aim to standardize data between different imaging centers and thus improve the generalizability of models.

Newer DL-based approaches also offer robust solutions for noise reduction and quality improvement. Techniques such as adversarial training, self-supervised learning and noise-tolerant algorithms enable models to learn effectively from imperfect data. Furthermore, the inclusion of expert validation loops during training helps to correct mislabeled samples and ensures that the models are trained on highly realistic data [112,113].

As medical imaging continues to evolve, adherence to strict standards for image quality and data integrity remains essential to realize the full potential of ML and DL in clinical practice.

3 MATERIALS AND METHODS

3.1 Materials

3.1.1 Rat PAH model

Pulmonary arterial remodeling in rats, including the modification of the ECM, additional deposition of collagen, hypertrophy of vascular smooth muscle cells, thickening and muscularization of the pulmonary arteries was studied using MTC based PAH model.

PAH was chemically induced in the rats by injections of MCT (Sigma-Aldrich) at a dose of 60 mg/kg according to the procedure described in the study [121]. This drug selectively damages the endothelial cells of the pulmonary arteries and thus triggers the onset of pulmonary hypertension [93]. The studies were performed on 64 outbred white rats (Wistar, male, 200 g – 250 g) purchased from the Experimental Animal Center of the Belarusian Medical Academy for Postgraduate Education. Twelve animals were separated into a control group and considered healthy animals, while 52 animals were sorted into a PAH group treated with MCT.

After the MCT injection, the rats in the PAH group were randomly divided into four groups: 2 weeks, 4 weeks, 6 weeks and 8 weeks. Each group consisted of 13 animals. The results were compared with the corresponding data for healthy animals.

3.1.2 Preparation of the rat lung tissue sections

Sections of lung tissue were fixed in 10% neutral formalin for 48 hours. They were then washed in a stream of water for 24 hours and dehydrated in ethanol of increasing concentrations (70%, 80%, 90%, absolute ethanol). Then the tissue samples were passed through ethanol-xylene, xylene, xylene-paraffin and finally embedded in paraffin. Slices of 3 μm – 4 μm thickness of the prepared tissue were stained with H&E [122].

3.1.3 Preparation of the papillary and follicular thyroid carcinoma tissue sections

Tissue sections of PTC and FTC nodules were prepared according to standard histologic procedure [123]. Proper alignment of thyroid tissue is crucial for the preparation of tissue sections that provide valuable diagnostic information and allow visualization of cellular and structural details required for accurate histologic analysis [124]. Either the entire thyroid gland or one of its lobes, which were surgically removed, were examined macroscopically to identify important anatomical features such as the thyroid capsule, isthmus and possible lesions. To demonstrate the relationship of a focal lesion to the thyroid capsule, the thyroid gland was sliced perpendicular to the long axis of each lobe. In our case, the tissue contained nodules, so these areas were positioned so that they were cut at their longest extent to capture the full extent

of the pathology. The prepared thyroid tissue blocks were embedded in paraffin. Then, the formalin-fixed, paraffin-embedded tissue blocks were cut into 4 μm – 7 μm thick sections, placed on glass slides, and stained with H&E. Although H&E staining affects SHG imaging [125] and leads to a decrease in average pixel intensity, this does not affect the quantitative analysis of SHG data.

3.2 Experimental methods

3.2.1 Immunohistochemical study of rat lung tissue sections

Immunohistochemical study of the expression levels of molecular markers (collagen I, III and metalloptotease (TIMP) -1) was carried out using the following monoclonal antibodies: (i) polyclonal rabbit IgG for Collagen type I (Abcam), (ii) polyclonal rabbit IgG for Collagen type III (Thermo Fisher scientific), (iii) monoclonal mouse IgG1 for TIMP-1 (Thermo Fisher scientific).

For IHC studies, tissue samples were deparaffinized in xylol by two-step washing for 10-15 minutes for each step. Then the slices were rehydrated in alcohols of decreasing concentrations followed by washing in distilled water. Then heat-induced epitope retrieval was performed in a microwave oven in 0.01 M citrate buffer pH 6.0 (Carl Roth GmbH) for 10 minutes, preheating the retrieval buffer in accordance with the standard protocol as described in [126]. The IHC staining was performed according to the following protocol. Incubation with primary antibodies was carried out in a humid chamber for about 1–2 hours at 37°C or 24 hours at 4°C. Then, the slices were treated by Polymer Helper and Polyperoxidase-Anti-Mouse/Rabbit IgG, containing a complex of secondary antibodies and chromogen diaminobenzidine (DAB) (Elabscience), at 37°C for 20 and 30 minutes, respectively. After each step, the slices were rinsed in phosphate buffered saline. After staining with DAB, sections were counterstained with H&E.

Then the slices were placed in absolute ethanol two times for 7 minutes, afterwards in xylene two times for 10 minutes, and then the slices were embedded in *Canadian balsam* (AppliChem). To control the activity of primary antibodies (to exclude false positive and false negative results), one negative and one positive control staining were performed in each experimental series. As a negative control, the slices pretreated with 1% bovine serum albumin solution (Helix) instead of incubation with the primary antibody were used. For positive control, lung (TIMP-1), kidney (collagen I), and ovarian (collagen III) tissues were examined.

DAB serves as a chromogen that enables the visualization of antibody-antigen reactions in IHC. At the site where the antibody binds to the antigen (collagen I, III and TIMP-1), a brown precipitate is formed in the tissue section by DAB. The brown colored areas of the tissue sections reflect the degree of protein expression in the tissue sample and can be quantified to determine the index of expression.

The quantitative assessment of the expression of biomolecular markers was carried out by analyzing digital images obtained with Leica DMLS microscope using the pre-installed software and a JVC digital camera (at 400× magnification, at least 30 view areas). The *positive pixel count* algorithm and software for morphometry *AperioImageScope12.1.0.5* were applied.

The image analysis yielded data on the prevalence and intensity of the brown color of the DAB reaction products in tissue slices. Digital images of non-overlapping areas with clearly defined nuclei, cells and vessels of the lungs were selected. The index of expression (IE) of biomolecular markers was calculated using the *Equation* 23:

Equation 23: $$IE = \frac{Number\ of\ positive\ pixels}{Total\ number\ of\ pixels} \times 100$$

3.2.2 Brightfield imaging

Lung tissue samples from healthy animals (control group) and from animals with PAH at weeks 2, 4, 6 and 8 of disease progression were examined by a combination of brightfield, TPEF and SHG microscopy of H&E-stained samples simultaneously on a single, custom-built wide-field nonlinear microscopy setup, described below. Bright-field color transmission images of lung tissue samples were acquired using a *DCC1645C-HQ* Complementary metal–oxide–semiconductor (CMOS*)* camera (Thorlabs).

Thyroid carcinomas tissue sections were imaged using a brightfield *Aperio LV1 IVD Whole Slide Scanner* (Leica Biosystems) with a 20× objective. Labeling of the bright-field images and identification of the ROI of capsular invasion were performed by experienced histopathologists.

3.2.3 Wide-field SHG and TPEF microscopy imaging

A custom-built wide-field non-linear microscopy setup based on a modular microscope (Applied Scientific Instruments), described in detail in [22], was used. A sample area of approximately 150 µm × 150 µm was illuminated with a FemtoLux3 laser (Ekspla) featuring a 1030 nm wavelength, circular polarization state, 262 fs pulse duration, 1 MHz pulse repetition frequency, and 1.5 W of average power at the sample. The excitation wavelength was chosen because its second harmonic can be easily detected by standard CMOS detectors. The laser power remained constant during measurements and was always set to the maximum generated by the laser. The laser beam was expanded 4× with an adjustable beam expander VEX18 (Optogama) and focused onto the back focal plane of the illumination objective using an achromatic lens with a 250 mm focal distance. The resulting image was detected in the forward direction using a UPlanFL N 40×/0.75 objective (Olympus) and a Neo 5.5 sCMOS camera (Andor Technology Ltd) with a pixel size of 6.5 µm, providing a lateral resolution of approximately 0.5 µm. The SHG signal was separated by a high-pass dichroic mirror T900LPXXRXT (Chroma Technology GmbH) and a bandpass

dielectric filter FF01-513/13-25 (Semrock), while fluorescence signals were separated using long-pass optical filters. The scheme of the SHG imaging setup is shown in *Figure 3.1*.

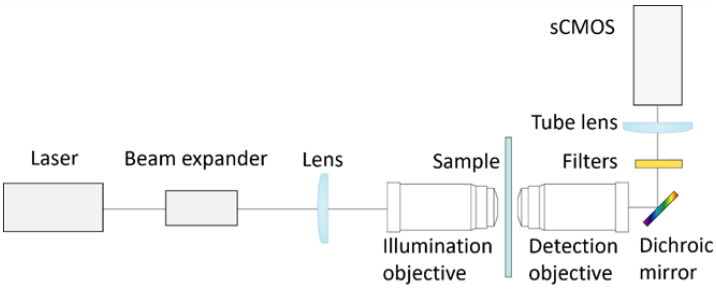


Figure 3.1 The scheme of the SHG imaging setup. Reprinted from [***Paper C***].

Tiled images of 150 µm × 150 µm of rat lung tissue sample areas were obtained by scanning the sample with a motorized mechanical stage S551-2201B (Applied Scientific Instruments). Both SHG and fluorescence image integration times were 0.5 s for lung tissue samples, and scanning a tiled image of 2.1 mm × 2.1 mm took approximately 2.6 minutes.

Large sample areas of thyroid tissue sections were scanned with the same motorized stage. The camera integration time was 0.1 s. The overall acquisition time of one sample was less than 40 minutes. In total, 5 sections of whole PTC nodules and 5 FTC tissue sections were imaged for this work.

3.2.4 Polarization-resolved SHG imaging

For polarization-resolved wide-field SHG imaging, a custom designed microscope, described in 3.2.3, was upgraded by integrating input laser polarization control. This involved incorporating a polarization state generator (PSG), which comprised a half-wave plate AHWP05M-980 (Thorlabs) and a quarter-wave plate AQWP05M-980 (Thorlabs) positioned in motorized rotation mounts PRM1 (Thorlabs). Prior to imaging, polarization state calibration was conducted for each desired linear polarization state using a polarimeter PAX1000IR1 (Thorlabs) in front of the illumination objective. The input laser polarization control process resembled the one outlined previously [127], resulting in values less than 0.1 for the ellipticity of the linear polarization states. The *LabVIEW* code, specifically designed for controlling the wide-field SHG microscope, was modified to facilitate the automated control of the PSG. Polarization-resolved image stacks were captured at various linear polarization angles ranging from 0° to 180° in 20° increments. Each polarization state was imaged with an exposure time of one second, with a 3-5 second delay time for switching between polarization states based on the transitions. A single PSHG stack, covering a 300 µm × 300 µm area, was recorded within 45 seconds. To encompass the entire thyroid nodule, a mosaic of 20×20 individual PSHG stacks was captured, taking approximately 8 hours. Automated sample movement was ensured by a

motorized XY stage S551-2201B (Applied Scientific Instruments). The imaging duration was limited not by the camera exposure itself but by the transition times between polarization states. While the recorded images were 2048 px × 2048 px only the central part of each stack of 1500 px × 1500 px corresponding to 219 µm × 219 µm was retained after cropping to remove any uneven illumination artifacts. The final mosaic covering the entire nodule was assembled using a custom-written ImageJ macro.

3.3 Data preprocessing algorithms and ROI selection

3.3.1 ROI Selection in SHG images of lung tissue samples

Due to the high signal intensity, no specific preprocessing of the images of lung tissue samples was performed. The selection of ROIs for analysis was performed according to the following procedure. The square ROIs with a size of 150 µm × 150 µm were selected so that they were no further than 500 µm from the blood vessel wall (see *Figure 3.2*). This selection was made to detect the remodeling of the collagen network in the lung tissue that accompanies the progression of PAH and leads to the main manifestations of the disease in the form of vasoconstriction, congestion, increase in blood pressure, etc. For each experimental group (samples from healthy animals and from animals at different stages of PAH progression), 50-80 non-overlapping ROIs were selected (see *Table 3-1*, for the number of ROIs selected for each experimental group).

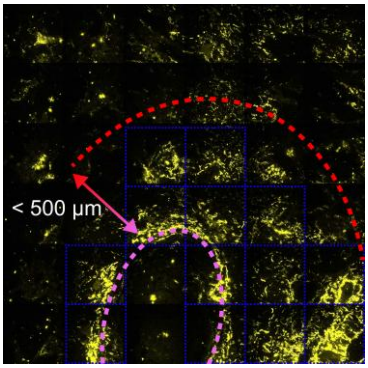


Figure 3.2 Scheme of ROI selection in large-scale SHG images of rat lung tissue samples. Dashed red line indicates the annular region for ROI selection around the vessel, blue squares indicate the ROIs. Dashed magenta line marks the blood vessel wall. Adapted from [***Paper A***].

Table 3-1 Number of ROIs selected for analysis. Adapted from [***Paper A***].

Experimental group	Number of ROIs
Control	48
PAH 2 weeks	78
PAH 4 weeks	52
PAH 6 weeks	43
PAH 8 weeks	47
All	268

3.3.2 Preprocessing of non-polarized SHG images of PTC/FTC

For PTC and FTC samples, the schematic workflow of the SHG image preprocessing is shown in *Figure 3.3*.

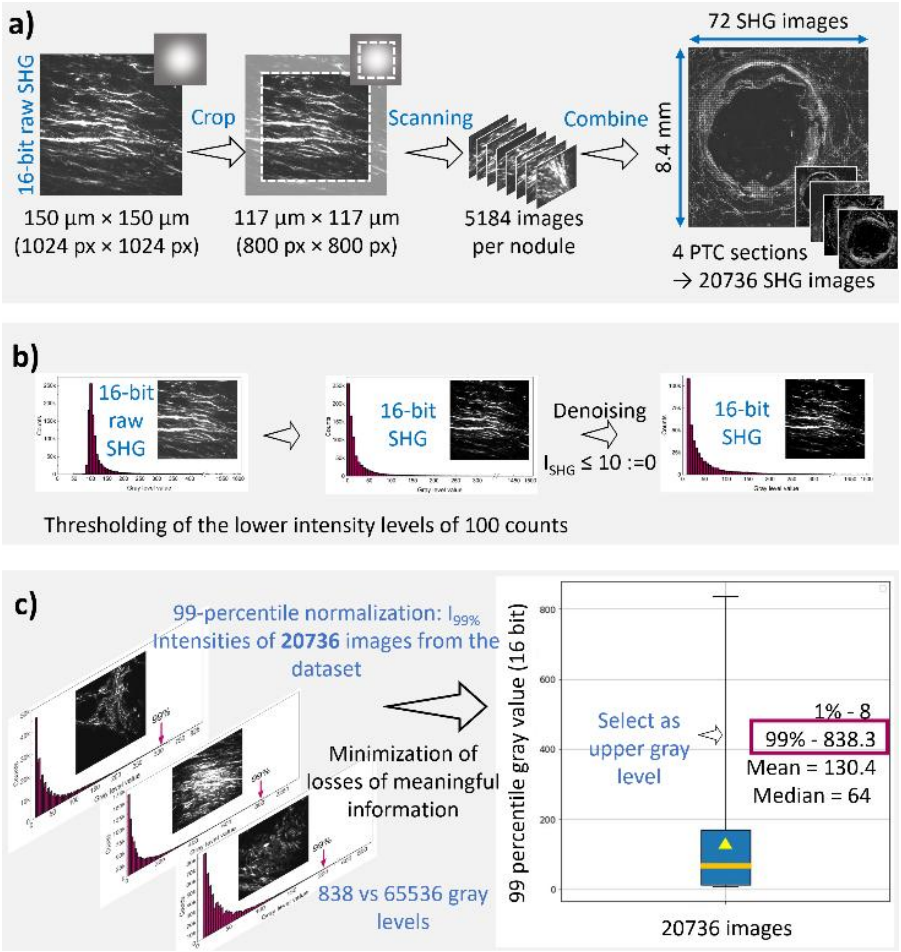


Figure 3.3 PTC/FTC SHG image preprocessing workflow: a) SHG scanning and cropping, b) thresholding and denoising, c) histogram normalization. Reprinted from [***Paper C***].

Each SHG image of a 150 µm × 150 µm sample area was centrally cropped to 117 µm × 117 µm to reduce the impact of illumination inhomogeneity due to the Gaussian laser beam intensity profile (*Figure 3.3a*). Then, images were thresholded at 110 counts, with 100 counts representing the electronic camera threshold, and an additional 10 counts allocated to account for the camera-associated background noise (*Figure 3.3b*).

For the subsequent texture analysis, the usual 16-bit image conversion to 8-bits was not performed. Instead, the reduction of the number of considered grey levels in an image was carried out in the following way. The 99th

percentile was calculated for each image from a set to remove random intensity outliers within each image [128]. Then the 99th percentile of the resulting distribution was calculated to exclude outlier images. The latter appeared to be 838 and was set as the upper threshold for all images (*Figure 3.3c*). Such a procedure enabled preservation of valuable texture information that is often lost during the 16-to-8-bit conversion.

3.3.3 PSHG Parameter map preprocessing

Each PSHG image stack was subjected to the processing steps described in detail in [123]. The experimental data fitted pixel-by-pixel using a theoretical curve that characterizes the changes in SHG intensity from collagen with the linear polarization of the input laser. This model for collagen, known as the single-axis molecule model [70], depicts the SHG intensity as follows:

Equation 24: $$I_{SHG} \sim [\chi_{15}^2 \cdot sin^2\, 2\,(\varphi - \alpha) + \\ +(\chi_{31} \cdot sin^2(\varphi - \alpha) + + \chi_{33} \cdot cos^2(\varphi - \alpha))^2].$$

Here, α represents the polarization orientation of the excitation beam, φ denotes the in-plane orientation of collagen, and χ_{15}, χ_{31}, χ_{33} are the only nonzero elements of the macroscopic nonlinear susceptibility tensor $(\chi^{(2)})$ assuming cylindrical symmetry of collagen.

For polarization resolved SHG image analysis, the preprocessing was performed differently due to the use of a modified experimental setup. The maps of polarization-related parameters of 30000 px × 30000 px, including χ_{31}/χ_{15}, χ_{33}/χ_{15}, χ_{33}/χ_{31}, and θ_e were subjected to a filtering process to exclude pixels with $R^2 < 0.8$. Subsequently, the maps were downsampled by calculating average parameter values within image tiles measuring 100 px by 100 px.

3.4 Feature extraction algorithms

3.4.1 FFT Analysis of SHG images of rat lung tissue samples

The anisotropy of the collagen fiber network was quantified by the orientation index (OI) of ROI images of lung tissue samples. The transformed image from each ROI (*Figure 3.4b*) was binarized with a threshold of 0.38, computed by the *Otsu method* [129]. Despite the presence of some residual noise, a binary image with a significant elliptical structure was obtained (*Figure 3.4c*). A predefined 2-D circular averaging filter for remaining noise removal (radius = 3 px) was applied (*Figure 3.4d*). The received ellipse was approximated by second order curve using a least squares fitting method (*Figure 3.4e*) [130].

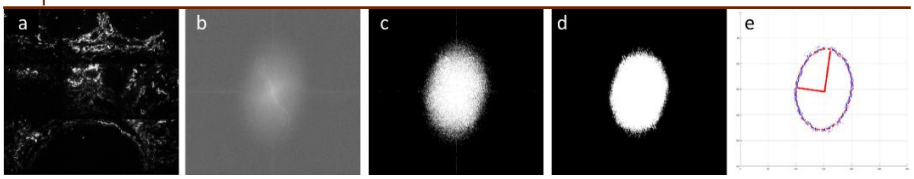


Figure 3.4 FFT processing and ellipse parameters calculation: a – image of a selected ROI, b – the result of FFT of the SHG image, c – result of binarization, d – 2-D circular averaging filter implementation, e – second order curve approximation. Reprinted from [***Paper A***].

OI is calculated through the long-to-short ellipse axis ratio obtained via FFT analysis of the ROI [131] by the following formula:

Equation 25: $OI = \left[1 - \left(\frac{short\ axis}{long\ axis}\right)\right]$.

That is the $OI = 0$ and $OI = 1$ correspond to the randomly and perfectly oriented collagen fibers, respectively.

3.4.2 FF-PSHG Analysis of PTC samples

FF-PSHG analysis [132] was employed to extract the collagen's biophysical parameters based on the proposed model. In brief, the FF-PSHG calculates the Discrete Fourier Transform (DFT) coefficients for each pixel in the PSHG image set, which are then utilized to determine the free parameters of the collagen model. Through this pixel-by-pixel fitting approach, three images were computed depicting ratios of $\chi^{(2)}$ elements (χ_{31}/χ_{15}, χ_{33}/χ_{15} and χ_{33}/χ_{31}), one revealing the collagen's angular distribution (φ), and another showing the orientation of the dominant axis of the hyperpolarizability tensor (θ_e). The latter was previously correlated with the helical pitch angle of the collagen triple helix and can be defined as follows [70]:

Equation 26: $\cos^2\theta_e = \frac{\chi_{33}/\chi_{15}}{2+\chi_{33}/\chi_{15}}$.

Three assessment methods for the fitting inherent to the FF-PSHG analysis were used to eliminate pixels unsuitable for statistical testing: the coefficient of determination (R^2), a signal-to-noise ratio (*SNR*) [133] and an experimental error (*ERR*). All were determined on a pixel-by-pixel basis, resulting in the generation of corresponding maps. *SNR* and *ERR* were calculated based on DFT coefficients having biophysical significance according to the collagen model [134].

3.4.3 FOS Analysis of PTC/FTC SHG images

FOS describes the distribution of pixel intensity values in an image without considering their spatial arrangement. The calculated FOS parameters are the following: mean (μ_1), standard deviation (σ), skewness or asymmetry

(g_1), and kurtosis (g_2) of the intensity distribution histogram. FOS parameters are extracted from the gray level histogram (inter-pixel correlations were ignored).

The different FOS parameters yield collagen content [135], its uniformity, and density [136]. Specifically, μ_1 and σ relate to the amount of collagen and homogeneity of its distribution in the sample. The g_1 value is a measure of asymmetry of the distribution of pixel intensities. The g_2 value indicates how closely the distribution of pixel intensities resembles the normal distribution. Strong SHG signals usually result in a wide distribution of pixel intensities, and consequently, low value of g_2 [135]. Detailed description and mathematical expressions of FOS parameters, as well as their relationship to the collagen fiber content and spatial organization are presented in *Table 3-2*.

Table 3-2 FOS parameters description with assignment to the collagen texture. Adapted from [***Paper C***].

Parameter	Mathematical expression	Description	Assignment to the collagen texture (examples)
Mean	$\mu_1 = \frac{1}{N^2}\sum_{i,j=0}^{N-1} I_{i,j}$	A mean value of image intensity.	High μ_1 – high collagen content (amount of collagen fibers is proportional to the detected SHG signal) [137]. Low μ_1 – thin and sparse fibrillar organization [35].
Standard deviation	$\sigma = \sqrt{\mu_2} = \frac{\sqrt{\sum_{i,j=0}^{N-1}(I_{i,j}-\mu_1)^2}}{N}$	A deviation of the image intensity from the mean value.	Overall contrast of the image [137].Low σ – homogeneous, high σ – heterogeneous collagen distribution in ROI.
Skewness	$g_1 = \frac{\mu_3}{\mu_2^{3/2}} = \sigma^{-3}\frac{\sum_{i,j=0}^{N-1}(I_{i,j}-\mu_1)^3}{N^2}$	A measure of the degree of asymmetry of the pixel intensity distribution, i.e., an extent of darker (left skewness) or brighter (right skewness) pixels compared to the mean value [35,137]	High g_1 (right-skewed intensity distribution) – presence of several thick fibrils [35]. g_1 allows detection of edges of fibers or level of difference from background [137], positive for darker and glossier surfaces [137].
Kurtosis	$g_2 = -3 + \frac{\mu_4}{\mu_2^2} =$ $= -3 + \sigma^{-4}\frac{\sum_{i,j=0}^{N-1}(I_{i,j}-\mu_1)^4}{N^2}$	A measure of the pixel intensity histogram sharpness. >0 – leptokurtic distribution (positive g_2, wide tails) =0 – mesokurtic distribution (the pixel intensity distribution coincides with the normal distribution) <0 – platykurtic distribution (negative g_2, narrow tails) [35,137]	Low kurtosis – developed collagenous networks covering large areas and generating strong SHG signals (a wide distribution of pixel intensities) [35].

Note: $\mu_k = \frac{1}{N^2}\sum_{i,j=0}^{N-1}(I_{i,j}-\mu_1)^k$ – the k[th] moment about the mean μ_1, where $I_{i,j}$ – intensity of a pixel, N^2 – number of pixels for a square image.

3.4.4 SOS Analysis of PTC/FTC SHG images

SOS has to do with the relative positions of the different gray levels within an image. The SOS parameters are obtained from the GLCM, which

represents the occurrence of two pixels with certain gray level values i and j at a distance d and at an angle θ relative to each other. For a rectangular image with M rows and N columns, having $M \times N$ pixels, the GLCM is defined by the following equation [138]:

Equation 27: $GLCM_{i,j}(\theta, d) =$

$$= \sum_{x=1}^{M} \sum_{y=1}^{N} \begin{cases} 1 \; if \; I(x,y) = i; \\ 1 \; if \; I(x + d\,cos(\theta), y + d\,sin(\theta)) = j. \\ 0 \; otherwise. \end{cases}$$

To eliminate the dependence of GLCM on the chosen direction θ, the GLCM was averaged over 4 directions of $\theta \in \{0°, 45°, 90°, 135°\}$:

Equation 28: $s_{i,j}(d) = \frac{1}{4} \sum_{\theta} GLCM_{i,j}(\theta, d)$.

The direction-averaged GLCM was calculated for 5 values of the distance $d \in \{1, 3, 6, 9, 12\}$. The step size of 3 px was chosen because it is commensurate with the optical resolution of the experimental setup. The texture parameters calculated from the GLCM are the following: energy or uniformity or angular second moment (E), contrast or inertia (I), correlation (C), local homogeneity or inverse difference moment (L), and entropy (H) [139]. Detailed description and mathematical expressions of SOS parameters, as well as their relationship to the collagen fiber content and spatial organization are presented in

Table 3-3 SOS parameters description with assignment to the collagen texture. Adapted from [**Paper C**].

Parameter	Mathematical expression	Description	Assignment to the collagen texture (examples)
Energy	$E(d) = \sum_{i,j=0}^{N_G-1} [s_{i,j}(d)]^2$	A measure of how often similar pixel values are found together at the distance *d*. Low *E* – less smooth ROI (uniformly distributed $s(d)$, no dominant levels of brightness, gray levels are equally probable [35]) [140–142].	Low *E* – disordered fibers, low uniformity [143].
Inertia	$I(d) = \sum_{i,j=0}^{N_G-1} (i-j)^2 s_{i,j}(d)$	A measure of the local intensity variation (heterogeneity). High *I* value – high variability of gray levels between neighboring pixel at the distance *d* (high-contrast ROI). Low *I* – homogenous ROI [35,140–142].	High *I* – highly dense collagen bundles ('a collagen network with well-formed fiber containing large amounts of collagen' [144]), foci of collagen synthesis, no preferential alignment of orientation [35], wavy fibers or randomly arranged fibers [143].
Correlation	$C(d) = = \frac{\sum_{i,j=0}^{N_G-1}(i-\mu_x)(j-\mu_y)s_{i,j}(d)}{\sigma_x \sigma_y}$	A metric of periodicity (correlation between gray levels of neighboring pixels at the distance *d* in two different directions) [35]. $C = 0$ for completely uncorrelated gray levels (no regular structure). $C = \pm 1$ for positively/negatively correlated levels (texture structures are repeated) [140–142]	High C – the presence of periodic structures (e.g., spatially ordered collagen fibers) [143] or one or several repeated patterns [35].
Local homogeneity	$L(d) = = \sum_{i,j=0}^{N_G-1} \frac{1}{1+(i-j)^2} s_{i,j}(d)$	A measure of the local homogeneity of an image. If the *L* value increases, the incidence of pixels' pairs co-occurrence is enhanced. High *L* – the image is homogeneous [140–142].	High *L* – dense or thick collagen fibers [35]. Low *L* – spreading of the network of thin and disordered collagen fibers and the development of fibrosis [145].
Entropy	$H(d) = = -\sum_{i,j=0}^{N_G-1} s_{i,j}(d) \log[s_{i,j}(d)]$	A measure of how random the pairs of pixels at distance *d* in the GLCM are distributed (spatial organization inside ROI). High *H* – rough, coarse-grained textures, high structural complexity [35]. Low *H* – smooth or homogeneous images [35], high regularity degree [140–143].	A degree of fiber organization [137]. Low *H* – poorly separated fibers, homogeneous local collagen morphology [35]. High *H* – bright and distinct but not necessarily ordered collagen fibers standing out from a homogeneous background, structural complexity [143]; Extensively disorganized collagen fibers [145], disordered collagen network [144].

Note: $s_{i,j}(d) = \frac{1}{4}\sum_\theta GLCM_{i,j}(\theta,d)$. $\sigma_x^2 = \sum_{i=0}^{N_G-1}(i-\mu_x)^2 \sum_{j=0}^{N_G-1} s_{i,j}(d)$, $\sigma_y^2 = \sum_{j=0}^{N_G-1}(j-\mu_y)^2 \sum_{i=0}^{N_G-1} s_{i,j}(d)$, $\mu_x = \sum_{i=0}^{N_G-1} i \sum_{j=0}^{N_G-1} s_{i,j}(d)$, $\mu_y = \sum_{j=0}^{N_G-1} j \sum_{i=0}^{N_G-1} s_{i,j}(d)$.

3.4.5 HOS Analysis of PTC/FTC SHG images

HOS describes the properties of groups of pixels with different gray level values occurring at specific locations relative to each other. The HOS parameters were calculated from the GLRLM, which represents the occurrence of a group of pixels of a certain size (run length) *j* with the same

gray level i in a certain direction θ [146]. For the same rectangular image ($M \times N$ dimension):

Equation 29: $GLRLM(g,r|\theta) = \sum_{i=1}^{M}\sum_{j=1}^{N}\delta\left(I_{i,j},g\right)\cdot\delta\left(L_{i,j}(\theta),r\right)$.

Here, g represents the gray level $g \in \{1,2,\ldots,N_g\}$, where N_g is the total number of gray levels; r is the length of a consecutive run of pixels with the same gray level $r \in \{1,2,\ldots,N_r\}$, where N_r is the maximum run length; $\delta(x,y)$ is the Kronecker delta function, which equals 1 if $x = y$ and 0 otherwise; $L_{i,j}(\theta)$ is the length of a continuous run of pixels with the same gray level g, starting from position (i, j) in the direction θ.

To eliminate directionality, GLRLM averaged over 4 directions:

Equation 30: $p_{i,j} = \frac{1}{4}\sum_{\theta} GLRLM(\theta)$.

The texture parameters calculated from the GLRLM are the following: short run emphasis (*SRE*), long run emphasis (*LRE*), gray-level non-uniformity (*GLN*), run length non-uniformity (*RLN*), and run percentage (*RP*) [139]. Detailed description and mathematical expressions of HOS parameters, as well as their relationship to collagen fiber content and spatial organization are presented in *Table 3-4*.

Table 3-4 HOS parameter description with assignment to the collagen texture. Adapted from [***Paper C***].

Extracted parameter	Mathematical expression	Description	Assignment to the collagen texture (examples)
Short Run Emphasis (*SRE*)	$SRE = \frac{1}{T_R}\sum_{i=0}^{N_G-1}\sum_{j=1}^{N_R}\frac{p_{i,j}}{j^2}$	A measure of the distribution of short runs (emphasizes short runs of pixels) [146]. The *SRE* value is high for fine-grained textures [147,148].	High *SRE* – a fragmented collagen network; small and fine bundles of fibers [144].
Long Run Emphasis (*LRE*)	$LRE = \frac{1}{T_R}\sum_{i=0}^{N_G-1}\sum_{j=1}^{N_R} j^2 p_{i,j}$	A Measure of the distribution of long runs (emphasizes long runs of pixels) [146]. The *LRE* is large for coarse structural textures [147,148].	High *LRE* – the presence of more long runs in the image, corresponding to coarse features, possibly due to large collagen bundles present in the network, with a certain level of orientation of bundles [144]
Gray Level Nonuniformity (*GLN*)	$GLN = \frac{1}{T_R}\sum_{i=0}^{N_G-1}\left(\sum_{j=1}^{N_R} p_{i,j}\right)^2$	A measure of the similarity of gray level values throughout the image. The *GLN* is small for gray levels that are alike throughout the image [147].	High *GLN* – regions of structural complexity or heterogeneity (for PET images, not collagen) [149]. Low *GLN* – a homogeneous image, a low amount of collagen, a loose network of the curled morphology [144].
Run Length Nonuniformity (*RLN*)	$RLN = \frac{1}{T_R}\sum_{j=1}^{N_R}\left(\sum_{i=0}^{N_G-1} p_{i,j}\right)^2$	A measure of the similarity of the length of runs. The *RLN* is small if the run lengths coincide throughout the image [146,147].	High *RLN* (in PET images of neuroblastoma) – high intratumor heterogeneity [149].
Run Percentage (*RP*)	$RP = \frac{1}{T_P}\sum_{i=0}^{N_G-1}\sum_{j=1}^{N_R} p_{i,j} = \frac{T_R}{T_P}$	A measure of the homogeneity and the distribution of runs of an image in a specific direction. The largest *RP* corresponds to the case when the length of runs is 1 for all gray levels in specific direction [147].	High *RP* – a large portion of the image is covered by runs [144].

Note: $p_{i,j} = \frac{1}{4}\sum_{\theta} GLRLM_{i,j}(\theta)$; $T_R = \sum_{i=0}^{N_G-1}\sum_{j=1}^{N_R} p_{i,j}$; T_P is the number of pixels in an image.

3.4.6 Statistical analysis of image parameter distributions for rat lung tissue samples

The statistical significance of the difference between the distributions of various parameters calculated on images from healthy control group and PAH animals was performed by statistical analysis of variance (one-way ANOVA) by applying an unpaired two-tailed Student's T-test.

3.5 Unsupervised ML

3.5.1 Data standardization and Pearson's correlation

Since the magnitudes of the different intensity and texture parameters are widely disparate, in order to be able to use the calculated distributions in the joint analysis, they were standardized using the Robust Scaler algorithm according to the formula:

Equation 31: $\tilde{X} = \frac{X - Median}{IQR}$,

where X is the parameter, *Median* is the median value of the distribution of the parameter, interquartile range (IQR*)* is the (1, 99) percentile range, and $\tilde{X}$ is the standardized parameter.

To simplify the analysis, first, redundant parameters with strong correlations were determined and then excluded from further analysis. For that, Pearson's correlation was calculated for all combinations of parameters according to the formula:

Equation 32: $Corr(\tilde{X}, \tilde{Y}) = \frac{Cov(\tilde{X}, \tilde{Y})}{\sigma_{\tilde{X}} \sigma_{\tilde{Y}}}$,

where $Corr(\tilde{X}, \tilde{Y})$ is a correlation matrix, $Cov(\tilde{X}, \tilde{Y})$ is the covariance matrix, and $\sigma_{\tilde{X}}$ and $\sigma_{\tilde{Y}}$ are the variances of the parameter distributions $\tilde{X}$ and $\tilde{Y}$, respectively. The Pearson's correlation is a measure of the strength of the linear relationship between two continuous variables and ranges from -1 to +1. A value of -1 indicates a perfect negative linear correlation, 0 - no correlation, and +1 - a perfect positive correlation. Correlation coefficients between 0.36 and 0.67 (absolute values) indicate moderately (positively or negatively) correlated variables [150]. Parameters with $Corr(\tilde{X}, \tilde{Y}) \geq 0.9$ were excluded from further analysis.

3.5.2 PCA

PCA is a non-parametric method for extracting valuable data from the original dataset. PCA was applied to the remaining set of intensity and texture parameters. PCA was performed using a singular value decomposition [151] of the data to project it into a low-dimensional space. PCA comprises a linear transformation of the original data set of correlated variables into a low-dimensional set of representative variables that together capture most of the information present in the original data by calculating the so-called principal components (PCs) [152–155]. The PCA is based on three main concepts: (i) eigenvalues of the covariance matrix of the original variables, which represent the data variance along each new dimension; (ii) orthogonal eigenvectors, which form a basis of the new data space and are expressed in terms of loadings related to the original variables; and (iii) scores, which represent the coordinates of the observations (data) in a new low-dimensional space. The number of PCs reflecting the most of the variability in the original data set was determined according to Kaiser's Rule with Jolliffe threshold [153]:

Equation 33: $\lambda_i > \frac{T}{n} \sum_{i=1}^{n} \lambda_i$,

where λ_i are the eigenvalues of the covariance matrix $Cov(\tilde{X}, \tilde{Y})$, $T = 0.7$ is the threshold value, and n is the number of eigenvalues.

3.5.3 *k*-Means clustering

k-Means clustering belongs to a group of unsupervised ML algorithms and allows data to be categorized into classes by grouping observations with similar sets of parameters. Each observation is considered as a point in the multidimensional space of parameters and the grouping of observations is achieved by minimizing the distance between points in the multidimensional Euclidean space [156]. In our case, each tile SHG image corresponds to one such observation.

The clustering of the data was performed with *k*-means based on Lloyd's algorithm [157]. The algorithm "*k*-means++" [158] was used to initialize the centroids of the clusters. The maximum number of iterations for a single run was set to 300. The efficiency of clustering and the meaningful number of clusters was assessed from the clustering metrics of [151] Silhouette coefficient (SC), Davies-Bouldin index (DBI), and Calinski-Harabasz index (CHI) Table 3-5.

Table 3-5 Metrics used for the evaluation of clustering algorithm efficiency. Adapted from [***Paper C***].

Metrics	Silhouette Coefficient	Davies-Bouldin index	Calinski-Harabasz index
Algorithm description	A measure of similarity of an object to its cluster in comparison with other clusters. It is composed of two scores: The mean distance between a sample and all other points in the same class, The mean distance between a sample and all other points in the next nearest cluster.	A measure that compares the distance between clusters with the size of the clusters themselves.	A measure of clustering quality based on the ratio of the sum of between-clusters dispersion and of within-cluster dispersion for all clusters.
Best separation	$s = 1$	$\rightarrow 0$	>>0 (dense and well separated clusters)
Not separated data	$s = -1$	>>0	Low values
Overlapping clusters	$s = 0$	–	–

3.6 Supervised ML

The schematic workflow outlining dataset preparation, strategies of handling with feature noise and label noise, model optimization, training, testing and generalization is shown in *Figure 3.5*.

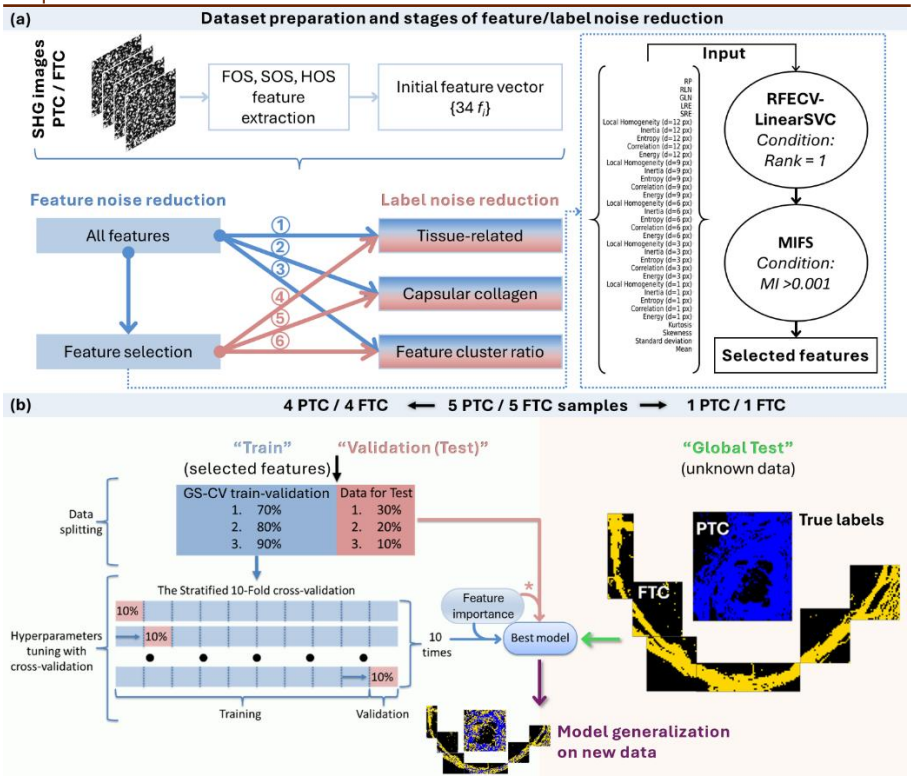


Figure 3.5 The schematic workflow of (a) dataset preparation, feature extraction, feature selection, and label noise reduction, and (b) optimization and generalization of the ML classification models. An asterisk indicates that, for PIA (feature importance analysis for C-SVC and MLP), the validation set was used. Numbers 1-3 correspond to dataset preparation with label noise reduction approaches without feature noise reduction (feature selection). Numbers 4-6 correspond to dataset preparation with feature selection followed by label noise reduction approaches. Feature selection consists of RFECV-LinearSVC for reducing redundant features and mutual information feature selector (MIFS) for removing target-irrelevant features. MI – mutual information parameter. Reprinted from [***Paper D***].

3.6.1 Managing label noise

Since the most obviously mislabeled data relates to the SHG images of glass slide, glass-related SHG images were excluded from further consideration except for the multiclass classification, where these data were added as a separate class.

Label noise reduction was done according to the following label correction approaches:

I. All tissue-related data (original dataset excluding images of glass and intensity outliers) was used, assuming that collagens in capsule and

surrounding tissue contain features relevant to target (label PTC or FTC) and reflect the tumor progression.

II. Only capsular collagen-related data was used with assumption that surrounding tissue does not contain any noticeable features of tumor progression. The capsular collagen was separated by applying the same algorithm (as in approach I) to all tissue-related data.

III. Multi-class classification, in which glass and non-capsular collagen data from both PTC and FTC were sorted in one class, with an assumption that these data are not relevant to the target (label PTC or FTC) and are identical in both carcinomas. An assumption that PTC and FTC collagen capsules may be heterogeneous in intensity and texture features extracted from SHG images was made. To take this into account, PCA and multi-cluster *k*-means were applied to segment both carcinomas according to patterns in texture features. The detailed description is provided in section 4.4.

Prior to further analysis, the datasets of the standardized features from the selected SHG images were reset to the original, non-standardized values. The scaling of the validation and the global test sets is further performed with the scaling parameters that are determined during the optimization of the ML classifiers made on the training set.

3.6.2 Managing feature noise

The feature noise reduction process aimed to remove redundant and non-target-related features was done through a two-step feature selection approach *Figure 3.5a*. First, Recursive Feature Elimination with Cross-Validation with LinearSVC estimator (RFECV-LinearSVC) was applied to remove redundant features. RFECV-LinearSVC feature selection enables the best classification performed on texture features of images compared to other feature selection techniques, including subgroup-based multiple kernel learning, RFE with naive Bayes/bagged trees/RF and LDA classifiers, etc. [159]. This method ranks features in descending order by recursively considering feature subsets of decreasing sizes [159]. Feature scores were averaged across cross-validation folds, and the optimal number of features was selected to maximize the cross-validation score. Only features with Rank = 1 were selected for further analysis [160].

Next, Mutual Information Feature Selector (MIFS) [161] was used to identify features relevant to the target labels (PTC or FTC) after RFECV-LinearSVC selection. MIFS allows measuring feature-target relations and significantly improves the classification results when applied together with feature selectors which use feature importance scores for selection [42]. Mutual information (MI) measures the dependency between features and labels, with MI = 0 indicating independence. Features with higher MI values are considered more relevant to a label [162]. A selection threshold of MI>0.001 (in nat units) was applied.

Feature selection was applied independently on the training sets for each label noise reduction approach (subsection 3.6.1) prior to optimizing the classifiers by stratified *k*-fold cross-validation (subsection 3.6.5). The most discriminative and target-relevant feature sets, selected by RFECV-LinearSVC and MIFS, were then used for the classifiers' optimization.

To estimate whether feature noise and/or label noise hamper the performance of the classification models, both initial datasets containing all features and datasets treated with RFECV-LinearSVC and MIFS were used in label correction approaches I-III (see *Figure 3.5a*).

3.6.3 Data separation

A total of 23652 SHG images were obtained for PTC and 21708 for FTC, ensuring that the initial datasets were balanced. One complete PTC and one complete FTC sample were set aside as unknown data (global test set) for the final validation and generalization of the trained ML classifiers.

The dataset of feature vectors from SHG images, representing 4 PTC and 4 FTC samples, was randomly split into a training set and a validation set using three different ratios: 70/30%, 80/20%, and 90/10%. Data was split in a stratified manner. Each training dataset was scaled using the Robust scaler algorithm with (1, 99) percentiles [128].

Since different strategies for managing label noise (see subsection 3.6.1) affect the number of SHG images available for analysis, these splits were applied separately for each label correction approach. The number of SHG images used for feature extraction in each approach and their corresponding splits are summarized in *Table 3-6*.

Table 3-6 Distribution of SHG images between training and validation datasets for each approach. Adapted from [***Paper D***].

Dataset	Dataset volume	Number of classes	Split set	Class name	Number of images per class (in corresponding train/validation split)		
					70/30	80/20	90/10
I. Tissue-related	19403	2	train	PTC	7947	9096	10251
				FTC	5635	6426	7211
				In total:	13582	15522	17463
			validation	PTC	3433	2284	1129
				FTC	2388	1597	812
				In total:	5821	3881	1941
II. Capsular-related	7695	2	train	PTC	3240	3695	4135
				FTC	2146	2461	2790
				In total:	5386	6156	6926
			validation	PTC	1351	896	456
				FTC	958	643	314
				In total:	2309	1539	771
III. Capsular-related, preliminary separated	13777	3	train	PTC	4521	5197	5848
				FTC	2456	2789	3131
				Non-capsular	2666	3035	3420
				In total:	9643	11021	12399
			validation	PTC	1953	1277	626
				FTC	1034	701	359
				Non-capsular	1147	778	393
				In total:	4134	2756	1378

3.6.4 ML Classifiers

The ML classifiers used in the study are listed in *Table 3-7*.

Table 3-7 ML classifiers. Reprinted from [***Paper D***]

Type	ML model	Abbreviation	Reference
Ensemble	Random Forest	RF	[114]
	Extreme Gradient Boosting	XGBoost	[115]
	Light Gradient-Boosting Machine	LightGBM	[116]
Monolithic	Logistic Regression	LR	[163]
	C-Support Vector Classifier	C-SVC	[117]
	Multilayer Perceptron	MLP	[118]

All models have been developed in the Python platform libraries scikit-learn 1.6.1 [151]. Depending on the task, all data were labelled either "PTC/FTC" or "PTC/FTC/Non-target" and the developed models aimed at either binary or multi-class classification, respectively.

3.6.5 Hyperparameter tuning

The hyperparameters are top-level parameters of the ML classifier that control the model development process and must be optimized before training the best/final model [164]. Since the default hyperparameter values for a given dataset are rarely optimal, a systematic search was conducted to determine the best configurations for each algorithm. The tuning process involved determining appropriate ranges for the most influential hyperparameters of

each model, followed by an exhaustive or grid-based search to evaluate their impact on performance. The specific hyperparameters tuned for each algorithm are described in detail below:

- **RF**: Three hyperparameters make a key contribution to model performance: the number of trees in the forest (n_estimators), The maximum depth of the tree (max_depth) and The number of features to consider (max_features) [165,166]. To find the best RF model, the following parameter grid was defined: the number of trees varied from 100 to 600 with a step of 50, the maximum depth of the tree varied from 10 to 100 with a step of 10 and the number of features varied from 2 to 34 with a step of 2.
- **XGBoost**: Four crucial hyperparameters were selected for initialization and evaluation based on their effectiveness [167,168]: The number of gradient boosted trees (n_estimators) varied from 100 to 600 with a step of 50. The step size shrinkage (learning_rate), which used in update to prevents overfitting varied from 0.1 to 0.9 with a step of 0.1. The minimum loss reduction (min_split_loss) varied from 0 to 10 with a step of 1 to obtain better control of algorithm. Maximum depth of a tree (max_depth) varied from 0 to 5 with a step of 1. When 0 means no limit on depth. Increasing this value makes the model more complex.
- **LightGBM**: Since LightGBM uses the leaf-wise tree growth algorithm, the three most important hyperparameters have been optimized within the following predefined ranges: The number of boosting trees to be set (n_estimators) was varied from 100 to 600 in step 50; the maximum number of tree leaves for the base learners (num_leaves) was varied from 70 to 180 in step 10; the boosting learning rate was varied from 0.05 to 0.25 in step 0.05.
- **LR**: In order to use all four regularization terms $r(w)$ via the penalty argument available in Scikit-learn ($None, l_1, l_2, ElasticNet$), the regularized logistic regression was solved using the SAGA algorithm [169]. The inverse of the regularization strength (C) and the tolerance for the stopping criterion (tol) ranged from 0.0001 to 10 on a logarithmic scale.
- **C-SVC**: As C-SVC highly depends on the kernel [170] four kernels ('rbf', 'linear', 'poly', 'sigmoid') were checked. The strength of the regularization, which is inversely proportional to Regularization parameter (C) ranged from 0.0001 to 1000 on a logarithmic scale. Tolerance for stopping criterion (tol) ranged from 0.0001 to 1 on a logarithmic scale.
- **MLP**: The number of hidden layers and the neurons they contain is the most important hyperparameter for MLP, as it determines the architecture of the neural network. Many hidden layers lead to overtraining of the model and more than two hidden layers are not necessary [171]. Therefore, MLPs with one and two hidden layers were considered. The number of

neurons in each layer varied from 10 to 100 in steps of 10. Four activation functions were considered for the hidden layer: Identity ($f(x) = x$), logistic ($f(x) = \frac{1}{1+exp(-x)}$), tanh ($f(x) = tanh(x)$) and ReLu ($f(x) = Max(0, x)$). The Adaptive Moment Estimation algorithm was used for weight optimization. The parameter for the penalty (regularization term) (alpha) was varied from 0.001 to 5. This range was divided into 20 values on a logarithmic scale (a geometric progression).

The hyperparameters of each ML model, tuned and used in this study, are detailed in *Table 3-8*. Hyperparameters not listed in *Table 3-8* were set to their default values.

Hyperparameter tuning was performed using either grid search or halving grid search. Grid search considers all possible combinations of hyperparameters and was used for models with a small number of hyperparameters (LR, C-SVC and LightGBM). Halving grid search, based on the successive halving algorithm [172], was used for models with larger hyperparameter combinations (RF, XGBoost and MLP) [173].

Since the imbalance of the PTC/FTC ratio could not be excluded after separation of the tissue-related data, a stratified 10-fold cross-validation [152] was performed for both tuning algorithms. The input dataset was equally divided into 10 stratified subsets, with 9 subsets used for training and the remaining subset used for validation. Each subset preserved the original class distribution. The training/validation process was repeated 10 times, with the validation subset changing each time [174]. Stratified 10-fold cross-validation can handle multi-class problems, so it was also applied in the "PTC/FTC/Normal tissue" classification [151], ensuring carcinoma-specific cluster ratios were preserved in both training and validation sets. The hyperparameter values for all optimized classifiers are summarized in Table S3 (Supplementary Material 1 of ***Paper D***).

The ML models with the highest accuracy were considered optimized and were retrained using the entire training dataset. The schematic workflow for model optimization, training, and testing is shown in *Figure 3.5b*.

Table 3-8 The ranged hyperparameters for tuning of the classifiers. Reprinted from [***Paper D***].

Model	The most important hyperparameters	Ranges of hyperparameters
LR	The inverse of the regularization strength	"C": [0.0001, 0.001, 0.01, 0.1, 1, 10];
	Penalty argument	"penalty": ['l1', 'l2', 'elasticnet', None];
	The tolerance for the stopping criterion	"tol": [0.0001, 0.001, 0.01, 0.1, 1, 10].
C-SVC	Regularization parameter	"C": [0.0001, 0.001, 0.1, 1, 10, 100, 1000];
	Kernel	"kernel": ['rbf', 'linear', 'poly', 'sigmoid'];
	The tolerance for stopping criterion	"tol": [0.0001, 0.001, 0.01, 0.1, 1].
MLP	The number of hidden layers and Number of neurons per each hidden layer	"hidden_layer_sizes": [(10), (20), (30), (40), (50), (60), (70), (80), (90), (100), (10, 10), (10, 20), (10, 30), (10, 40), (10, 50), (10, 60), (10, 70), (10, 80), (10, 90), (10, 100), (20, 10), (20, 20), (20, 30), (20, 40), (20, 50), (20, 60), (20, 70), (20, 80), (20, 90), (20, 100), (30, 10), (30, 20), (30, 30), (30, 40), (30, 50), (30, 60), (30, 70), (30, 80), (30, 90), (30, 100), (40, 10), (40, 20), (40, 30), (40, 40), (40, 50), (40, 60), (40, 70), (40, 80), (40, 90), (40, 100), (50, 10), (50, 20), (50, 30), (50, 40), (50, 50), (50, 60), (50, 70), (50, 80), (50, 90), (50, 100), (60, 10), (60, 20), (60, 30), (60, 40), (60, 50), (60, 60), (60, 70), (60, 80), (60, 90), (60, 100), (70, 10), (70, 20), (70, 30), (70, 40), (70, 50), (70, 60), (70, 70), (70, 80), (70, 90), (70, 100), (80, 10), (80, 20), (80, 30), (80, 40), (80, 50), (80, 60), (80, 70), (80, 80), (80, 90), (80, 100), (90, 10), (90, 20), (90, 30), (90, 40), (90, 50), (90, 60), (90, 70), (90, 80), (90, 90), (90, 100), (100, 10), (100, 20), (100, 30), (100, 40), (100, 50), (100, 60), (100, 70), (100, 80), (100, 90), (100, 100)];
	Regularization term	"alpha": [0.0001, 0.00018, 0.00031, 0.00055, 0.00097, 0.00172, 0.00305, 0.00538, 0.00951, 0.01682, 0.02973, 0.05253, 0.09285, 0.16409, 0.29000, 0.51252, 0.90579, 1.60082, 2.82915, 5];
	Activation functions	"activation": ['identity', 'logistic', 'tanh', 'relu'].
RF	Number of trees in the forest	"n_estimators": [100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600];
	Maximum depth of a tree	"max_depth": [10, 20, 30, 40, 50, 60, 70, 80, 90, 100];
	Number of features to consider	"max_features": [2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34].
XGBoost	Step size shrinkage	'learning_rate': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9];
	Minimum loss reduction	'min_split_loss': [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10];
	Maximum depth of a tree	'max_depth': [0, 1, 2, 3, 4, 5];
	Number of gradient boosted trees	'n_estimators': [100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600]
LightGBM	Number of boosting trees	'n_estimators': [50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600];
	Boosting learning rate	'learning_rate': [0.05, 0.1 , 0.15, 0.2 , 0.25];
	Maximum number of tree leaves for the base learners	'num_leaves':[70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170].

3.6.6 Classifier performance evaluation

The performance of the developed models was estimated using the confusion matrix.

In binary classification, samples labelled as PTC were treated as the positive class, and samples labelled as FTC as the negative class. In this context, false positives (FP) and false negatives (FN) refer to incorrect predictions, while true positives (TP) and true negatives (TN) refer to correct predictions for PTC and FTC, respectively. The model performance was quantitatively compared using accuracy, precision, recall, and F1-score, which were calculated from the confusion matrix elements according to [175].

The accuracy, precision, recall, and F1-score metrics were calculated using fixed thresholds, set to 0.5 for the predicted class probabilities in the current study. To evaluate the performance of the developed binary classification models across a range of thresholds for sensitivity and specificity, Receiver Operating Characteristic (ROC) analysis was performed [176]. The performance of the classifier was represented by the area under the ROC curve (AUC) values [175].

In multiclass classification, evaluation metrics included accuracy, precision "macro" (unweighted mean of precision for each label), recall "micro" (global recall calculated by counting the total TP, FN and FP), F1 "weighted" (average weighted F1 score for each label by the number of true instances for each label) [151].

3.6.7 Feature importance analysis and interpretation

To identify the features which contribute most to the classifiers' decision-making, a feature importance analysis was performed. Since the texture features of SHG images are expected to be correlated [177], there is no single universal approach for estimating the most important features. For instance, multicollinearity may affect the results of permutation importance analysis (PIA) if applied to LR model or tree-based classifiers. Thus, feature contribution to decision-making was treated independently using model-specific approaches, as detailed in *Table 3-9*.

Table 3-9 Algorithms for feature importance analysis. Adapted from [***Paper D***].

Model	Function	Feature importance analysis method
LR	built-in	For binary classification – module of the coefficients of the features in the decision function. For multi-class – mean of modules of the coefficients of the features in the decision function
RF	built-in	The impurity-based feature importances (also known as the Gini importance)
XGBoost	built-in	Feature importance - the number of times a feature is used to split the data across all trees (for multi-class the feature importance is "averaged" over all targets)
LightGBM	built-in	Feature importances contain numbers of times the feature is used in a model
MLP	external algorithm	Permutation feature importance
C-SVC	external algorithm	Permutation feature importance

3.7 Computation

The calculations were performed in Python v3.9 with an Intel i7-13700KF CPU with 16 cores and 24 threads; 32 GB random-access memory; Nvidia GeForce RTX 3060 Ti graphics card with 4864 cores.

3.8 Ethical statement

The animal experiments were conducted in accordance with the principles of bioethics. The study adhered to the requirements of the European Convention for the Protection of Vertebrate Animals used for Experimental and other Scientific Purposes and to legal, scientific and methodological guidelines and reference materials for the husbandry, feeding, removal from the experiment and subsequent disposal of the animals. The permit was issued by the Ethics Committee of the Belarusian Academy of Postgraduate Education (approval number 4 dated 23.09.2020).

The use of the thyroid tissue samples for research purposes was approved by the Carol Davila University Central Emergency Military Hospital, Bucharest, Romania (protocol number 380/09.06.2020). Written informed consent was obtained from patients, and all samples were anonymized before analysis. All experiments were performed according to the relevant guidelines and regulations and in accordance with the Declaration of Helsinki.

4 RESULTS AND DISCUSSION

4.1 Statistical intensity and texture-based analysis of fibrosis progression in PAH in rats

This section is dedicated to the quantitative and qualitative analysis of the development of fibrosis accompanying the progression of MCT-induced PAH in rats based on the wide-field SHG images of lung tissue sections. The results, presented in this section, were published in ***Paper A*** and presented in ***Conferences 1*** and ***2***.

The development of sensitive, label-free and fast imaging techniques with high resolution and robust statistical image analysis is of paramount importance in early diagnosis of PAH progression. Wide-field SHG microscopy of lung tissue in combination with quantitative analysis of SHG images may significantly simplify the detection of fibrosis progression during PAH, allow objective assessment of the stage of the pathology and overcome an error-prone human judgment as in the case of histological analysis.

To demonstrate this, SHG images of rat lung tissue samples were scanned as described in subsections 3.2.1, 3.2.2, 3.2.3 and then analyzed over the manually selected ROIs. ROI images were quantitatively characterized using FFT and texture analysis, where texture analysis consisted of FOS and SOS *Figure 4.1*.

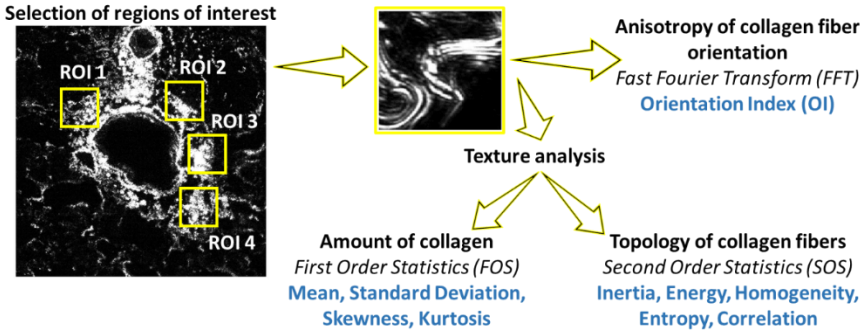


Figure 4.1 Procedures of SHG image analysis. Reprinted from [***Paper A***]

Additionally, SHG image analysis data were compared to the results of IHC. The distributions of the parameters obtained from IHC, FFT, SOS and FOS are further represented by Beeswarm boxplots. The 75th and 25th percentiles are labeled as the top and bottom of each rectangular box, respectively. The median is shown inside the box. The whiskers are shown as 1.5 times the IQR below and above the box.

4.1.1 IHC Analysis of rat lung tissue

Type I and III fibrillar collagen accounts for over 90% of collagens in the lung parenchyma [178]. The remodeling of ECM during fibrosis is

characterized by fluctuations in its content [179], and the increased collagen levels may indicate the stage of progression of the pathology [180]. Metalloproteinases (MPP) are responsible for collagen degradation [181], and the imbalance of MPP levels may contribute to fibrosis development in PAH [182]. TIMP-1 is an inhibitor of all types of MPP, which are responsible for the degradation of most fibrillar collagens. An elevated circulating level of TIMP-1 therefore indicates a disruption of collagen degradation and the associated accumulation.

Quantitative analysis of IHC makes it possible to show the changes in the expression levels of collagen I and III as well as TIMP-1, which are molecular markers of fibrosis development associated with the progression of PAH. The corresponding IEs are shown in *Figure 4.2*; the statistical significances for the comparison of all experimental groups are summarized in Table S2 (Supplementary Material of ***Paper A***).

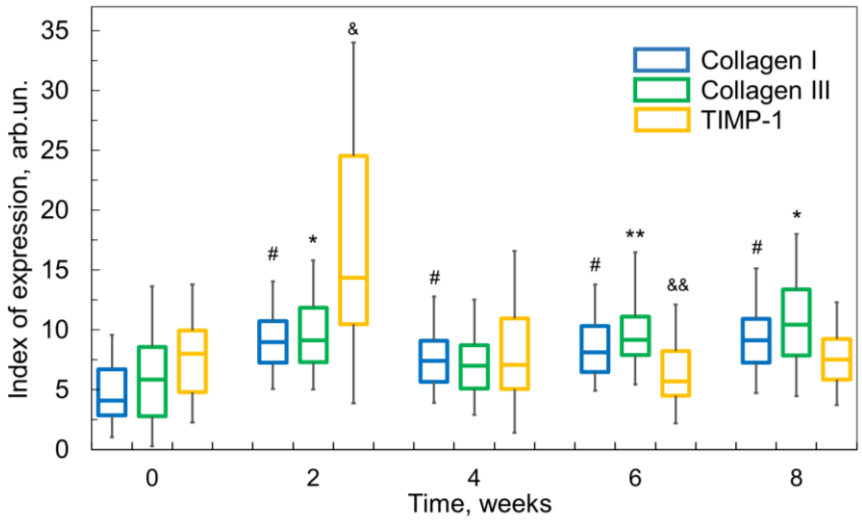


Figure 4.2 Expression indices (IEs) of collagen I, collagen III and TIMP-1 during PAH progression. Statistical difference: # $p < 0.001$ for collagen I; * $p < 0.001$ and ** $p < 0.01$ for collagen III; & $p < 0.001$ and && $p < 0.01$ for TIMP-1 compared to the corresponding control groups. Reprinted from [***Paper A***]

The expression of collagen I in the lung tissue of rats increased significantly in all experimental groups (2, 4, 6 and 8 weeks of PAH progression) compared to the control group. The expression of collagen III is periodic. A significant increase is observed after 2 weeks of PAH progression ($p < 0.001$). Thereafter, a normalization of collagen III levels to control levels is recorded, with a further slight increase after 6 weeks of PAH progression (*Figure 4.2*).

The expression of TIMP-1 is time-dependent with a 2-fold increase after 2 weeks of PAH progression in rats compared to the control group of healthy animals. However, in contrast to the increase in collagen I and III levels in the later stages of PAH, TIMP-1 expression is downregulated to control levels throughout the remaining observation period.

In addition, fibrogenesis begins after 2 weeks of MCT-induced PAH (*Figure 4.2*). The maximum increase in TIMP-1 expression in the 2nd week of the experiment ($p < 0.001$ vs. all groups) promotes fibroblast synthesis activity and a change in the balance between collagen I and III synthesis. Further progression of PAH (6-8 weeks) is characterized by a decrease in MPP inhibitor expression ($p < 0.001$), an increase in type III collagen synthesis from 7.0 to 10.41 ($p < 0.001$) from the 4th to the 8th week of the experiment ($p < 0.001$). As a result of ECM remodeling, a favorable environment for the initiation of fibrogenesis in the vessel wall is created in the local microenvironment of the pulmonary arteries.

4.1.2 SHG/TPEF Imaging of rat lung tissue

The combination of SHG and TPEF imaging allows visualization of both the fibrotic structures and the outer tissue, providing a complete picture of the branching and enlargement of collagen fibers in the tissue [183]. Endogenous TPEF signal often originates from metabolic compounds present in tissue sections, such as nicotinamide adenine dinucleotides, flavins, tryptophan and tyrosine in proteins, serotonin, phycoerythrin, etc. [184,185]. It is important to note that the H&E staining used for brightfield microscopy also contributes to the TPEF signal, as eosin is fluorescent and this has already been demonstrated for TPEF visualization of tissues [186]. The images of H&E-stained lung tissue sections are shown in *Figure 4.3*, where SHG of collagen is colored yellow and TPEF is colored red. In the control sample, it can be seen that collagen surrounds the wall of the blood vessels (marked with arrows in *Figure 4.3*) but is not present in the lung tissue. The H&E images of the PAH samples show only a thickening of the blood vessel walls compared to the control sample, while the SHG/ TPEF images show a clear collagen overproduction. *Figure 4.3* shows that collagen content increases over time and that highly assembled long fibers form dense networks both around the blood vessels and in the surrounding tissue, extending deep into the alveolar region (marked with arrows). This confirms our previous findings [187] and the IHC data indicating a significant increase in collagen expression after 4-8 weeks of PAH progression.

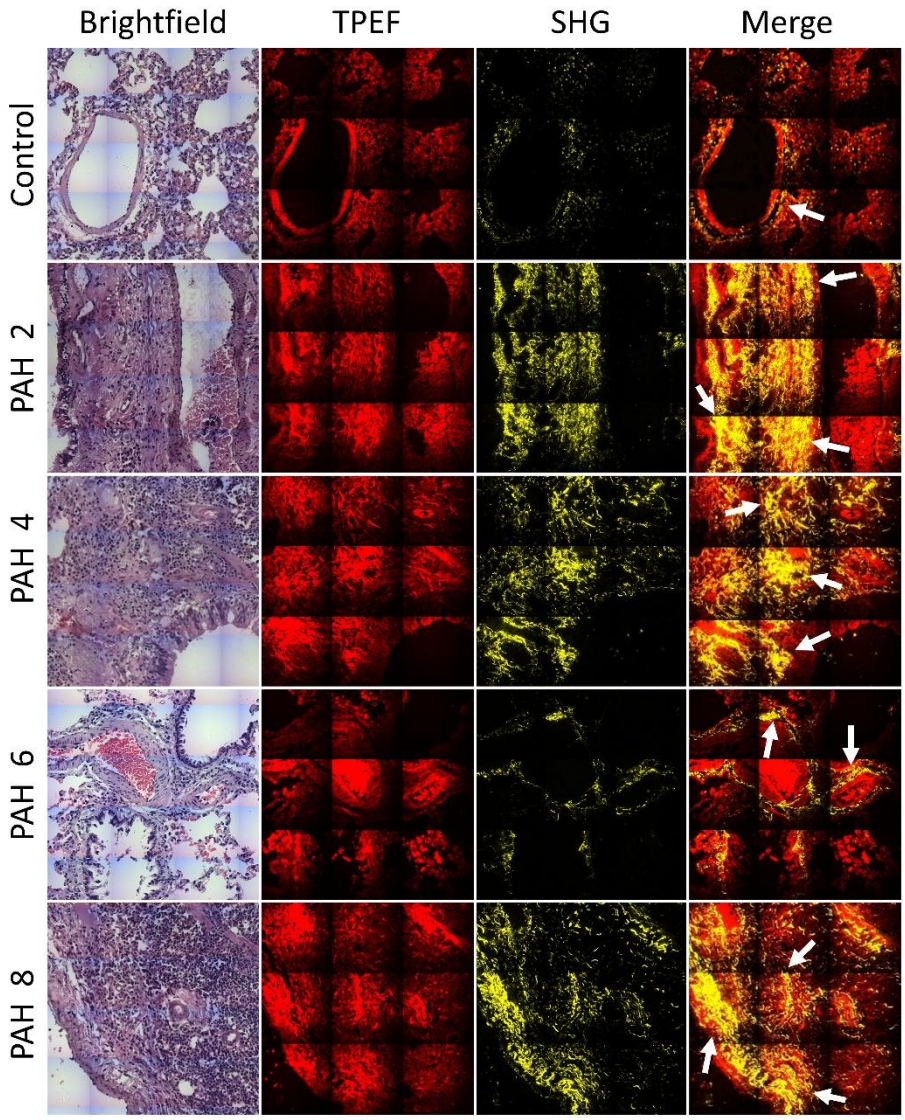


Figure 4.3 Brightfield images, TPEF images, SHG images and combined TPEF and SHG images of H&E-stained rats lung tissue of control group and of rats on the 2nd, 4th, 6th, and 8th week of PAH progression. Image size is 450 µm × 450 µm. Adapted from [***Paper A***]

4.1.3 Anisotropy of collagen fiber orientation of rat lung tissue

Information about the anisotropy of collagen fiber orientation is extracted from SHG images using FFT. The temporal dependence of the *OI* is shown in *Figure 4.4*. The lower the *OI*, the worse the arrangement of the collagen fibers. An isotropic collagen structure is characterized by *OI* = 0 and a circular FFT image [188]. In contrast, the higher the *OI*, the more pronounced the anisotropy of collagen fiber orientation. The lung tissue of healthy rats was

found to have a lower *OI* than the tissue during PAH progression. This suggests that PAH is associated with collagen fiber arrangement. In the second week of PAH-associated fibrosis development, *OI* increases significantly, indicating elongation of fibers and their expansion in lung tissue. The subsequent decrease in *OI* indicates a reorganization of the collagen fibers and the formation of a more isotropic collagen network.

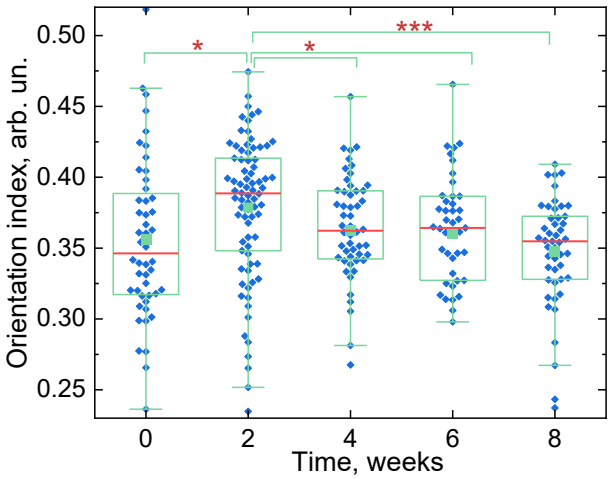


Figure 4.4 OI values calculated on SHG images of control samples (0 weeks) and during PAH progression. The statistical significance of the difference between the different distributions is p<0.1*, and p<0.001***. Adapted from [***Paper A***].

4.1.4 Intensity and texture analysis of SHG images of rat lung tissue

The distributions of μ_1 and σ are shown in *Figure 4.5a-b*. The increase in both parameters in the early phase of disease progression (2 weeks) indicates an increase in collagen production and an increasing heterogeneity of collagen distribution in the tissue. Both parameters decrease in the 4th and 6th week and then increase sharply in the final stage of PAH progression in the 8th week. The decrease in collagen content after week 2 may indicate that the organism is struggling to dampen the effects of rapid PAH progression and restore the balance between collagen synthesis and degradation. This hypothesis is supported by the IHC data and confirms the evolution of the levels of collagen I, collagen III and TIMP-1.

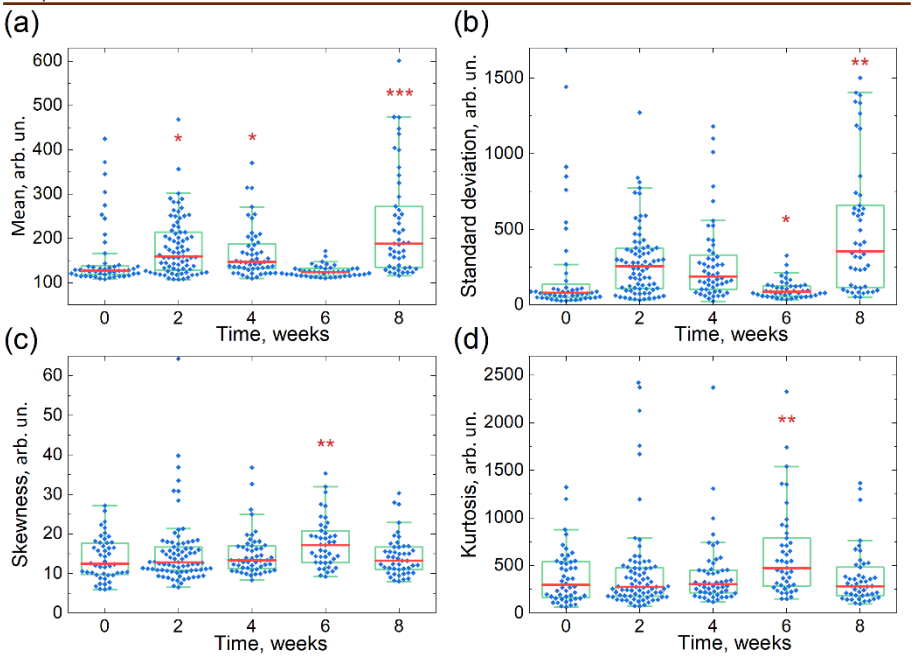


Figure 4.5 FOS parameters determined inform SHG images of control samples and during the PAH progression: (a) – Mean, (b) – Standard Deviation, (c) – Skewness, (d) – Kurtosis. The statistical significance of the difference of different distributions relative to the control is $p < 0.1$ *, $p < 0.01$ **, $p < 0.001$ ***. Adapted from [***Paper A***].

The values for g_1 and g_2 remained constant throughout the experiment and were consistent with the control group, with the exception of the 6th week of PAH progression, when an increase in these parameters was observed (see *Figure 4.5c-d*).

The SOS parameters are extracted from the GLCM and provide information on collagen distribution, spatial fiber organization, uniformity, etc. All calculated SOS parameters are shown in *Figure 4.6*.

The value of *I* increases at an early stage of PAH progression (week 2), indicating an increase in the number of areas of high contrast compared to the control. This is probably related to the formation of high-density collagen bundles and the appearance of foci of collagen synthesis. A subsequent decrease in contrast indicates a homogenization of collagen distribution in the 4th and 6th week of PAH progression. However, the final phase of PAH is accompanied by a significant increase in the number of high-contrast areas and is due to the thickening of the collagen bundles. A low *E* in the 2nd and 8th week thus indicates the order and spread of the collagen fibers. The calculated *C* is in the range of 0.9-1.0, indicating the presence of some periodic structures, both in the healthy rats of the control group and in the rats with PAH at different stages of progression. Examples of such periodic structures are shown in *Figure 4.6f*.

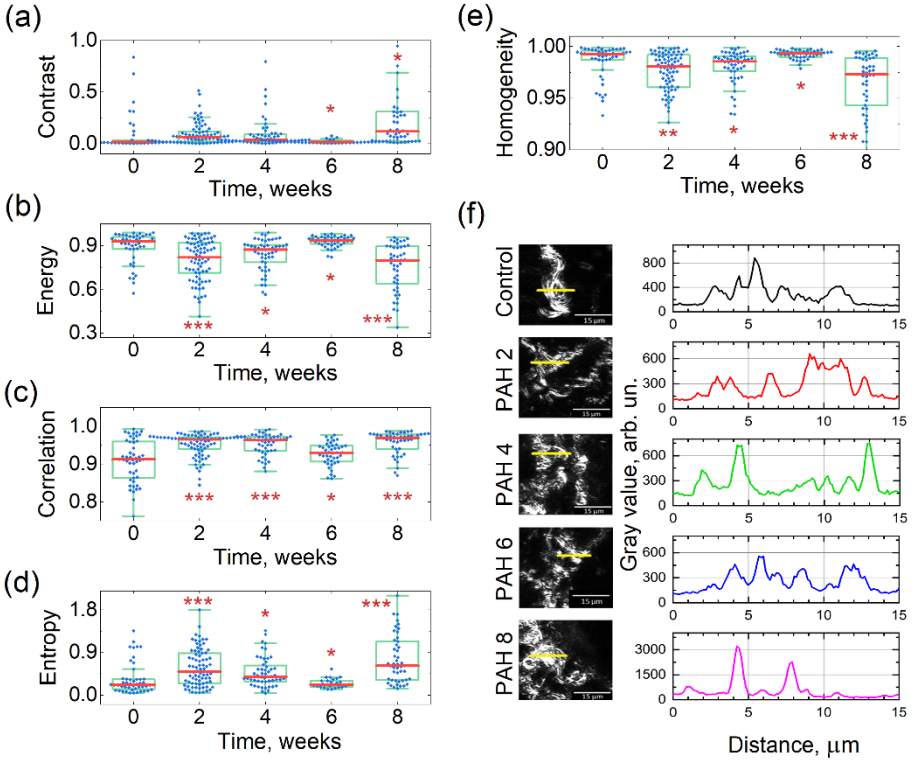


Figure 4.6 SOS parameters calculated on images of control samples of healthy lung tissue and at different stages of PAH progression: (a) – inertia, (b) – energy, (c) – correlation, (d) – homogeneity, (e) – entropy, (f) – typical SHG collagen structures (scale bar 15 µm). The statistical significance of the difference of distributions relative to the control is $p < 0.1$ *, $p < 0.01$ **, $p < 0.001$ ***). ROI size is 150 µm × 150 µm. Adapted from [***Paper A***].

These data suggest that despite the changes in fiber orientation anisotropy (*Figure 4.4*), collagen packing in fibers characterized by periodicity is not significantly disrupted during PAH progression, as indicated by the small variations in correlation values. Nevertheless, the latter are significantly distinguishable *Figure 4.6c*. The development of the *H* values *Figure 4.6d* correlates with the data discussed above. The low *H* of the control samples and lung tissue at week 6 of PAH-associated fibrosis development suggests that the fibers in the selected area are less clearly separated compared to other stages of disease progression. This could be caused, for example, by fiber swelling during inflammation in the 6th week of PAH progression [189]. High *H* values in the 2nd, 4th and 8th week of PAH are caused by bright and distinct, but not necessarily ordered, collagen fibers that stand out from a homogeneous background. A decrease in *L* values during PAH progression *Figure 4.6e* indicates the expansion of the network of thin and disorganized collagen fibers and the development of fibrosis. These data agree well with the *E* values and

confirm the growth of collagen fibers in the lung tissue both near and far from the blood vessel walls.

The significance of the differences between all statistical parameters of the ROIs of the experimental groups of animals with PAH and the control group of healthy animals is shown in *Table 4-1*.

Table 4-1 The significance difference between statistical parameters of ROIs of experimental groups of animals with PAH and control group of healthy animals, performed by one-way ANOVA applying an unpaired two-tailed Student's T-test. Adapted from [***Paper A***]

	μ_1	σ	g_1	g_2	*I*	*C*	*E*	*L*	*H*	*OI*
Con. - PAH2	<0,1	0,551	0,285	0,308	0,717	<0,001	<0,001	<0,01	<0,001	<0,1
Con. - PAH4	0,366	0,669	0,302	0,691	0,615	<0,001	<0,1	<0,1	<0,1	0,510
Con. – PAH6	<0,01	<0,1	<0,01	<0,01	<0,1	<0,1	<0,1	<0,1	<0,1	0,687
Con. – PAH8	<0,001	<0,01	0,566	0,926	<0,1	<0,001	<0,001	<0,001	<0,001	0,418
PAH 2 – PAH 4	0,245	0,912	0,792	0,382	0,704	0,869	<0,1	0,130	0,127	<0,1
PAH 2 – PAH 6	<0,001	<0,001	<0,1	0,627	<0,001	<0,001	<0,001	<0,001	<0,001	<0,1
PAH 2 – PAH 8	<0,01	<0,001	0,532	0,297	<0,001	0,260	<0,1	<0,1	<0,1	<0,001
PAH 4 – PAH 6	<0,001	<0,001	<0,1	<0,1	<0,01	<0,001	<0,001	<0,001	<0,001	0,810
PAH 4 – PAH 8	<0,01	<0,01	0,634	0,637	<0,01	0,259	<0,01	<0,001	<0,01	<0,1
PAH 6 – PAH 8	<0,001	<0,001	<0,01	<0,01	<0,001	<0,001	<0,001	<0,001	<0,001	0,162

Physiologically, the observed collagen changes during the 5-week period of PAH progression are associated with, among other things, a gradual accumulation of inflammatory markers in lung tissue, including monocyte chemoattractant protein 1 (MCP-1) and interleukin-6 (IL-6), as shown in [190]. Treatment with silibinin (C-X-C chemokine receptor type 4 inhibitor) reduces PAH symptoms in the first two weeks due to downregulation of gene expression of inflammatory markers in the pulmonary arteries but not in the lung tissue. In later stages of PAH, however, silibinin no longer succeeds in relieving symptoms. At this point, it becomes an irreversible condition as MCP-1 and IL-6 accumulate in the tissue. Elevated MCP-1 induces collagen synthesis by lung fibroblasts and also recruits monocytes to the site of inflammation [191]. Elevated IL-6 induces the production of collagen I by dermal fibroblasts [192] and contributes to the deposition of ECM [193].

Based on the SHG image analysis results and considering the physiological background and the results of IHC analysis, the following stages of development/progression of PAH can be proposed:

I. The pathology develops rapidly at the beginning: in the 2nd week, the μ_1, σ, *OI*, *I* and *H* increase, but the *E* and *L* decrease, indicating the initiation of collagen accumulation, stretching of collagen fibers and their spread in the lung tissue.

II. The organism tries to regulate the progression of the disease: In the 4th week, the decrease in μ_1 and σ as well as the high H values indicate a possible inflammation-related disruption of the collagen structure as well as the activation of the collagen degradation system, which is supposed to correct the imbalance between collagen synthesis and degradation.

III. The phantom recovery of the organism at week 6, as the μ_1 and σ decrease and H is low; however, an increase in g_1 and g_2 indicate a significant redistribution of collagen, suggesting a thickening of collagen fibers and deep penetration into lung tissue, and thus could represent a "point-of-no-return" in PAH pathogenesis;

IV. Finally, total tissue failure occurs at the 8th week of pathology, which is the result of a significant increase in collagen content and thickening of collagen bundles, characterized by the increase in μ_1, σ, I, H and decrease in E and L.

Overall, SHG imaging provides a complete picture of morphologic collagen changes during PAH-associated fibrosis progression. The evolution of the different FOS and SOS parameters indicates the same characteristic changes in collagen fiber structure and network organization that are also consistent with the results of IHC. This is beyond the scope of the present work, but with the collection of further data from a larger number of samples, SHG image analysis could also provide reliable signatures for the different stages of PAH pathogenesis.

This was summarized in the first statement of the thesis: *Statistical analysis of wide-field SHG images of lung tissue sections reveals and qualitatively and quantitatively describes characteristic changes in collagen organization, morphology and collagen content associated with the different stages of pulmonary arterial hypertension.*

4.2 ML-based analysis of collagen ultrastructure in PTC capsular invasion based on wide-field PSHG microscopy

This section describes the application of PSHG for studying changes in collagen ultrastructure in the areas of capsular invasion and intact collagen capsule around PTC utilizing single-axis molecule model of collagen fibers and demonstrates the advantages of using unsupervised ML algorithms for PSHG image analysis. The results from this section were published in ***Paper B*** and presented at ***Conference 6***.

4.2.1 Wide-field PSHG imaging of thyroid nodule section

To illustrate the ability of PSHG wide-angle imaging to quantitatively assess capsular invasion, a PTC node was selected in which differential analysis was performed between invasion-proximal areas and control areas on the same histologic slide. An expert pathologist evaluated the overall appearance *Figure 4.7a* of the histologic slide and diagnosed the nodule as an

apparently encapsulated papillary thyroid microcarcinoma with capsular microinvasions. The nodule shows a trabecular and microfollicular growth pattern, accompanied by nuclei that are enlarged, elongated or angulated (irregular contour of the nuclear membrane). The nuclei are clustered, overlapping, vesicular and have clarified chromatin, commonly referred to as "Orphan Annie's eyes" type nuclei. The nucleoli are displaced at the periphery, and occasionally incisions and pseudo-inclusions are seen, resulting from invaginations of the cytoplasm into the nucleus. The cytoplasm appears amphophilic and moderately rich. In addition, the collagenous nodular capsule appears thickened with extensive subcapsular calcifications (i.e., the irregular clumps of basophilic material present mainly on the left and upper left side of the nodule). The surrounding thyroid parenchyma consists of normal-sized follicles characterized by cuboidal and flat epithelia, fibrous septa, and a minimal diffuse lymphocytic inflammatory infiltrate.

Control and invasion regions were selected for further analysis: two regions were selected as control regions within the nodule capsule (i.e. *Ctrl-1* and *Ctrl-2*), while *Ctrl-1* was selected near the subcapsular calcifications, *Ctrl-2* is located within a normal capsule, away from the calcifications. Two other ROIs with focal tumor infiltration were identified as *Inv-1* and *Inv-2* in *Figure 4.7a*.

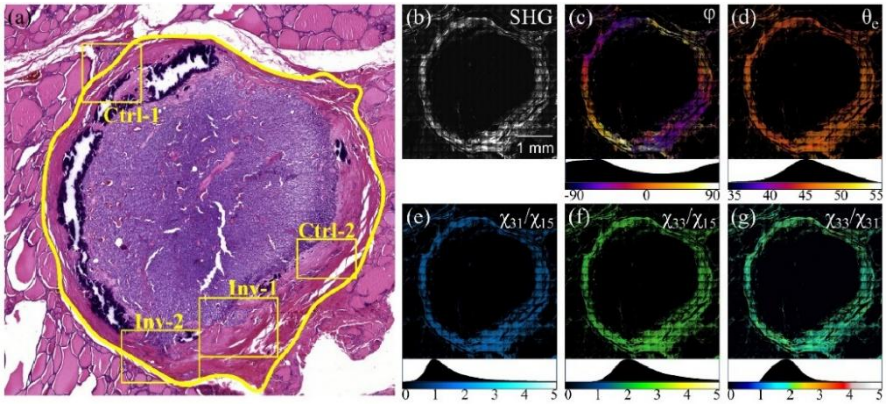


Figure 4.7 Wide-field PSHG imaging of an entire thyroid nodule. (a) H&E-stained tissue image acquired with a whole slide scanner. Annotations represent areas of the nodule capsule away from potential invasion sites (ROI1 and ROI2) and nodule capsule invasion sites (ROI3 and ROI4). (b) corresponding wide-field SHG image generated as the intensity average over the PSHG image stack. (c) the orientation map of collagen around the nodule capsule. (d) helical pitch angle map (θ_e). (e)-(g) $\chi^{(2)}$ elements ratios maps for χ_{31}/χ_{15}, χ_{33}/χ_{15} and χ_{33}/χ_{31}, respectively. Reprinted from [***Paper B***]

The PTC node was imaged with a wide-field PSHG microscope. The individual wide-field PSHG tiles were stitched together to form a complete PSHG image set depicting the entire PTC node *Figure 4.7b*. The application of FF-PSHG analysis to the polarization-resolved image set facilitated the

generation of several maps, including the collagen orientation map *Figure 4.7c*, the helical tilt map *Figure 4.7d*, and maps depicting the ratio of $\chi^{(2)}$ elements *Figure 4.7f-g* for further quantitative analysis.

The results of the FF-PSHG analysis for the ROIs highlighted in *Figure 4.7a*, which correspond to both the control areas and the invasion sites of the nodule capsule, are shown in *Figure 4.8*.

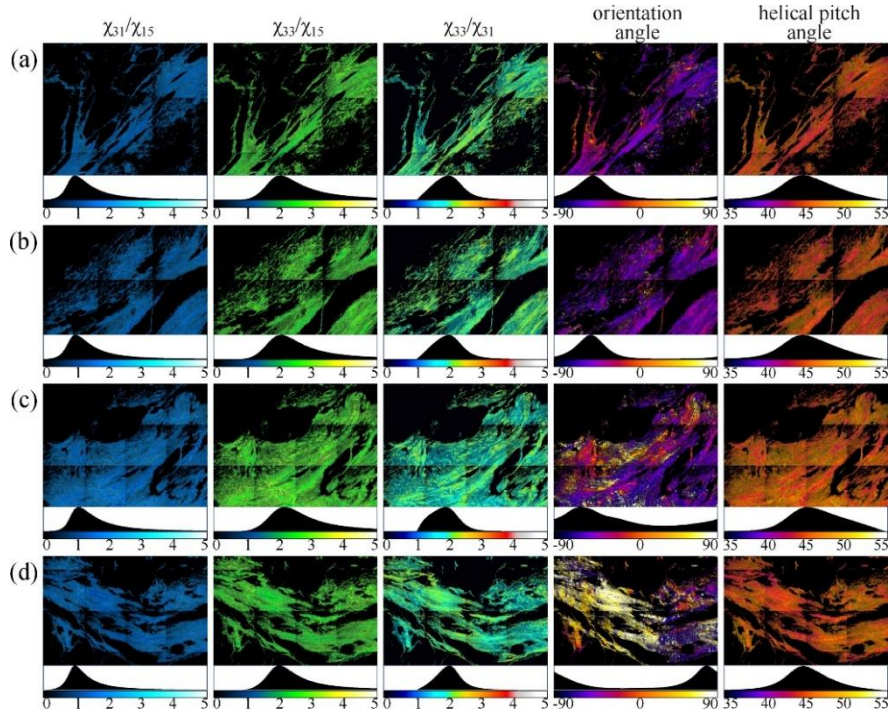


Figure 4.8 Quantitative analysis of *wide-field* PSHG image stacks. The collagen orientation maps, the helical pitch angle and ratios for the $\chi^{(2)}$ elements are displayed for the ROIs of interest: (a) Ctrl-1; (b) Ctrl-2; (c) Inv-1; (d) Inv-2. Reprinted from [***Paper B***]

4.2.2 Unsupervised ML analysis of wide-field PSHG of thyroid nodule section

To further visually improve the identification of potential microinvasion sites, analysis using unsupervised ML techniques was performed. Classification of the data from the maps of all polarization-related parameters, with the exception of φ, resulted in segmentation of the data into $k = 6$ clusters. The total number of downsampled images classified by k-means was 90000. The data from these averaged parameter maps were standardized using the Robust Scaler algorithm. The choice of the number of clusters was made on the basis of two validation metrics SC and BDI (see *Figure 4.9a*).

The separation into two clusters is trivial as it allows a simple separation of collagenous and non-collagenous areas (*Figure 4.9b*), although such binary

k-means segmentation is quite commonly used for the characterization of PSHG images of cancer tissues [194]. The segmentation into several clusters makes it possible to recognize the differences in the collagen structures. The next optimal set of SC and DBI corresponds to $k = 6$ (*Figure 4.9a*): The local DBI minima at $k = 6$ and $k = 9$ are the same, but the SC for $k = 6$ is higher than that for $k = 9$.

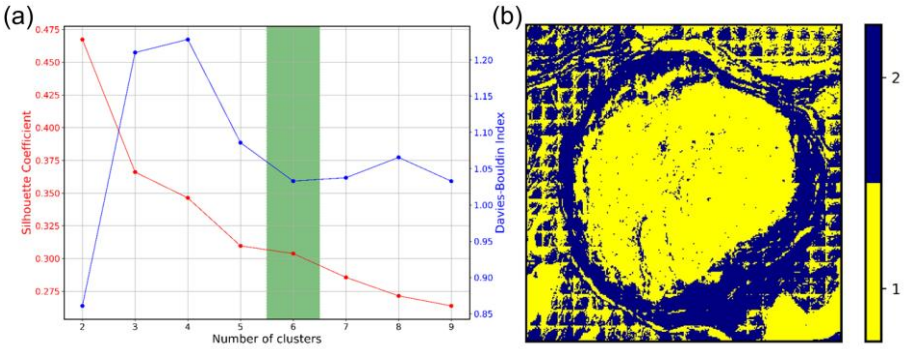


Figure 4.9 (a) The dependence of the evaluation metrics for k-means clustering on the number of clusters k. The green rectangle marks the SC and DBI corresponding to the appropriate number of clusters for image segmentation. (b) The resulting cluster map of the entire nodule for k = 2. Cluster 1 (yellow) corresponds mainly to the non-collagenous area, cluster 2 (navy blue) corresponds to the collagen-containing areas. Adapted from [***Paper B***]

The distribution of the downsampled images between the clusters and their affiliation to collagenous and non-collagenous regions is summarized in *Table 4-2*.

The resulting cluster map for the entire nodule shows the spatial distribution of the characteristic combinations of parameter values representing the different clusters *Figure 4.10a*.

Table 4-2 Distribution of downsampled images in clusters. Adapted from [***Paper B***]

	The number of the cluster	The number of images in the corresponding cluster	The number of images in the corresponding cluster, %
Collagen related clusters	1	20356	22.62
	3	13558	15.06
	5	19867	22.07
Non collagen clusters	2	9362	10.40
	6	26706	29.67
Outliers	4	151	0.17

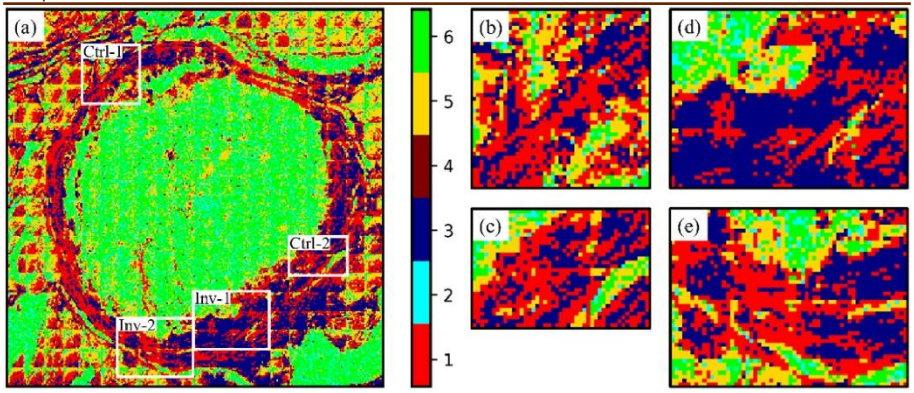


Figure 4.10 Clustering map for data of polarization-related parameters. (a) whole thyroid nodule; (b) Ctrl-1; (c) Ctrl-2; (d) Inv-1; (e) Inv-2. Reprinted from [***Paper B***]

These combinations or centroids obtained by *k*-means *Figure 4.11* can be used to characterize the local properties of the collagen and provide an interpretation of the clustering. Clusters 1, 3 and 5 were assigned to collagenous areas, while 2 and 6 were assigned to non-collagenous areas mainly related to follicles with glandular secretions, cells, etc. Cluster 4 corresponds to the outlier values of PSHG parameters and comprises 0.2% of all data (*Figure 4.11*).

From the visual analysis of the whole capsule, cluster 1 mainly forms the capsule and is also distributed outside the capsule; cluster 3 is mainly located in the inner part of the capsule and is only found in some specific areas; cluster 5 mainly lines the inner and outer sides of the capsule and such collagens probably form numerous septa surrounding groups of follicles. Further examination of cluster formation within the intact capsule in ROIs *Ctrl-1* and *Ctrl-2* *Figure 4.10b,c* and within the microinvasion sites in ROIs *Inv-1* and *Inv-2* *Figure 4.10d,e* reveals distinct quantitative differences in cluster formation (*Table 4-3*). A notable feature of invasion is the increased occurrence of cluster 3 and decreased occurrence of clusters 1 and 5. Consequently, a quick visual inspection of the cluster map facilitates the identification of regions within the nodule capsule that may warrant suspicion of invasion.

Table 4-3 Percentages of collagen-related clusters in intact capsule and at sites of invasion. Adapted from [***Paper B***]

The number of the cluster	Control [%]	Invasion [%]
1	47.1	32.7
3	31.1	50.7
5	21.8	16.6

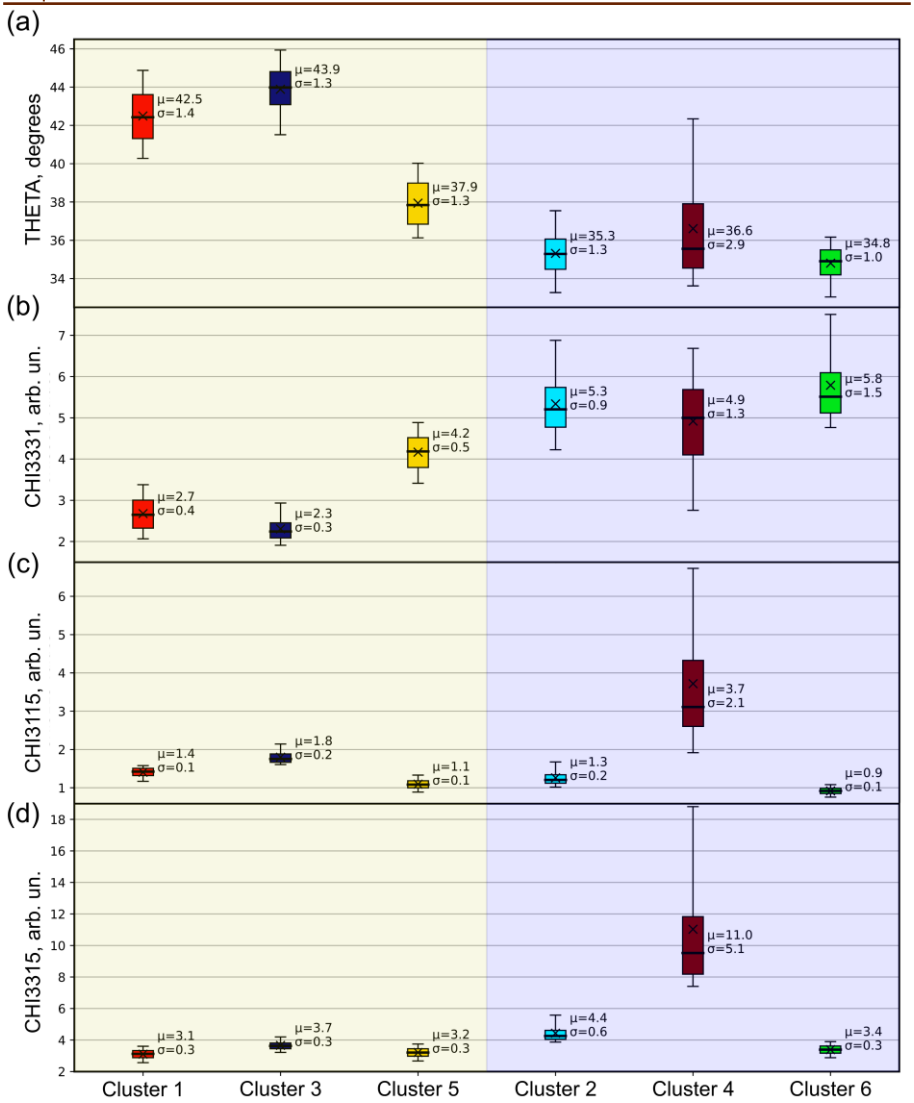


Figure 4.11 The centroids of each cluster for (a) THETA values, (b) CHI3331 values, (c) CHI3115 values, (d) CHI3315 values. Multiplication symbols (×) represent mean values, solid horizontal lines (—) are the median values. Adapted from [***Paper B***].

The mean helical pitch angles of the collagens in clusters 1, 3 and 5 are 42.5°, 43.9° and 37.9°, respectively, indicating severe changes in the molecular structure of the collagen in the invaded capsule compared to the control areas (*Figure 4.11*). Such changes in the triple helix structure may influence the interaction of collagen with normal and cancer cells [195]. Recently, collagens with a tightly packed triple helix (with a helical pitch angle of 43.9° [195] as in cluster 3) were shown to have a higher binding

efficiency to cancer cells compared to normal cells and thus may promote cancer progression and metastasis [196].

A higher angular expansion was observed in the vicinity of suspicious and invasion sites. Conversely, other parameters related to collagen structure studied here show their potential in highlighting collagen changes at the pixel level (i.e., the ratio of $\chi^{(2)}$ elements and the helical pitch angle). The biological significance of susceptibility ratios should be interpreted with caution and depending on several factors, including the theoretical model of collagen, image resolution, and others. If all collagen molecules within the fibrils are aligned in the same direction, the second-order susceptibility should reflect the first-order hyperpolarizability, a tensor that describes the response of the molecule to the applied electric field in terms of second-order nonlinear optical effects. Therefore, the susceptibility ratio should be similar to that of first-order hyperpolarizability. Furthermore, the pitch angle of the collagen molecule can be estimated by relating the ratio of the $\chi^{(2)}$ elements to the orientation angle of the emitting dipole. Care should be taken when interpreting the results and checking whether the original assumptions (e.g. Kleinman symmetry) are fulfilled. In this study, although the χ_{31}/χ_{15} distributions are close to unity, there is considerable variation, suggesting that the absolute results should be taken with a grain of salt. One of the $\chi(2)$-element ratios, namely χ_{33}/χ_{31}, provides insight into the anisotropy of collagen fibrils within the focal volume and is known as an anisotropy parameter [197]. In the literature, χ_{33}/χ_{31}values between 1.2 and 2.6 are reported, depending on the tissue type (lower values for organized collagen in tendons with straight fibrils within the focal plane) and the spatial resolution used [198]. Conversely, collagen showed higher susceptibility values in tissues in which the molecule is oriented at a larger angle within the fibril [199]. Although the interpretation of the obtained susceptibility ratios in a biological context is challenging without the consideration of numerous variables, the differences between the susceptibility ratios may provide more meaningful insights into the changes occurring with pathology under consistent technical conditions.

This was summarized in the second statement of the thesis: *k-Means clustering of cylindrical model parameters extracted from wide-field polarization-resolved SHG images of whole thyroid nodule sections allows differentiation between areas of capsular invasion and unaffected regions of the capsule surrounding cancer cells by revealing patterns in the ultrastructure of collagen.*

4.3 ML-based diagnostics of capsular invasion in thyroid nodules with wide-field SHG microscopy

This section demonstrates the application of the unsupervised ML algorithms for detailed analysis of the wide-field SHG images of collagen

capsules surrounding PTC, which enables the interpretation of changes in collagen structure within the capsule and the detection of areas of microinvasion or areas requiring additional investigation. The results from this section were published in ***Paper C*** and presented at ***Conferences 3***, ***4***, ***5***, ***8*** and ***10***.

4.3.1 Wide-field SHG imaging of thyroid nodule sections

SHG microscopy approach enabled large-scale imaging collagen distribution in whole PTC nodules. An example of an SHG image of an entire nodule is shown in *Figure 4.12a*.

For comparison, a bright-field image is presented in *Figure 4.12b*. PTC can be seen as a lump of cancer cells surrounded by a collagen capsule. Healthy thyroid tissue comprising follicles is outside the capsule. Dark purple formations inside the PTC nodule and the wall of the capsule are calcifications. There are also 2 sites of capsular invasion that were annotated by the pathologist. By comparison, in the SHG image, collagen structures are visible with much better contrast and in more detail. The capsule appears to be rather heterogeneous and of varying thickness. At the same time, almost no SHG signal was detected inside the capsule, since there are almost no collagen structures in that area. Also, at sites of the noted capsular invasion, the integrity of the capsule is clearly compromised by the cancer cell escape pathways. Some collagen structures that are not visible in the bright-field can be seen inside the calcifications. ROIs containing calcifications are shown in *Figure 4.13*.

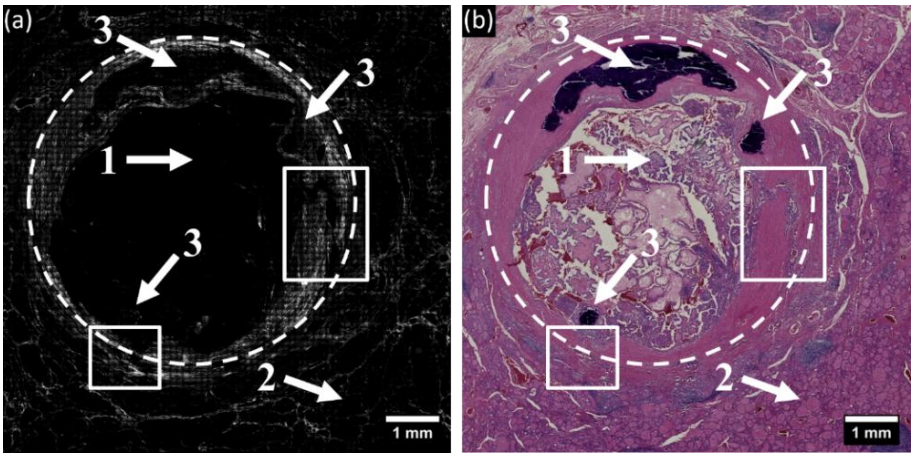


Figure 4.12 Images of an H&E-stained section from an entire nodule of encapsulated PTC. (a) SHG image of collagen distribution tiled from the wide-field SHG images. (b) Bright-field image. The different tissue structures are designated as: 1 – carcinoma cells, 2 – normal thyroid follicles, 3 – calcifications. White boxes indicate invasions annotated by the pathologist. The white dashed circle designates the capsule of the PTC nodule. Reprinted from [***Paper C***].

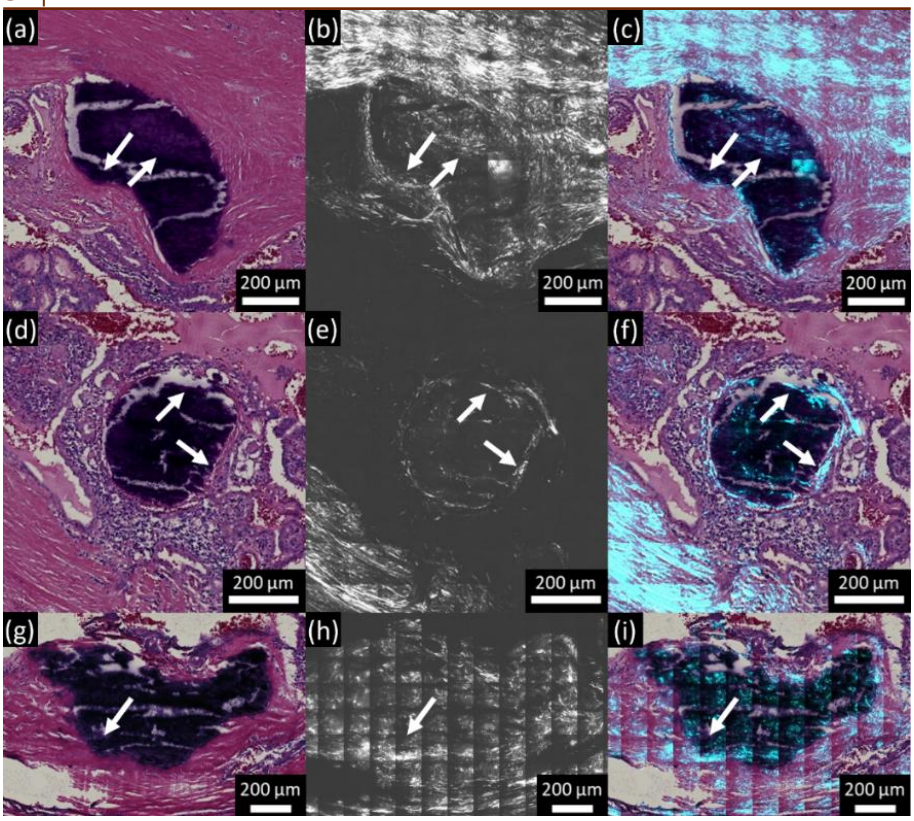


Figure 4.13 Calcifications in the PTC nodule: (a, d, g) – bright-field images of a H&E-stained sections; (b, e, h) – SHG images; (c, f, i) – merged bright-field and SHG images of the same area. Arrows mark the clear boundary between calcification and H&E-stained tissue in the brightfield images and the corresponding areas in the SHG images. Reprinted from [***Paper C***].

Micro- and macrocalcifications are considered a distinct diagnostic feature of PTC [200], although some authors suppose their association with benign conditions [201]. Round or oval calcifications that are concentrically laminated and less than 1 mm in diameter are psammoma bodies, larger calcifications are referred to as macrocalcifications [202]. *Figure 4.13* shows the brightfield image of an H&E-stained tissue section containing calcifications and the corresponding SHG images. Conventional H&E-staining of the tissue limits visual analysis of the calcifications (dark purple, *Figure 4.13a,d,g*). SHG imaging reveals specific collagen structures within the calcifications. Outside the capsule, collagen network of fibrous septa around the follicles is clearly visible. The colloid inside the follicles is seen as patches of uniformly low intensity, which is most probably due to a bleed-through from the TPEF of the Eosin stain.

4.3.2 Intensity and texture parameters of wide-field SHG imaging of thyroid nodule sections

From each recorded SHG image the following intensity and texture parameters were calculated: μ_1, σ, g_1, g_2 (FOS); *E*, *I*, *C*, *L*, *H* (SOS); *SRE*, *LRE*, *GLN*, *RLN*, *RP* (HOS).

Interpretation of each parameter and examples of their assignments to collagen structures are collected in *Table 3-2, Table 3-3* and *Table 3-4*. To address possible long-range order of the collagen structure, the GLCM-related parameters were calculated with a few values of the distance parameter d = 1, 3, 6, 9, and 12 px. Therefore, 34 features were extracted from each individual SHG image, and with the total of 20736 recorded images this resulted in a data set of 34×20736 parameter values. The obtained parameter distributions were, first, standardized using the Robust Scaler algorithm. Then, the parameter correlation matrix (Table S3 in Supplementary Material of ***Paper C***) to identify parameters that could be correlated and therefore redundant were calculated. Here, SOS parameters calculated with different values of d turned out to be strongly correlated.

The local homogeneity *L* with d = 1 and 3 px exhibited strong correlation with virtually all parameters in the set, while *L* with d = 6, 9, and 12 px were moderately correlated with other parameters and highly correlated with each other. Excluding parameters with $Corr(\tilde{X}, \tilde{Y}) \geq 0.9$ left only 13 features left (Table 4-4) to be considered in further analysis.

Table 4-4 Correlation of intensity and texture parameters. Adapted from [***Paper C***].

	$\tilde{\mu}_1$	$\tilde{g}_1$	$\tilde{g}_2$	$\tilde{E}_1$	$\tilde{C}_1$	$\tilde{H}_1$	$\tilde{I}_1$	$\tilde{C}_3$	$\tilde{L}_6$	$\tilde{C}_9$	$\widetilde{\mathrm{LRE}}$	$\widetilde{\mathrm{GLN}}$	$\widetilde{\mathrm{RLN}}$
$\tilde{\mu}_1$	1.00	-0.47	-0.18	0.34	0.20	0.74	0.83	-0.08	0.58	-0.27	-0.31	0.87	0.18
$\tilde{g}_1$	-0.47	1.00	0.80	-0.28	0.07	-0.62	-0.43	0.14	-0.65	0.23	0.51	-0.60	-0.53
$\tilde{g}_2$	-0.18	0.80	1.00	-0.12	-0.07	-0.26	-0.17	-0.08	-0.28	-0.02	0.21	-0.24	-0.23
$\tilde{E}_1$	0.34	-0.28	-0.12	1.00	-0.07	0.39	0.32	-0.19	0.50	-0.26	-0.29	0.33	0.31
$\tilde{C}_1$	0.20	0.07	-0.07	-0.07	1.00	-0.08	0.17	0.75	-0.22	0.45	0.41	0.08	-0.29
$\tilde{H}_1$	0.74	-0.62	-0.26	0.39	-0.08	1.00	0.79	-0.51	0.89	-0.73	-0.68	0.86	0.51
$\tilde{I}_1$	0.83	-0.43	-0.17	0.32	0.17	0.79	1.00	-0.24	0.55	-0.45	-0.39	0.70	0.13
$\tilde{C}_3$	-0.08	0.14	-0,08	-0.19	0.75	-0.51	-0.24	1.00	-0.52	0.89	0.66	-0.23	-0.42
$\tilde{L}_6$	0.58	-0.65	-0.28	0.50	-0.22	0.89	0.55	-0.52	1.00	-0.70	-0.74	0.78	0.74
$\tilde{C}_9$	-0.27	0.23	-0.02	-0.26	0.45	-0.73	-0.45	0.89	-0.70	1.00	0.76	-0.43	-0.50
$\widetilde{\mathrm{LRE}}$	-0.31	0.51	0.21	-0.29	0.41	-0.68	-0.39	0.66	-0.74	0.76	1.00	-0.44	-0.60
$\widetilde{\mathrm{GLN}}$	0.87	-0.60	-0.24	0.33	0.08	0.86	0.70	-0.23	0.78	-0.43	-0.44	1.00	0.49
$\widetilde{\mathrm{RLN}}$	0.18	-0.53	-0.23	0.31	-0.29	0.51	0.13	-0.42	0.74	-0.50	-0.60	0.49	1.00

Note: $Corr(\tilde{X}, \tilde{Y})$ [0, 0.36) – low correlated (painted green), [0.36, 0.67) – moderately correlated (painted yellow), and [0.67, 0.9) – strongly correlated (pained orange) [150]. Subscripts for SOS parameters correspond to the GLCM distance d.

4.3.3 ML-based analysis of parameters of wide-field SHG imaging of thyroid nodule sections

The main phases of the whole cycle of data analysis are schematically shown in *Figure 4.14*. In total, 4 (FOS) + 5×5 (SOS) + 5 (HOS) = 34 intensity and texture parameters were considered in the image analysis.

To classify the recorded SHG images into different categories according to their intensity and texture parameters, a method of unsupervised machine learning of the *k*-means was employed. The cluster analysis was performed in two stages: first, to delineate the collagen capsule of the PTC from the surrounding tissue, and then to reveal the possible differences of the collagen structure between the areas of annotated capsular invasion and the intact capsule. The results of the clustering were also validated by the pathologist to exclude the under- or over-clustering. Prior to the *k*-means, the data was processed with the PCA to reduce the complexity of the available parameter set and also to be able to trace back which parameters have the most weight in separating the images to different classes. This way, the clustering was performed on the PCs rather than on the parameters directly.

In the first stage of the analysis, according to the Kaiser-Jolliffe rule [203] (*Figure 4.15e*) the first 4 PCs that cover 89% of the total data variance (*Figure 4.15a*) could be considered sufficient for a reduced data set.

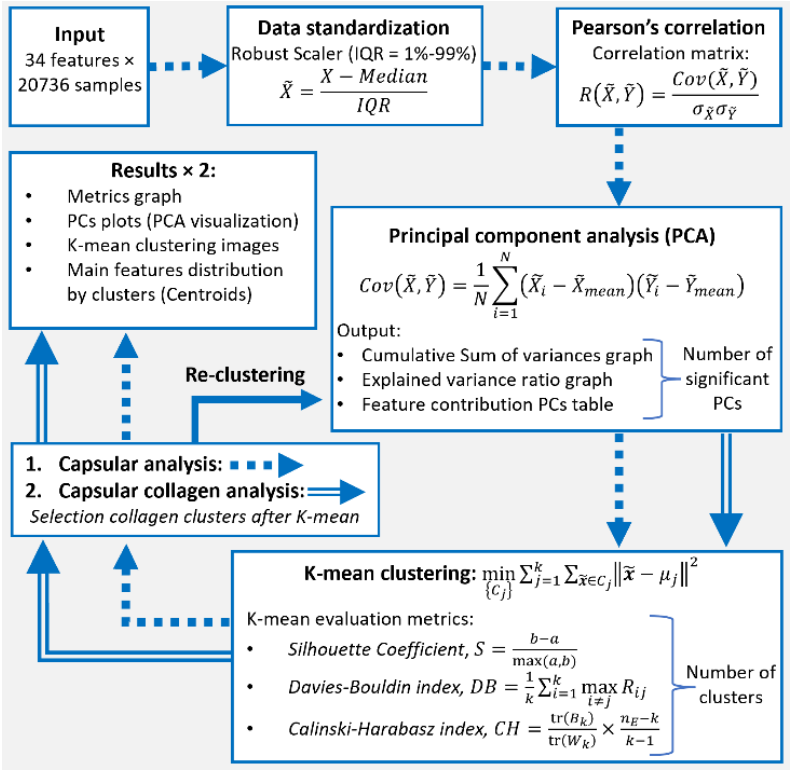


Figure 4.14 Sequence of data processing and ML-based analysis. Reprinted from [***Paper C***].

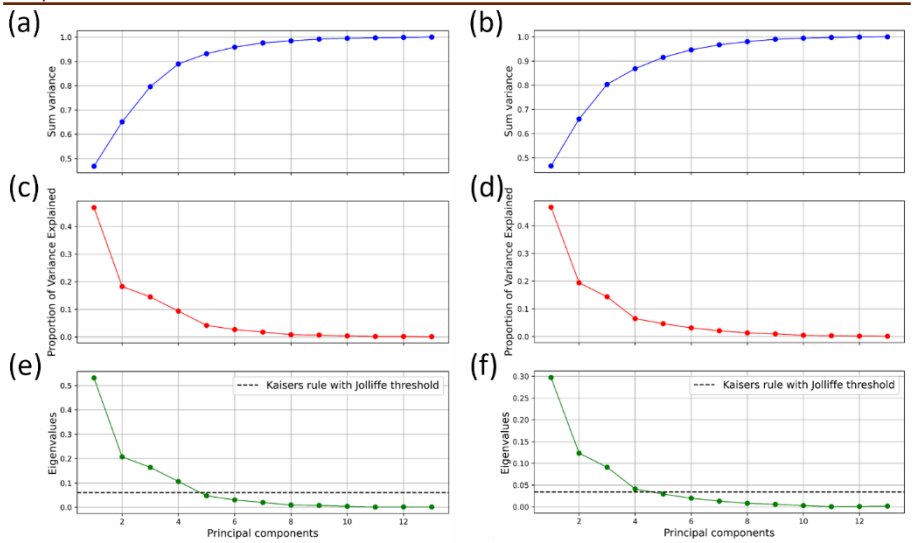


Figure 4.15 The number of PCs needed to explain variance: (a, b) – cumulative explained variance; (c, d) – the proportion of explained variance; (e, f) – eigenvalues plotted as a function of PCs. The PCA of the entire data set corresponds to (a), (c), (e). The PCA of the collagen-related data corresponds to (b), (d), (f). The dashed lines in (e) and (f) are Kaiser-Jolliffe thresholds for the selection of PCs to be retained. Reprinted from [***Paper C***].

The loadings that represent the contributions of the different features in the selected PCs are presented in *Table 4-5*. PC1 is responsible for 46.8% of the total data variance and, as will be shown below, it separates the SHG images of collagen from the rest of the tissue. Parameters with the relative weight in the PC1 larger than 5% are $\tilde{E}_1$, $\tilde{L}_6$, $\tilde{H}_1$, $\widetilde{GLN}$, $\widetilde{RLN}$, $\tilde{C}_9$, and $\widetilde{LRE}$. This implies that the data distribution along the PC1 is mainly due to the differences in the SHG image texture: roughness, short- and long-range order in the distribution of pixel intensities, structural complexity (*Table 3-3*). On the other hand, since the relative weight of $\tilde{\mu}_1$ to PC1 is less than 5%, the intensity of the SHG signal associated with the collagen content plays a minor role in this regard.

The largest loadings of the PC2 correspond to $\tilde{E}_1$, $\tilde{C}_9$, $\tilde{C}_3$, $\widetilde{LRE}$, and $\tilde{C}_1$. This implies that the separation along the PC2 occurs between regularly structured and uniform images such as arrays of collagen fibers versus glass or disordered collagen.

The FOS intensity parameters $\tilde{\mu}_1$, $\tilde{g}_1$, and $\tilde{g}_2$ have a major contribution to PC3 and PC4, and as will be shown further, determine the segmentation within the collagen capsule. Then, the number of clusters k of the k-means was determined as follows. The k-means metrics were calculated for different values of k = 2-10 (*Figure 4.16a, c, e*).

Table 4-5 Loadings of the PCs of the PCA of the whole dataset. Adapted from [***Paper C***].

PC 1 (46.8%)			PC 2 (18.3%)			PC 3 (14.5%)			PC 4 (9.4%)		
Feature	Loading	Rel. weight, %	Feature	Loading	Rel. weight, %	Feature	Loading	Rel. weight, %	Feature	Loading	Rel. weight, %
$\tilde{E}_1$	0.512	26.2	$\tilde{E}_1$	0.774	59.9	$\tilde{g}_2$	0.601	36.2	$\tilde{g}_2$	0.475	22.6
$\tilde{L}_6$	0.376	14.2	$\tilde{C}_9$	0.349	12.2	$\tilde{E}_1$	0.509	25.9	$\tilde{\mu}_1$	0.474	22.5
$\tilde{H}_1$	0.345	11.9	$\tilde{C}_3$	0.332	11.0	$\tilde{g}_1$	0.406	16.5	$\tilde{C}_1$	0.026	0.1
$\widetilde{\mathrm{GLN}}$	0.259	6.7	$\widetilde{\mathrm{LRE}}$	0.246	6.0	$\widetilde{\mathrm{RLN}}$	0.271	7.4	$\tilde{I}_1$	0.015	0.0
$\widetilde{\mathrm{RLN}}$	0.224	5.0	$\tilde{C}_1$	0.235	5.5	$\widetilde{\mathrm{LRE}}$	0.236	5.6	$\widetilde{\mathrm{GLN}}$	0.015	0.0
$\tilde{I}_1$	0.188	3.5	$\tilde{\mu}_1$	0.038	0.1	$\tilde{L}_6$	0.107	1.1	$\tilde{g}_1$	-0.019	0.0
$\tilde{\mu}_1$	0.176	3.1	$\tilde{g}_1$	0.020	0.1	$\tilde{H}_1$	0.102	1.0	$\tilde{H}_1$	-0.031	0.1
$\tilde{C}_1$	-0.091	0.8	$\tilde{I}_1$	-0.013	0.0	$\tilde{I}_1$	0.099	1.0	$\widetilde{\mathrm{LRE}}$	-0.045	0.2
$\tilde{g}_2$	-0.135	1.8	$\widetilde{\mathrm{GLN}}$	-0.044	0.2	$\tilde{\mu}_1$	0.020	0.0	$\tilde{C}_3$	-0.135	1.8
$\tilde{g}_1$	-0.197	3.9	$\tilde{g}_2$	-0.050	0.3	$\tilde{C}_9$	-0.058	0.3	$\tilde{L}_6$	-0.326	10.6
$\tilde{C}_3$	-0.218	4.8	$\tilde{L}_6$	-0.111	1.2	$\widetilde{\mathrm{GLN}}$	-0.113	1.3	$\tilde{C}_9$	-0.357	12.7
$\widetilde{\mathrm{LRE}}$	-0.288	8.3	$\widetilde{\mathrm{RLN}}$	-0.119	1.4	$\tilde{C}_3$	-0.134	1.8	$\tilde{E}_1$	-0.379	14.4
$\tilde{C}_9$	-0.312	9.8	$\tilde{H}_1$	-0.146	2.1	$\tilde{C}_1$	-0.137	1.9	$\widetilde{\mathrm{RLN}}$	-0.387	15

Note: Loadings of PCs are arranged in descending order. The loadings with the relative weight above 5% are highlighted green.

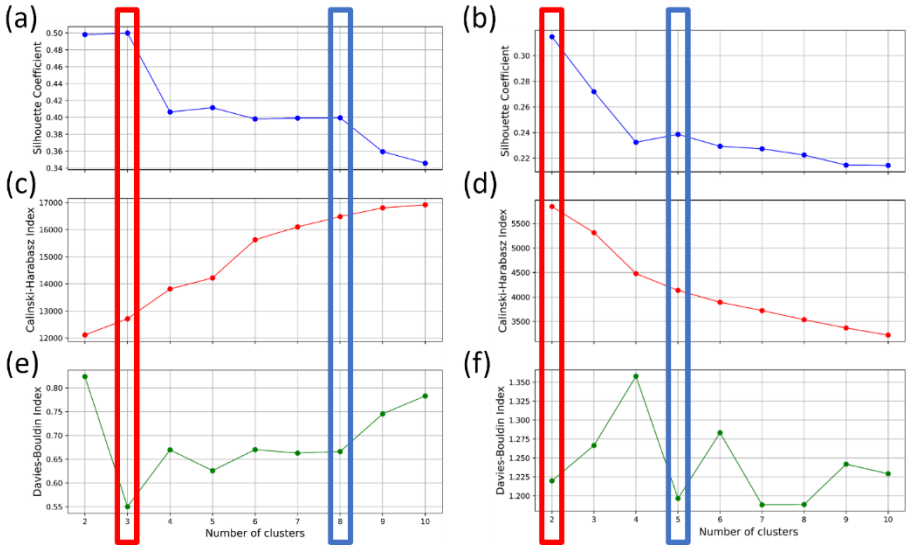


Figure 4.16 Evaluation metrics for k-means clustering: (a, c, e) – clustering of the entire dataset; (b, d, f) – clustering of collagen-related SHG images. The rectangles show the optimal selection of the number of clusters according to the combination of metrics: red rectangles correspond to the trivial data segmentation; blue rectangles correspond to the number of clusters used for *k*-means. Reprinted from [***Paper C***].

The best combination of high SC and low DBI was achieved with $k = 3$ (*Figure 4.16a, e* red rectangles). However, such clustering is trivial since it categorizes SHG images into outliers, those of the glass slide, and then the rest without separating the capsule (*Figure 4.17*).

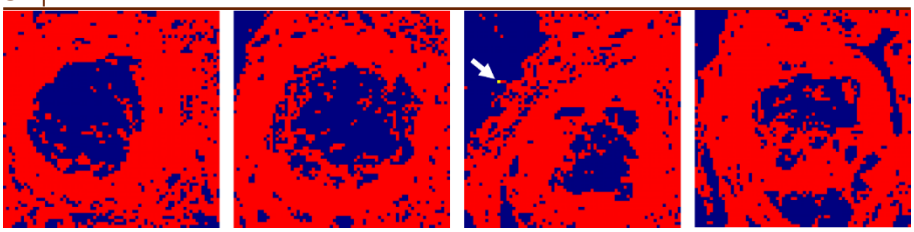


Figure 4.17 *k*-Means clustering of all examined PTC nodule sections with trivial clustering (k = 3). The arrow shows the outlier tile image (colored yellow). The size of the cluster maps is 8.4 mm × 8.4 mm. Reprinted from [***Paper C***].

With larger k = 4-8 the SC and DBI are both fairly constant, while the CHI monotonously increases with increasing k. We, therefore, settled for k = 8, which is the number of clusters that appears to be appropriate to account for all the different outliers, sample areas without tissue, and also to segregate the remaining tissue images into different physically meaningful categories.

The detailed results of the k-means clustering with k = 8 for the PTC nodule sections are shown in *Figure 4.18c, f, i, l*.

Of the 8 clusters, two (clusters 6 and 7) encompass the outlier data that are scarce and characterized by extreme values of individual parameters as compared to the rest of the dataset. SHG images of the glass slide and cancer cells generating background levels of SHG signal fall into cluster 5.

Then, by looking at the localization of the different clusters in the samples (*Figure 4.18*), it appears that the rest of the clusters, albeit somewhat conditionally, could be assigned to collagen around large vessels near the capsule (cluster 0, 93 SHG images, *Figure 4.18*), the collagen capsule surrounding the PTC nodule and collagen spreading to normal tissue (clusters 1 and 2), normal follicles (cluster 3), and possibly inflamed tissue (cluster 4).

The SHG images of clusters 0, 3-7 are shown in Figure S7 Supplementary Material of ***Paper C***.

The centroids of the standardized parameters attributed to the different clusters are presented in Figure 4.19*g*. The score plots of the different pairs of PCs are shown in Figure 4.19*a-f*. The 3D score plot of the first three PCs is presented in Visualization 1 [***Paper C***]. The data points are colored the same as their assigned clusters in the clustering maps (*Figure 4.18*). It appears that the collagen-related clusters 0-2 are well separated from the remaining clusters 3-5 along the PC1 (Figure 4.19*a, c, e*).

According to PC1 loading (*Table 4-5*), the separation of the capsule from the rest of the tissue is mainly due to the variance in SOS parameters related to coarseness/smoothness of the SHG image and the long-range order parameters. Indeed, as it follows from cluster centroids (Figure 4.19*g*), collagen-related clusters are characterized by high positive values of $\tilde{E}_1$, $\tilde{L}_6$, $\widetilde{H}_1$, $\widetilde{GLN}$ and negative value of $\widetilde{LRE}$, while for the other clusters these parameters have opposite signs. In terms of image morphology, high $\tilde{E}_1$ indicates the presence of dominant levels of brightness. High $\tilde{L}_6$ implies

frequent occurrence of pixel pairs of the same gray level at a distance d = 6 px.

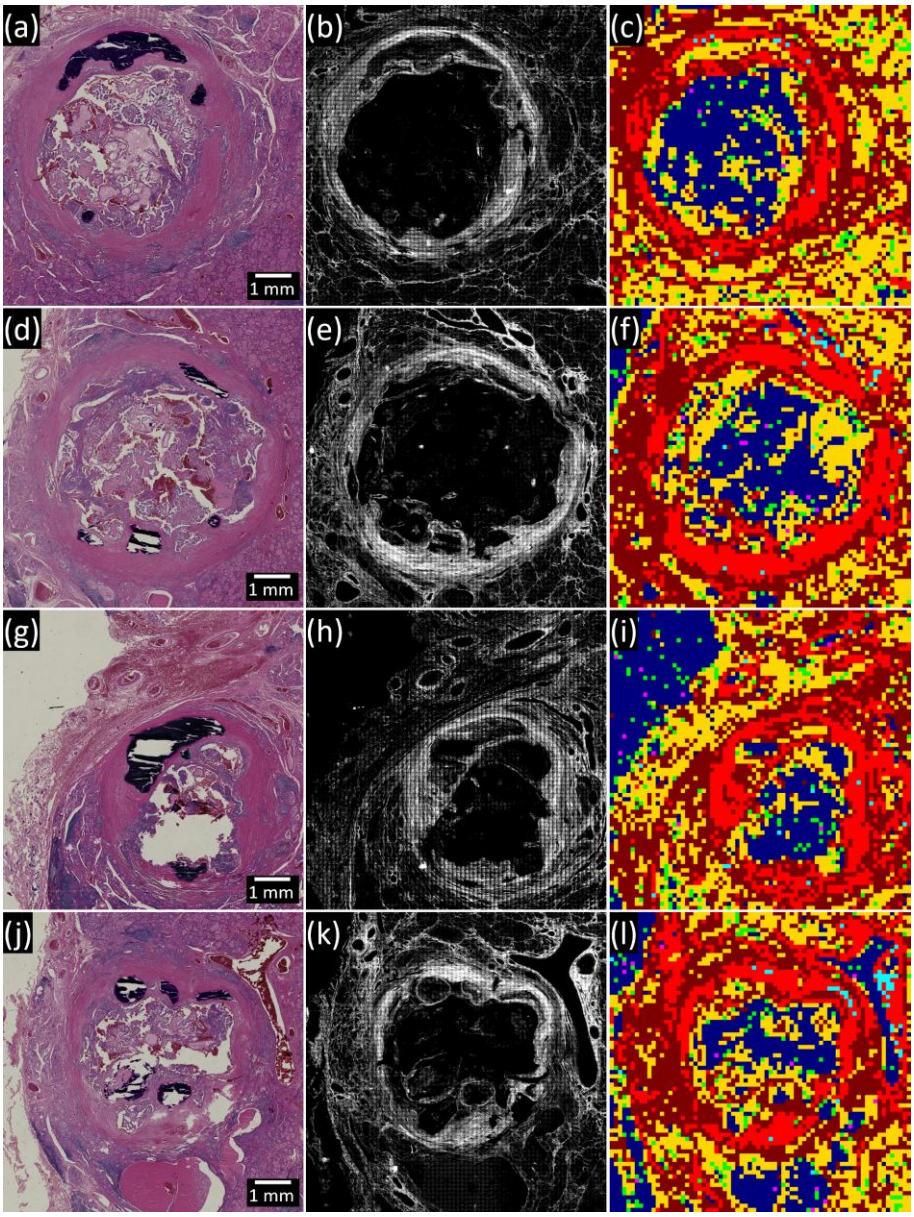


Figure 4.18 *k*-Means clustering of the analyzed PTC nodule sections: (a, d, g, j) – bright-field images; (b, e, h, k) – SHG images; (c, f, i, l) – k-means clustering, k = 8. Adapted from [***Paper C***].

High $\widetilde{H}_1$ indicates coarse texture, and high $\widetilde{GLN}$ means significant non-uniformity in the grey level runs and complexity of the texture. Unsurprisingly, in contrast to this, non-collagenous clusters of thyroid colloid

and glass are characterized by high $\widetilde{LRE}$, indicating long pixel runs and large homogeneous areas.

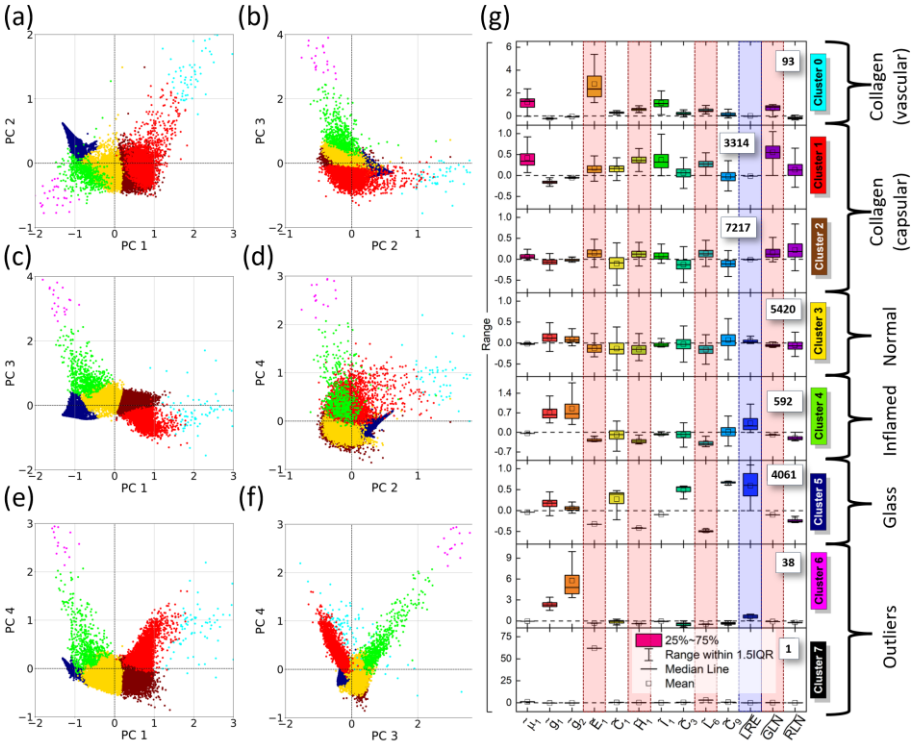


Figure 4.19 *k*-Means clustering of an SHG image of PTC nodule sections based on PCA. (a)-(f) Score plots. (g) Centroids of standardized parameters of all clusters. The red and blue dashed rectangles mark the parameters with the major positive and negative contributions, respectively, to the data separation along the PC1. The data points in (a)-(f) and the clusters in *Figure 4.18c, f, i, l* have the same color as the clusters in (g) and are categorized identically. Adapted from [***Paper C***].

The reconstructed original, unstandardized parameters can be found in Figure S8 Supplementary Material of ***Paper C***.

4.3.4 ML-based analysis of parameters of wide-field SHG imaging of collagen capsules.

To focus on the specific changes in the collagen capsule around PTC, two collagen clusters (1 and 2) mainly related to the capsule were selected and their additional clustering was performed. All other data (clusters 0, 3-7) were excluded from the analysis and colored black in the resulting cluster maps.

The parameters of the selected SHG images were re-standardized within the limits of a new, reduced data set and PCA was repeated. Four PCs covering 86.9% of the data variance were selected using the Kaiser-Jolliffe rule (see

Figure 4.15b, d, f) and used for clustering. The PC loadings for the new dataset are shown in *Table 4-6*.

Table 4-6 Loadings of the PCs for PCA of the collagen-related dataset. Adapted from [***Paper C***]

PC 1 (46.7%)			PC 2 (19.4%)			PC 3 (14.3%)			PC 4 (6.5%)		
Feature	Loading	Rel. weight, %	Feature	Loading	Rel. weight, %	Feature	Loading	Rel. weight, %	Feature	Loading	Rel. weight, %
$\widetilde{GLN}$	0.440	19.4	$\widetilde{RLN}$	0.559	31.2	$\tilde{C}_3$	0.601	36.2	$\tilde{C}_9$	0.475	22.5
$\tilde{H}_1$	0.353	12.4	$\tilde{E}_1$	0.340	11.6	$\tilde{C}_9$	0.509	26.0	$\widetilde{LRE}$	0.474	22.4
$\tilde{\mu}_1$	0.337	11.3	$\tilde{L}_6$	0.339	11.5	$\tilde{C}_1$	0.406	16.5	$\tilde{C}_3$	0.026	0.1
$\tilde{L}_6$	0.291	8.5	$\tilde{C}_9$	0.156	2.4	$\tilde{g}_1$	0.271	7.3	$\tilde{\mu}_1$	0.015	0.0
$\tilde{C}_3$	0.239	5.7	$\widetilde{GLN}$	0.002	0.0	$\tilde{g}_2$	0.236	5.6	$\tilde{L}_6$	0.015	0.0
$\tilde{C}_1$	0.230	5.3	$\tilde{C}_3$	-0.035	0.1	$\widetilde{LRE}$	0.107	1.1	$\tilde{E}_1$	-0.019	0.0
$\tilde{I}_1$	0.230	5.3	$\tilde{g}_2$	-0.132	1.8	$\tilde{E}_1$	0.102	1.0	$\tilde{H}_1$	-0.031	0.1
$\tilde{C}_9$	0.136	1.8	$\tilde{g}_1$	-0.165	2.7	$\tilde{L}_6$	0.099	1.0	$\tilde{I}_1$	-0.045	0.2
$\tilde{E}_1$	0.122	1.5	$\tilde{\mu}_1$	-0.183	3.4	$\widetilde{RLN}$	0.020	0.0	$\widetilde{GLN}$	-0.135	1.8
$\widetilde{RLN}$	0.009	0.0	$\widetilde{LRE}$	-0.209	4.4	$\tilde{\mu}_1$	-0.058	0.3	$\tilde{g}_1$	-0.326	10.6
$\widetilde{LRE}$	-0.272	7.4	$\tilde{H}_1$	-0.260	6.8	$\tilde{I}_1$	-0.113	1.3	$\tilde{g}_2$	-0.357	12.8
$\tilde{g}_2$	-0.306	9.4	$\tilde{C}_1$	-0.317	10.0	$\tilde{H}_1$	-0.134	1.8	$\tilde{C}_1$	-0.379	14.3
$\tilde{g}_1$	-0.346	12.0	$\tilde{I}_1$	-0.377	14.2	$\widetilde{GLN}$	-0.137	1.9	$\widetilde{RLN}$	-0.387	15.0

Note: Loadings of PCs are arranged in descending order.

k-Means clustering was performed according to the previously described procedure. The number of clusters $k = 5$ was estimated based on the most optimal set of cluster metrics (see *Figure 4.16b, d, f*). The cluster maps of collagen distribution in all 4 analyzed PTC sections, the score plots and the corresponding centroids are shown in *Figure 4.20*. The relevant 3D score plot of the first three PCs is presented in Visualization 2 [***Paper C***]. The centroids of the non-standardized parameters are shown in Figure S9 Supplementary Material of ***Paper C***.

Cluster maps reveal significant heterogeneity of the collagen structure in the capsule surrounding PTC (*Figure 4.20a-d*). The capsule consists mainly of three clusters (1-3). Cluster 1 (green in *Figure 4.20*) forms the core part of the capsule and is virtually absent in normal tissue with follicles. Although it tends to form a continuous core of the entire capsule, there are clearly recognizable areas where it is replaced by cluster 2 or 3 (red and brown in *Figure 4.20*, respectively) or their mixture. Cluster 4 (yellow in *Figure 4.20*) is adjacent to the capsule from the outside. Cluster 5 (dark blue in *Figure 4.20*) is located almost entirely outside the capsule and probably comprises the septa of normal follicles. Calcifications are represented by collagens mainly assigned to clusters 2, 4 and 5 (marked with arrows in *Figure 4.20a-d*).

PCA allowed revealing the main structural and textural characteristics of collagen, which determined the segmenting of the capsule into clusters.

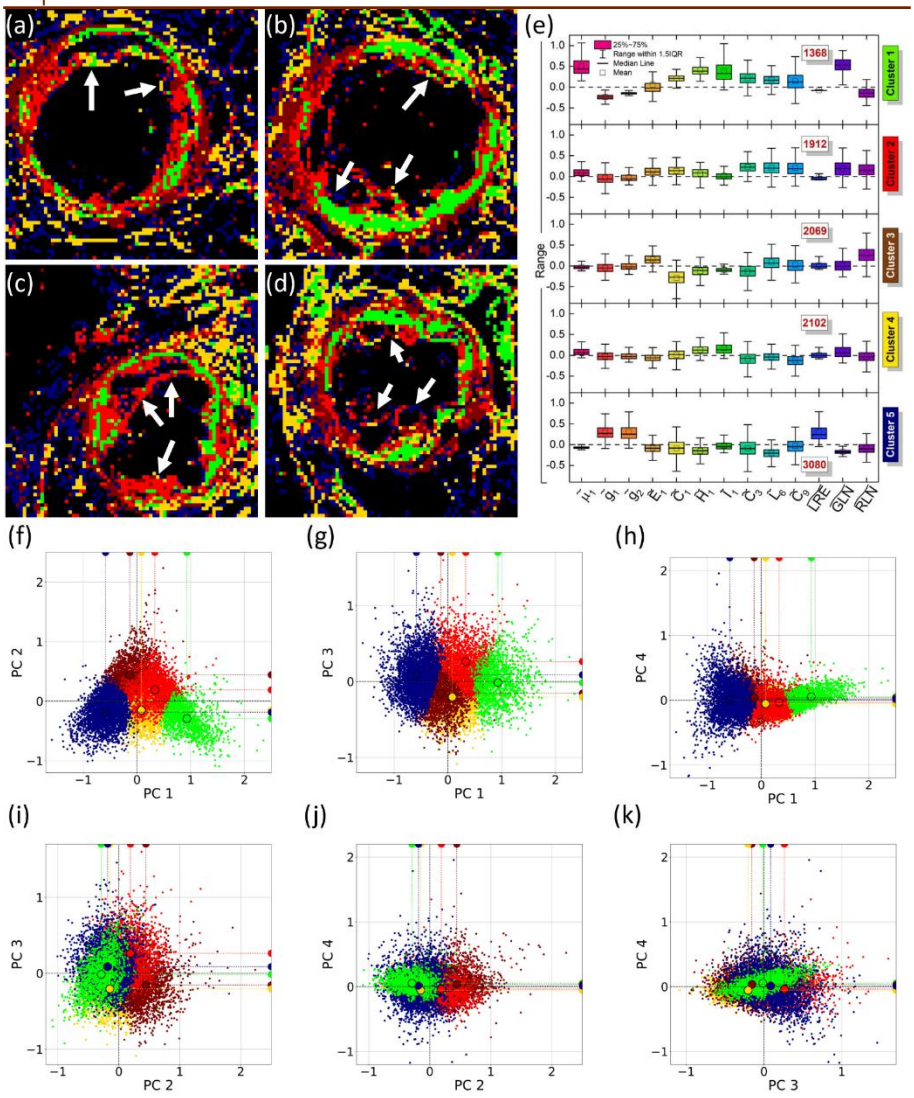


Figure 4.20 *k*-Means clustering of collagen-related clusters only. (a-d) Cluster maps of 4 analyzed PTC nodule sections. (e) Centroids composed of standardized parameters corresponding to each cluster. (f-k) Score plots. The reconstructed original, unstandardized parameters can be found in Figure S9 Supplementary Material of ***Paper C***. The colored circles indicate the position of the centroid of the cluster (center of mass of the cluster) in PC space; the colored semicircles on the abscissa and ordinate axes mark the projections of the centroids. The size of the cluster maps is 8.4 mm × 8.4 mm. The data points in (f)-(k) and the clusters in (a)-(d) have the same color as the clusters in (e) and are categorized identically. Reprinted from [***Paper C***].

PC1 separates clusters 1 and partially cluster 2 from other clusters. The data are mainly segmented by the significant influence of $\widetilde{GLN}$, $\tilde{H}_1$, $\tilde{\mu}_1$, $\tilde{L}_6$, $\tilde{C}_3$,

$\tilde{C}_1$, $\tilde{I}_1$ with positive sign and $\widetilde{LRE}$, $\tilde{g}_2$, $\tilde{g}_1$ with negative sign in the PC1 loadings (*Table 4-6*). High centroid value of entropy $\tilde{H}_1$ and the non-uniformity of runs of grey levels $\widetilde{GLN}$ of clusters 1 and 2 (*Figure 4.20e*) indicate coarse-grained textures and the presence of clearly separated elongated fibers. High mean $\tilde{\mu}_1$ values, high local homogeneity $\tilde{L}_6$ and inertia $\tilde{I}_1$ indicate high collagen content, high contrast of the image, and are probably related to the presence of highly dense collagen bundles [35]. High correlations $\tilde{C}_3$, $\tilde{C}_1$ indicate the periodicity (regularity) of the collagen fibers. Negative values of $\widetilde{LRE}$ denote the absence of large repetitions of pixels with the same grey levels, which usually originate from empty glass or uniformly distributed collagens that occupy the entire imaged area. The non-standardized values g_1 and g_2 are positive for all clusters, so that the gray level distributions are leptokurtic and right-skewed [204] in all cases. However, the g_2 value for cluster 1 is smaller than for the other clusters, indicating a broader intensity distribution which can be explained by a higher heterogeneity of the images. Although cluster 2 has high values of the correlations of distant and neighboring pixels $\tilde{C}_3$, $\tilde{C}_9$, $\tilde{C}_1$ indicating similarity of cluster 2 to tight collagen of cluster 1, the g_1 and g_2 values are higher indicating collagen disorder (or degradation). Clusters 3-5 consist of low-contrast fibers that produce an SHG signal with a narrower leptokurtic right-skewed intensity distribution, and form disordered (complex) networks that only partially cover the scanned area or contain evenly distributed collagen (cluster 3, *Figure 4.21*).

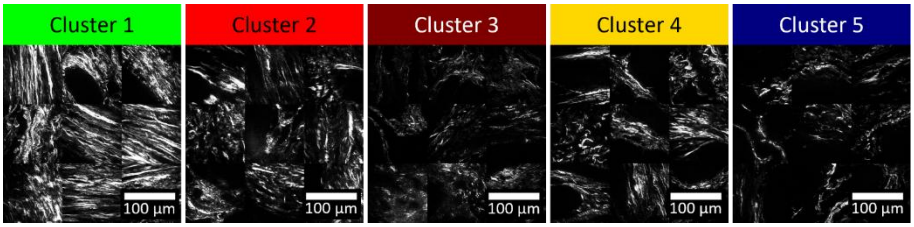


Figure 4.21 SHG images of typical collagen structures related to clusters separated by *k*-means in PTC capsule. Images consist of nine randomly selected SHG tile images from the datasets corresponding to each cluster. Reprinted from [***Paper C***].

PC2 separates cluster 3 (brown) and partially cluster 2 (red) from the other clusters (*Figure 4.20f, i, j*) due to the large values of PC2 loadings of $\widetilde{RLN}$, $\tilde{E}_1$, $\tilde{L}_6$ and $\tilde{H}_1$, $\tilde{C}_1$, I_1 , with positive and negative signs, respectively (*Table 4-6*). Cluster 3 is characterized by high positive values for $\widetilde{RLN}$, $\tilde{E}_1$, $\tilde{L}_6$ and negative values for $\tilde{H}_1$, $\tilde{C}_1$, and $\tilde{I}_1$ (*Figure 4.20e*). Structurally, this indicates that cluster 3 consists of collagen fibers with high variability of length and dominant levels of brightness.

Collagen networks have lower contrast and are more uniform. Such parameter values may indicate fragmentation of collagen fibers and an

increased level of collagen degradation in cluster 3. In contrast to cluster 3, cluster 2 is partially separated by PC2. It is characterized by a more uniform distribution of the lengths of collagen fibers (lower $\widetilde{RLN}$ values), a lower uniformity in fiber spatial distribution (more disordered network) (lower $\tilde{E}_1$) and the presence of some elongated, clearly separated fibers (higher values of $\tilde{L}_6$, $\tilde{H}_1$, $\tilde{C}_1$, I_1), although the fibers are mainly fragmented (*Figure 4.21*).

PC1 and PC2 cover 66.1% of the data variance and reveal the main differences in collagen structures within the capsule. PC3 allows only partial separation of clusters 2 and 5 from clusters 1, 3 and 4 (*Figure 4.20g, i, k*), although all the clusters overlap significantly, and direct interpretation of the morphological features is complicated. PC4 does not reveal any significant data separation (*Figure 4.20h, j, k*).

Comparison of the clustering maps of collagen distribution in whole PTC capsules and H&E-stained BF images annotated by the pathologist revealed a correlation between the spatial distribution of clusters and the invasive/non-invasive areas in all tissue samples. Areas with dominant cluster 1 surrounded by clusters 2 and 3 were not labeled as invasion in any of the samples examined (white doubled rectangles, *Figure 4.22*). Most of the annotated invasions (rectangles with solid line, *Figure 4.22*) are mainly composed of cluster 2 or 3 (or both) with minor inclusions of clusters 4 and 5 and with some rare inclusions of cluster 1 which do not form the continuous core. There are two exceptions, which can be seen in *Figure 4.22* (marked with an asterisk), where cluster 1 forms a continuous barrier. However, these annotations can be classified as invasions that have not yet occurred [205]. During primary analysis, some of the areas represented by clusters 2 and/or 3 and missing cluster 1 were not categorized as invasion areas by the pathologist (*Figure 4.22*, dotted rectangles). These areas were marked as suspicious areas requiring additional analysis. Subsequent examination of one of these suspicious areas, revealed by ML-assisted wide-field SHG microscopy, allowed clarification of the diagnosis and classification of this area as a microinvasion (marked with arrows in *Figure 4.22*). An additional examination of two exceptions among the annotated invasion clusters proved the relationship of these clusters to microinvasions.

The presence and dominance of collagen structures classified as cluster 1 (*Figure 4.20e, Figure 4.21*) were thought to be associated with a non-invasive capsule and corresponding collagen structures may prevent invasion (double rectangle, *Figure 4.22*) [206].

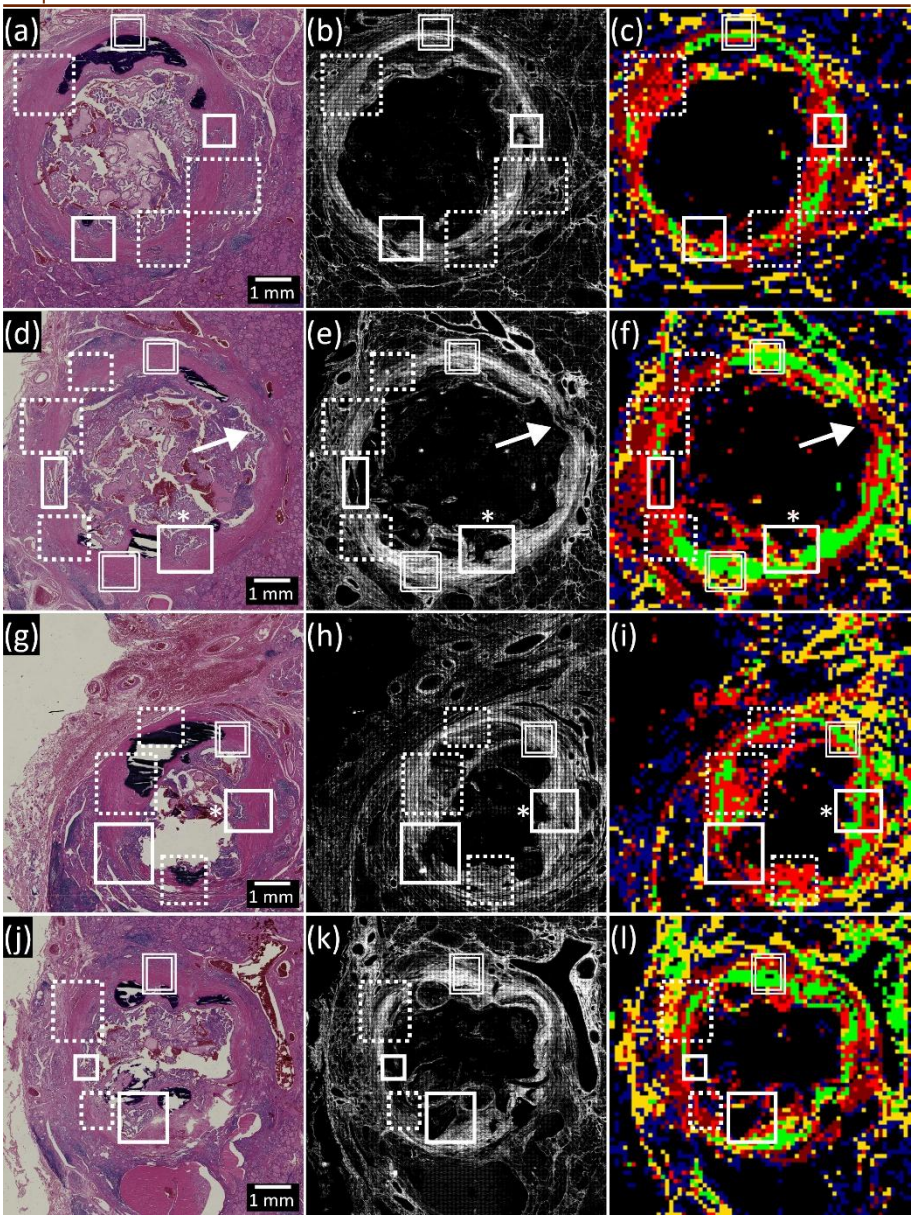


Figure 4.22 Annotations of invaded, affected, and non-affected areas in PTC collagen capsules. (a, d, g, j) Bright-field image of H&E-stained tissue section. (b, e, h, k) SHG image. (c, f, i, l) Cluster map. White rectangles denote the following specific regions in the PTC sections: capsular invasion annotated by pathologist (solid), suspicious or affected by PTC progression, annotated by *k*-means (dotted), non-invaded, annotated by *k*-means (double). Adapted from [***Paper C***].

Dense packing of long, aligned collagen fibers in the area of carcinogenesis has been shown to be one of the mechanisms for preventing

tumor growth, as the alignment of collagen fibers facilitates the motility of the key players in the anti-cancer immune response, CD8+ T cells [207], and their infiltration into cancer cells. If the alignment of the fibers is disturbed and the collagen fiber network is tangled, the movement of CD8+ T cells is impeded [208], which promotes tumor growth. One could hypothesize that the latter is reflected in the replacement of cluster 1 collagens in the cluster maps with collagens classified as cluster 2 and cluster 3, indicating the development of invasion or highlighting areas that are suspicious (or susceptible) to invasion. Clusters 2 and 3 are represented by a sparse network of thin fibers. Collagen degradation, fiber realignment, random orientation and fragmentation were shown to support the migration of invading cancer cells and promote metastasis. Such remodeling is associated with the activity of cancer-associated fibroblasts [209] and TAM [210], which create a pro-invasive and immunosuppressive microenvironment [211].

Our findings indicate that analyzing images of entire tissue sections from surgically removed thyroid nodules provides a comprehensive picture of the PTC progression. ML-enhanced SHG wide-field microscopy of the entire PTC nodule sections, in contrast to previously published SHG imaging of small-size areas [20,212,213], allowed for revealing the local heterogeneity of capsular collagen, closely related to the invasion or the susceptibility of the capsule to invasion. Data segmentation within the collagen clusters occurs mainly along PC1 and PC2. PC1 isolates tight collagen of the inner core which is associated with encapsulated PTC (cluster 1), while PC2 isolates disordered, fragmented collagen which is associated with already invaded or suspicious parts of the capsule (clusters 2 and 3). This shows that the proposed approach, in addition to traditional analysis, can help to clarify and/or confirm a diagnosis, identify overlooked, poorly distinguishable invasions, and highlight suspicious areas that require closer examination.

Identification of areas of invasion in thyroid nodules is critical for both diagnosis and treatment, as it has significant implications for determining the nature of the nodule and subsequent clinical decisions.

The primary benefit of identifying invasion is to differentiate between benign and malignant thyroid nodules [214]. Benign nodules typically do not invade the surrounding tissue, whereas malignant nodules, such as papillary or follicular thyroid carcinomas, often show invasive characteristics. The presence of areas of invasion is a key histopathologic criterion for the diagnosis of malignancy in follicular adenomas and carcinomas of the thyroid. These two pathologies share the same cytologic features, and differentiation can only be made by identifying capsular and/or vascular invasion.

On the other hand, the extent of invasion can provide information about the aggressiveness of the tumor [215]. Tumors that invade far into surrounding tissue such as muscles or vascular structures are often more aggressive and are associated with a poorer prognosis. This information is important for predicting outcomes and planning appropriate treatment strategies.

In terms of treatment, the presence and extent of areas of invasion are critical to the surgical approach [216]. For example, a nodule with localized invasion may require a more extensive resection than just a lobectomy, in which case a total thyroidectomy is required. Cases of extensive invasion, such as extrathyroidal extension or lymph node metastases, may require adjuvant therapies such as radioactive iodine ablation to target the residual malignant tissue. This would not be the case for fully encapsulated nodes with no signs of invasion.

Finally, the presence of invasion also affects the monitoring of patients after surgery and possible adjuvant therapies. For invasive or high-risk tumors, more aggressive surveillance is required to early detect recurrence or metastases.

Thus, any progress in increasing the accuracy of invasion identification increases the quality of diagnosis, treatment and patient survival. The results and the ML-based approach proposed in the current study make an important contribution to this progress.

Although traditional histopathology remains the golden standard for the diagnosis of capsular invasion, SHG microscopy used in this study offers several technical benefits for data accumulation. SHG microscopy provides information on the collagen within the thyroid nodule capsule. Not only the collagen capsule, but also the cellular context both inside and outside the nodule are important for malignant/benign classification using random forest [20] and DL approaches, as recently shown using optical microscopy datasets [217]. Hence, adding cellular context to the collagen capsule might improve the classification strategy. In the context of nonlinear optical imaging, this might come straightforward as SHG-related techniques can be easily implemented with SHG imaging: THG, TPEF and CARS. Such approaches have already been applied to demonstrate qualitative characteristic of various thyroid pathologies [218]. The next step would be to find suitable classification approaches dealing with such multidimensional data. In this context one of the possible approaches is related to image fusion techniques. For example, a fusion autoencoder [219] would take a stack of SHG, THG, and TPEF images as inputs and provide feature maps with the fused information. Subsequently, a classifier would divide the feature maps into benign and malignant categories.

To sum up, the proposed approach of automated ML-based selection of collagen-related SHG images and the findings on the heterogeneity of the PTC capsule can very likely form the basis for the development of new effective models of automatic diagnosis based on unsupervised ML algorithms with further extension to supervised ML models. The results of the current research indicate that the traditional approach with manual selection of ROIs for training the supervised ML models is inconsistent due to the heterogeneity of PTC capsules. Their heterogeneity, which cannot be completely detected during manual inspection, can cause the largest error in the set of selected ROIs of the "intact" (control) capsule, as they contain both normal and

unidentified 'suspicious' areas of the capsule. As a result, the supervised ML model cannot be trained correctly because one of the data sets for training ('intact' capsule) is ambiguous. The approach proposed in the present study, potentially complemented with "cellular context" from THG, TPEF and CARS can be an efficient method for automated approach for cancer diagnosis.

This was summarized in the third statement of the thesis: *Unsupervised machine learning improves SHG image analysis, reveals the textural heterogeneity of papillary thyroid carcinoma capsule, and enables identification of capsular invasion, poorly distinguishable microinvasions and regions requiring additional examination based on the specific sets of image parameters.*

4.4 Supervised ML for thyroid carcinoma diagnosis using wide-field SHG microscopy

The results from this section were demonstrated in ***Paper D*** and in ***Conferences 7*** and ***9***.

All FTC and PTC samples were imaged using the SHG microscopy setup and the combined images of all samples are shown in Figures S1, S2 from ***Paper D***. To ensure diverse sample description, 34 intensity and texture features (4 FOS, 25 SOS and 5 HOS) extracted from each 117 µm × 117 µm SHG image tile were used for further analysis. However, not all the features are necessarily highly discriminative and relevant to the target, which may affect the classification results. To evaluate the impact of feature redundancy and irrelevance on classification, feature selection was performed using RFECV-LinearSVC followed by MIFS. The results of feature selection for each label correction approach are shown in *Table 4-7*.

Table 4-7 Features selected by RFECV-LinearSVC and MIFS within each label correction approach. Adapted from [***Paper D***]

Label correction approach	Split	Excluded by RFECV-LinearSVC; (excluded f_i, No.)	Excluded by MIFS; (excluded f_i, No.)	No. of remained f_i
I. Tissue-related	70/30	E_{12}, C_1; (2)	C_{12}, I_9, C_6, I_3, C_3; (5)	27
	80/20	*RLN*, I_9, H_1, C_1; (4)	C_{12}, C_6, I_3, C_3; (4)	26
	90/10	*RLN*, E_{12}, I_9, L_6, E_6, H_1, E_1; (7)	C_{12}, C_6, I_3; (3)	24
II. Capsule-related	70/30	*LRE*, L_{12}, E_{12}, E_9, H_6, E_6, I_3, E_1, μ_1; (9)	(0)	25
	80/20	*GLN*, *LRE*, L_{12}, C_{12}, E_{12}, E_9, H_6, E_6, I_3, E_3, E_1, μ_1; (12)	(0)	22
	90/10	*LRE*, E_{12}, E_6, μ_1; (4)	(0)	30
III. Multi-class, accounts for capsular heterogeneity	70/30	E_9, E_6; (2)	(0)	32
	80/20	I_9; (1)	(0)	33
	90/10	I_9, E_9, E_6, E_3; (4)	(0)	30

Footnote: Lower index indicates the step (in px) used for calculating GLCM. f_i – features.

All FTC and PTC samples were imaged using the SHG microscopy setup and the combined images of all samples are shown in Figures S1, S2 from

Paper D. To ensure diverse sample description, 34 intensity and texture features (4 FOS, 25 SOS and 5 HOS) extracted from each 117 µm × 117 µm SHG image tile were used for further analysis. However, not all the features are necessarily highly discriminative and relevant to the target, which may affect the classification results. To evaluate the impact of feature redundancy and irrelevance on classification, feature selection was performed using RFECV-LinearSVC followed by MIFS. The results of feature selection for each label correction approach are shown in *Table 4-7*.

4.4.1 Tissue-related wide-field SHG images of PTC and FTC

The tissue-related SHG images were separated from the non-tissue-related images using PCA of the texture feature vectors extracted from SHG images. Binary *k*-means clustering, based on the first five PCs covering more than 92% of data variance, enabled segmentation of the tissue section image into tissue- and non-tissue related SHG images [177]. A typical SHG image segmentation is shown in *Figure 4.23*. Such clustering separates tissue and non-tissue related points in two well-defined clusters. As previously shown [16], tissue- and non-tissue-related data points (glass) have projections with opposite signs on PC1 in the score plots and can be clearly visualized (*Figure 4.23*).

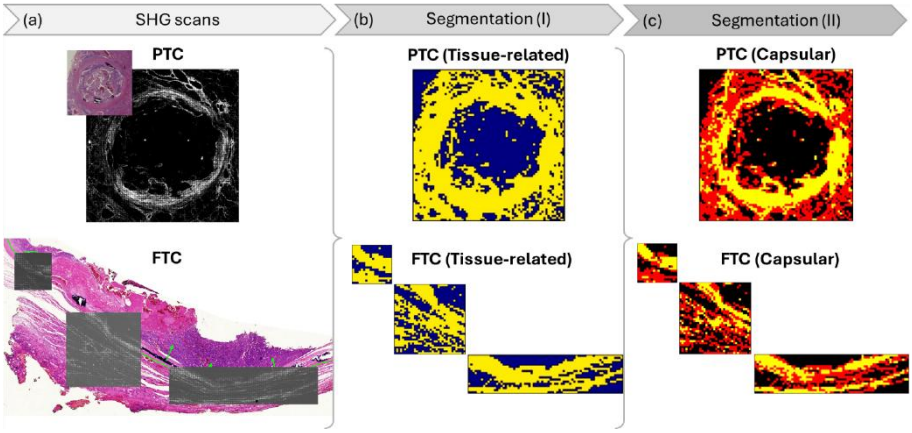


Figure 4.23 Typical SHG and brightfield images of PTC and FTC nodule sections (a) and the result of the separation of tissue-related (b) or capsule-related (c) SHG images based on k-means clustering performed on the PCs PC1-PC5. The images separated for the analysis are colored yellow in (b) and (c). Reprinted from [***Paper D***]

Either complete or reduced feature vectors (via RFECV-LinearSVC/MIFS) of tissue-related SHG images were used for classifier optimization.

The predictive performance results of all ML classifiers are summarized in Table S5 and Figures S1a1-18 in ***Paper D***. The highest accuracy values were achieved with a 90/10 (training/validation) data split and classifier

optimization using the complete feature vectors (*Table 4-8* "I. Tissue-related", *Figure 4.23*). All ensemble ML models (RF, XGBoost, and LightGBM) show signs of overfitting, suggesting data leakage, which led to overly optimistic results on the training dataset but poor performance on the validation dataset. Although the LR classifier had relatively satisfactory accuracy, its other metrics were significantly worse than those of the other models. MLP and C-SVC demonstrated the best performance on the validation set (*Figure 4.24a - c*), with C-SVC surpassing MLP in all metrics. The lower recall and F1-score values for MLP indicate a higher rate of false negatives (PTC predicted as FTC) and false positives (FTC predicted as PTC) as compared to C-SVC.

Table 4-8 Numerical estimation of the optimized model performance (based on maximized accuracy) obtained for data split training/validation 90/10 for MLP and C-SVC models. Reprinted from [***Paper D***].

Label correction approach	ML model	Accuracy (validation), %	Accuracy (train), %	Recall	Precision	F1	AUC	Accuracy (FTC test), %	Accuracy (PTC test), %	Accuracy (PTC* test), %	Comment
I. Tissue-related	MLP	75.88	76.77	0.588	0.781	0.671	0.850	54.57	56.82	62.40	+++
	MLP*	78.00↑	79.00↑	0.687	0.763	0.723	0.862	63.94	44.69↓	49.91	↑++↓
	C-SVC	81.71	87.09	0.767	0.789	0.778	0.881	72.13	45.09	48.84	++
	C-SVC*	80.31↓	82.85↓	0.736	0.780	0.757	0.870	70.76↓	42.01↓	46.13	↓
II. Capsule -related	MLP	81.94	83.95	0.745	0.798	0.771	0.898	64.67	38.05	40.91	++
	MLP*	80.00	82.15	0.707	0.781	0.742	0.881	67.89	44.93	46.61	++↑
	C-SVC	82.07	88.36	0.770	0.785	0.778	0.901	71.35	43.28	46.19	++
	C-SVC*	82.20	87.61	0.748	0.802	0.774	0.886	65.69	52.74	56.16	+++↑

Footnote: * – indicates that feature selection was performed prior to the optimization of the hyperparameter configurations of the used classifiers; good accuracy validation/training, good Recall/Precision/F1/AUC, poor for real test set; +++ good accuracy validation/training, "classified" for real test set. The arrow (↑) indicates the improvement of the model performance and (↓) indicates the decrease in the model performance after the removal of redundant and irrelevant features.

On the unknown test set, MLP demonstrates correct classification rates slightly above 50% (*Figure 4.24e*), while C-SVC exhibited low discriminative power for PTC but performed well in distinguishing FTC (*Figure 4.24d*). Visual inspection of the classified PTC images revealed that the normal tissue surrounding the circular PTC capsule was frequently misclassified as FTC, resulting in false negatives. Restricting the analysis area closer to the PTC capsule improved the classification performance, increasing the proportion of true positives to 48.84% for C-SVC and 62.40% for MLP.

Two distinct regions within the collagen capsule of PTC were consistently classified as FTC by both C-SVC and MLP. Optical and SHG images of the capsule suggest that these areas correspond to calcifications (*Figure 4.13*, *Figure 4.18* and *Figure 4.22*).

Calcifications are more frequently observed in PTCs than in FTCs and are generally accepted as a reliable indicator of PTC [202]. Calcifications of sizes less than 1 mm are called microcalcifications and can be referred to as stromal calcification, bone formation, or psammoma bodies, whereas

calcifications > 1 mm are macrocalcifications. Calcifications are believed to form due to necrosis, haemorrhage and subsequent fibrosis within the tumour. Collagen I serves as a scaffold for mineralisation – deposition of mineral salts [220], such as calcium carbonate phosphates [221], calcium hydroxyapatite [202], or other calcium compounds within the fibrous extracellular matrix. Mineral deposits do not generate SHG [222,223], indicating that SHG signal detected in PTC samples originate from fibrillar collagens which favoured the formation of macrocalcifications.

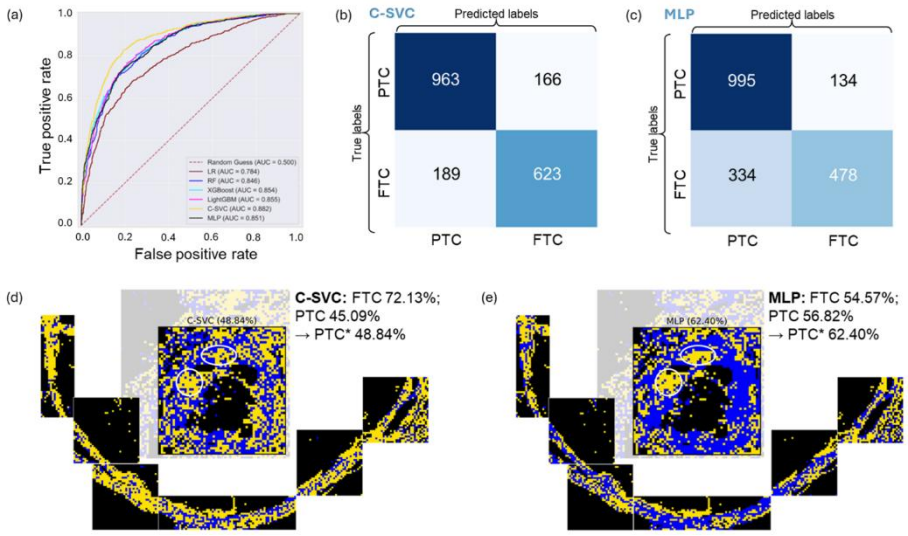


Figure 4.24 Performance of the best developed ML models optimized on the basis of accuracy: (a) ROC curves of all ML models, (b) confusion matrix for C-SVC; (c) confusion matrix for MLP; (d) C-SCV classification on new dataset (test set); (e) MLP classification on new dataset (test set). Blue colored tile images mark images classified as PTC, yellow – classified as FTC. The percentage for PTC indicates the proportion of correctly predicted PTC tiles in the PTC sample that includes the surrounding tissue. An asterisk marks the percentage of correctly predicted PTC tiles in the PTC sample excluding the surrounding tissue. White circles mark the areas with calcifications. Reprinted from [***Paper D***].

The areas of the PTC capsule which are associated with calcifications (*Figure 4.13*, *Figure 4.18* and *Figure 4.22*) were classified as FTC (*Figure 4.24*) indicating that the collagen texture features in calcifications resemble those of either normal tissue or FTC. This introduces another potential source of data errors, categorized as mislabelling: despite being a characteristic feature of PTC, calcifications possess texture features that align with other targets. One possible solution is to create an additional class for calcifications and perform multi-class classification. However, due to limited sample size, attempting to separate the dataset this way would result in unbalanced data.

Both C-SVC and MLP classify the samples based on a combination of approximately half of all features, as indicated by the PIA of C-SVC and MLP (*Figure 4.25*). PIA reveals how the model's accuracy is disrupted when one feature is randomly changed, providing insight into the model's reliance on specific features [114]. C-SVC uses more texture features for training than MLP and significantly more than the tree models. Consequently, C-SVC may be better suited to reveal hidden relationships between the texture features of SHG images within PTC and FTC samples.

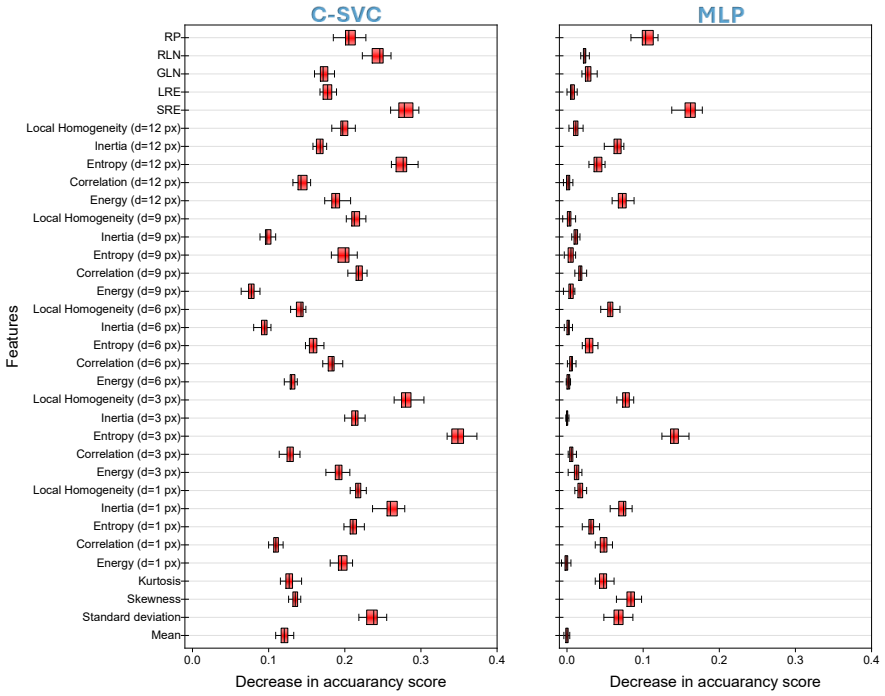


Figure 4.25 PIA of C-SVC and MLP models. Reprinted from [***Paper D***].

Adding a feature selection step prior to the optimization and training of the classifiers, results in the removal of both redundant features (from 2 to 7 features) and irrelevant features to the target (from 3 to 5 features), indicating that data is corrupted by both feature and label noise (*Table 4-7*). However, the performance of the classifiers decreases in all cases, indicating that there is another, more significant source of noise beyond feature noise, e.g. label noise (Table S5 and Figures S1b1–18 in ***Paper D***). The proportion of mislabelled data is likely considerable, and the classification models mainly fail in handling this mislabelled data.

This noise may originate from non-capsular collagen structures in the tissue surrounding the PTC and FTC nodules, as well as from calcifications. Removing normal tissue surrounding the neoplasm from the analysis is considered as a possible approach to increase the classification accuracy. The perinodular tissue may contain signatures of tumour progression, such as

collagen network remodelling induced by increased MMP9 secretion from macrophages recruited to the tumour growth sites [224], providing additional information for classification models. However, since these changes may be similar in both carcinoma types [225], labelling surrounding tissues as PTC or FTC could lead to mislabelling and misclassification.

4.4.2 Capsule-related SHG images of PTC and FTC

To minimize data overlap between PTC and FTC samples caused by surrounding tissue, only SHG images of nodule capsules were selected. To automate label noise reduction, capsule images separation was performed using an unsupervised ML approach [177]: PCA was performed to the SHG dataset, followed by binary *k*-means clustering on the acquired PCs. A typical result is shown in *Figure 4.23c*. While neither manual labelling nor this method ensures perfect capsule separation, *k*-means clustering based on feature variance differences provides a more objective segmentation than visual inspection. Further analysis was performed on the capsule-related SHG images.

Filtering out non-capsule SHG images significantly reduced the training/validation datasets and caused slight but manageable class imbalance (*Table 3-6*). Compared to the tissue-related SHG image dataset, RFECV-LinearSVC significantly reduced the number of features, while MIFS removed none. The absence of features removed by MIFS indicates that all features selected with RFECV-LinearSVC were relevant for distinguishing PTC and FTC capsules.

While feature selection can improve model performance, a significant dataset reduction may negatively impact classifier performance [226]. Thus, classifier performance results are presented below for both the full and reduced feature set.

The predictive performance of all ML models trained on the full feature set is summarised in Table S6 and Figures S2a-1–18 in ***Paper D***. Unlike the all-tissue approach, accuracy remains consistent across 70/30, 80/20 or 90/10 (train/validation) splits, with two best (MLP and C-SVC) shown in *Table 4-8* "II. Capsule-related". Overall, models perform better than those trained on all tissue-related data (*Figure 4.26a*). However, ensemble models (RF, XGBoost and LightGBM) remain overfitted, showing overly optimistic results on the training dataset but failing on the validation dataset. RF achieves satisfactory accuracy, but low recall and F-1 score. Feature importance analysis (Figure S2a-13 in ***Paper D***) shows that low coefficients were assigned to most features, leading to the classification of PTC and FTC capsules as identical in the unknown dataset.

For MLP and C-SVC, accuracy on the validation set improves to 81.94% and 82.07% respectively (*Table 4-8* "II. Capsule-related", *Figure 4.26*). In addition, the recall and F-1 score metrics increase for the MLP model, indicating enhanced performance when trained on capsule-related datasets.

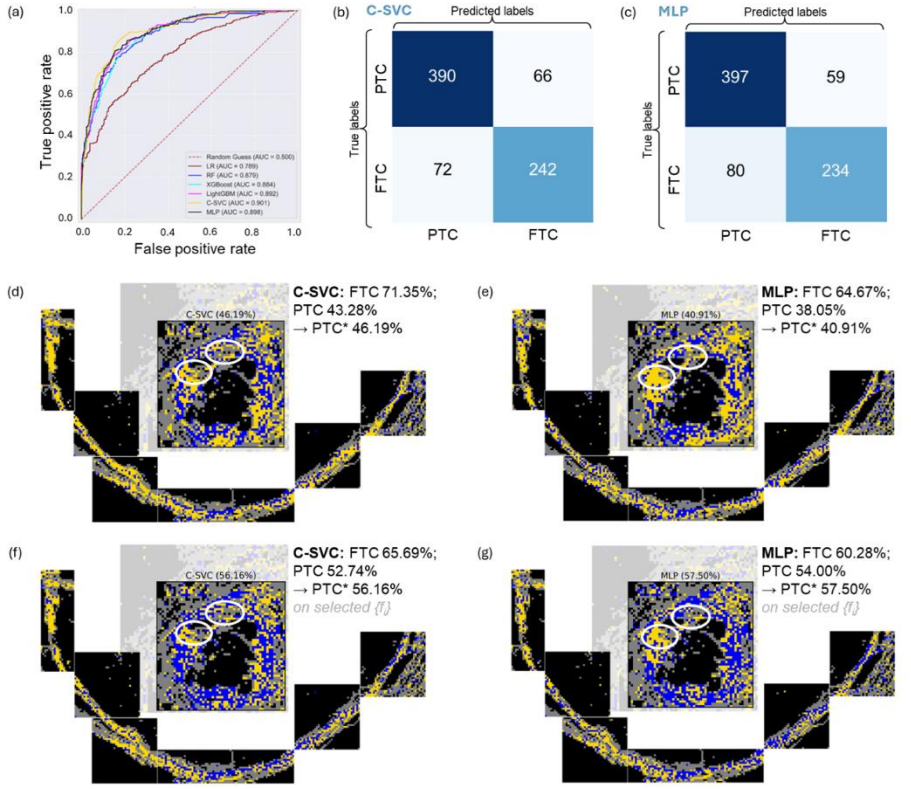


Figure 4.26 Performance of the ML models developed on capsule-related datasets for a split of (90/10): (a) ROC curves of all ML models, (b) confusion matrix for C-SVC; (c) confusion matrix for MLP; (d) C-SVC classification performed on the new data set (test set); (e) MLP classification performed on the new data set (test set); (f) C-SCV (trained on a reduced set of features) classification performed on the new data set (test set); (g) MLP (trained on a reduced set of features, 70/30) classification performed on the new data set (test set). Blue colored tile label images classified as PTC, yellow – classified as FTC. Percentage for PTC indicates the portion of correctly predicted PTC tiles in the PTC sample, which includes surrounding tissue. Asterisk marks the portion of correctly predicted PTC tiles in the PTC sample excluding surrounding tissue. White circles mark the areas of calcifications. Reprinted from [***Paper D***].

Both MLP and C-SVC not only showed improved performance on the validation set, but also correctly classified FTC samples in the global test dataset (*Figure 4.26d, e*). The C-SVC model's classification results for PTC were similar to those from the all-tissue-related datasets (*Figure 4.26d*), suggesting that C-SVC is able to manage the overlap in all tissue-related datasets and focus solely on the patterns associated with the capsular collagen. In contrast, the MLP classification accuracy for PTC decreases when trained

with the capsule-related datasets compared to the all tissue-related datasets (*Figure 4.26*d).

PIA reveals that C-SVC is better at identifying feature relevance and addressing feature multicollinearity (*Figure 4.27*). While the importance of individual features changed between tissue-related and capsule-related datasets, the model's overall performance remained stable. In contrast, MLP struggles with the capsule-related datasets and/or suffered from high feature multicollinearity. The very small fluctuations in MLP accuracy in response to random changes in feature values revealed by the PIA may indirectly indicate this issue (*Figure 4.27*).

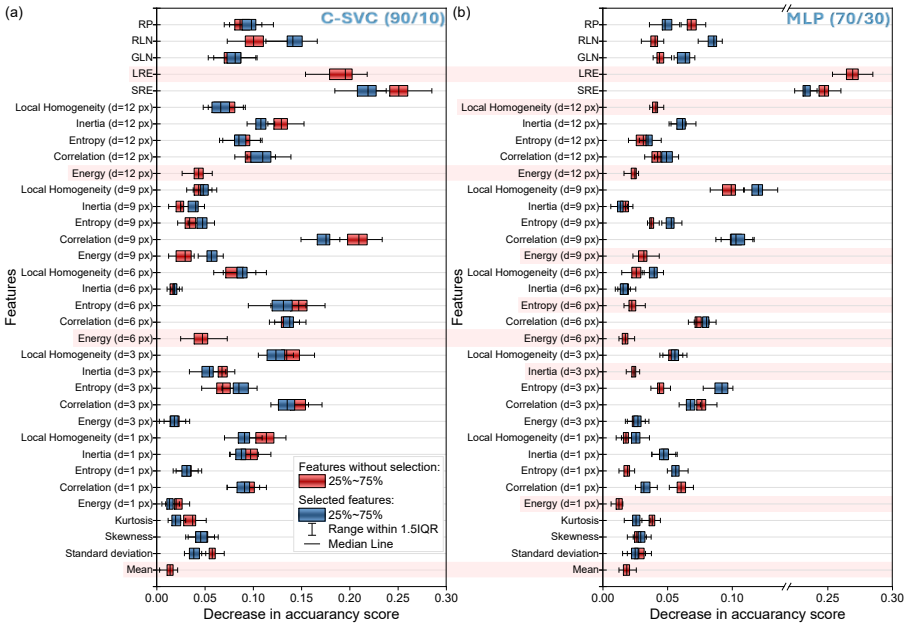


Figure 4.27 PIA of the performance of (a) C-SVC (90/10) and (b) MLP (70/30) classifiers, optimised based on the complete (red) and reduced (blue) feature sets. Reprinted from [***Paper D***].

Feature selection with RFECV-LinearSVC slightly decreased accuracy for all classifiers, but significantly improved performance on the unknown global test dataset for all splits (*Table 4-7 Features* selected by RFECV-LinearSVC and MIFS within each label correction approach. "II. Capsule-related", Table S6 in ***Paper D***, rows marked with asterisks). The only exception is the LR classifier, which showed no change in performance between full and reduced feature sets, suggesting that feature redundancy is not the reason for the model's failure. The tree classifiers and C-SVC benefited most from feature selection, with C-SVC reaching 65.69% and 56.16% accuracy for FTC and PTC, respectively, on the training/validation split, outperforming all previous approaches (*Figure 4.26f*). MLP, which appeared

to be sensitive to the size of the training dataset, showed improved classification, reaching 57.50% and 60.28% for PTC and FTC, respectively, on the global test set (*Figure 4.26g*). The classification results for all classifiers are summarized in Figures S2b-1-18 in ***Paper D***.

PIA of C-SVC and MLP, trained on the reduced feature sets (*Figure 4.27*), showed that the features important to C-SVC before feature selection, remained significant after feature selection. Increased contributions from HOS (*RP*, *RLN*, *GLN*) and SOS (d = 6 px, 9 px) more likely explains the improvement in C-SVC performance. The exclusion of features E_{12}, E_6 and μ_1 by RFECV-LinearSVC had minimal effect. Although *LRE* contribution was relatively high in all previous approaches, its removal with RFECV-LinearSVC did not affect the classification performance.

For MLP, removing redundant features (L_{12}, $E_{1,6,9,12}$, I_3, H_6 and μ_1) increased the contribution of almost all remained SOS parameters and *SRE*. Both C-SVC and MLP classifiers focused on HOS (*RP*, *RLN*, *GLN* and *SRE*) and SOS, suggesting that these features capture the main structural differences in PTC and FTC capsular collagen networks, while FOS parameters provided no valuable information.

To sum up, ML training and testing using capsular collagen-related SHG images selected by two-step binary clustering improved the accuracy of MLP and C-SVC estimated on the validation set, with accuracies reaching 81.94% and 82.07%, respectively. Feature selection, performed prior to classifier optimization (excluding LR), significantly improved performance on the global test set.

However, accurate PTC classification remains challenging, despite improvements with label and feature denoising. This could be due to the high heterogeneity of collagen features along the PTC capsule and similarity between certain PTC and FTC capsule segments, leading to higher accuracy in identifying FTC and lower accuracy for PTC (*Figure 4.26d-g*). While FTC tends to have a more uniform capsular structure, some PTC capsule areas may share structural similarities with FTC, possibly due to common stromal response pathways or similar collagen alignment, density, or biochemical properties. Histopathological studies have shown that thyroid tumour capsules are heterogenous, with variations in collagen composition and structure influenced by tumour subtype, growth patterns, and interaction with host tissue. Further supporting these observations, advanced imaging techniques, such as SHG microscopy bundled with AI methods for image analysis, can provide quantitative insights into these variations.

4.4.3 Multiclass classification based on the specific ratio of clusters describing PTC and FTC

The tissue surrounding the nodules in both carcinoma types does not provide relevant information for classification based on SHG image texture features. However, texture features like *LRE*, E_{12}, E_6, μ_1, removed during feature selection when only capsule collagen was considered, likely explain

the differences between the adjacent tissue and the nodules. This suggests that adjacent tissue could form an additional class, helping address mislabelling in the tissue-related approach.

Segmentation based on intensity and texture features was performed as described in [177] to prove the similarity of perinodular tissue. PCA and multi-class *k*-means clustering results (*Figure 4.28*) show that while capsular collagen is heterogeneous in both PTC and FTC, adjacent tissue is separated in one class (coloured magenta, *Figure 4.28b-c*).

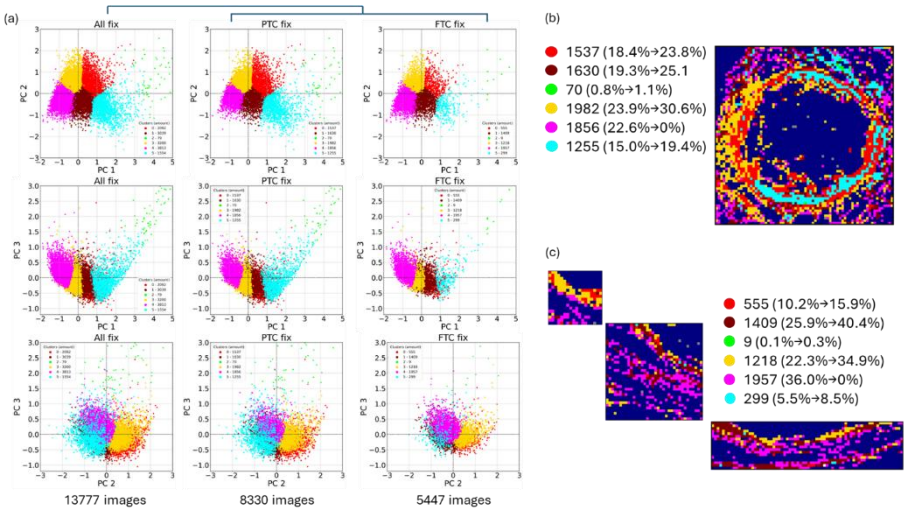


Figure 4.28 PCA analysis of the feature datasets of PTC and FTC samples and their clustering using k-means: (a) score plots of PC1 vs. PC2, PC3 vs. PC1 and PC3 vs. PC2 for all data and separately for PTC and FTC; (b) cluster map of random PTC and FTC samples. The numbers indicate the percentage of each cluster in the corresponding data set. The second number in parentheses is the percentage of each cluster within the capsule. The 'magenta' cluster was assigned to normal collagen surrounding normal tissue follicles and was therefore excluded from the 'capsular collagen' class and added to a separate class combining glass and normal tissue images present in both PTC and FTC samples. Reprinted from [***Paper D***].

Both carcinoma capsules consist of the same clusters, which complicates classification of PTC and FTC capsules even when adjacent tissue is excluded from the analysis (e.g., *Figure 4.26d-g*). Despite shared cluster composition, the cluster ratios differ between carcinoma types. PCA score plots representing all data (*Figure 4.28a*) and examples of segmented SHG scans (*Figure 4.28b-c*) show that FTC capsules are dominated by brownish and yellow clusters, while PTC capsules are more heterogeneous. The former likely explains the better classification of FTC by C-SVC classifier in previous approaches, while the latter probably led to a higher error rate for PTC.

The higher heterogeneity of collagen capsules surrounding PTC nodules, compared to FTC nodules may be due to differences in growth rates of the

nodules. PTC tends to grow more slowly, while FTC exhibits increased aggressiveness and a higher tendency for metastasis [227]. FTC often presents with larger nodules at diagnosis [228], contributing to its faster growth. Furthermore, FTC has a tendency for hematogenous spread, contrasting with the lymphatic spread more commonly associated with PTC [227], which influences clinical management and prognosis.

These differences may enhance classification results, as both the clusters and their ratios describe the capsules of PTC and FTC nodules.

Prior to the stratified 10-fold cross-validation for model optimization, the ratio of clusters, which was identified for the whole train dataset via *k*-means, was fixed. SHG images of adjacent tissue and glass were added to a "non-target" class to avoid preprocessing steps aimed at the removal of SHG images which are irrelevant to the target and thus that could introduce label noise. The "FTC", "PTC" and "non-target" classes were balanced prior to classifier optimization, though some data disproportion remained. This reduced dataset size compared to both all-tissue and capsular-related approaches could affect the classifier performance *Table 3-6*.

Multi-class classification, which includes all SHG images and tends to correct the mislabelling by adding a "non-target class", also lead to a slight reduction in redundant features, with all remaining features being relevant for the target.

The results of classifier optimization for all data splits are shown in Table S7 and Figures S3a1-18 in ***Paper D***. The 70/30 split results for MLP and C-SVC shown in *Table 4-9*, *Figure 4.29* and *Figure 4.30* are more representative, since this split includes more data in the validation set and preserves cluster ratios.

Table 4-9 Numerical estimation of the optimized model performance (based on maximized accuracy) obtained for data split training/validation 70/30 for datasets considering the ratio of clusters in PTC and FTC samples for MLP and C-SVC models. Reprinted from [***Paper D***].

ML model	Accuracy (validation), %	Accuracy (train), %	Precision (macro)	Recall (micro)	F1 (weighted)	Accuracy (FTC test), %	Accuracy (PTC test), %	Comment
MLP	81.76	83.16	0.816	0.817	0.811	42.98	63.73	++
MLP*	81.32	83.98	0.810	0.813	0.814	65.12	50.22	+++
C-SVC	84.73	89.30	0.843	0.847	0.847	63.70	52.23	+++
C-SVC*	84.80	89.50	0.844	0.848	0.847	68.65	51.66	+++

Confusion matrices (*Figure 4.29*) show improved performance of all ensemble classifiers (RF, XGBoost, LightGBM) when considering the cluster ratios within the capsules of each type, though RF and LightGBM are still overfitted (Table S7 in ***Paper D***). LR performs well with non-target data but fails in classification of PTC and FTC capsules. Although it successfully separates non-target data from capsular collagen, it makes many false

positives. MLP classifies PTC better than FTC (*Figure 4.29*), while C-SVC outperforms other classifiers achieving 84.73% accuracy on the validation set (*Figure 4.29*, *Table 4-9*).

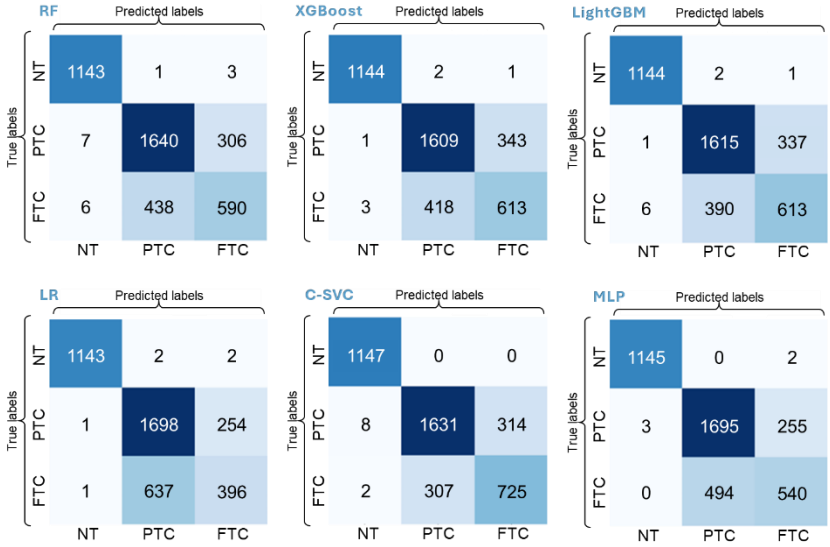


Figure 4.29 Confusion matrices calculated for all classifiers developed with a 70/30 split (optimized by maximum accuracy) for datasets considering the ratio of clusters in PTC and FTC samples. Reprinted from [***Paper D***].

The C-SVC classifier performs better on the unknown test set compared to the all tissue and capsule-related approaches without feature selection (*Figure 4.30a*), but worse than the capsule-related approach with feature selection (*Figure 4.26f*). On the contrary, the performance of the MLP has deteriorated.

Feature selection in a multi-class label correction approach resulted in removal of few features and had little impact on the classification performance of practically all classifiers (Table S7, Figures S3b1-18 in ***Paper D***). However, C-SVC generalization performance was significantly improved and correct predictions increased up to 68.65% and 55.26% for FTC and PTC, respectively (*Figure 4.30b*).

Areas of calcification are still misclassified as FTC, as their texture and intensity features resemble those of FTC capsules rather than PTC capsules. To address this, a "calcifications" class could be added, but due to limited data, this isn't feasible at this stage.

PIA shows that C-SVC (*Figure 4.30c*) relies on the full feature set, with high μ_1 and σ (FOS) contributions distinguishing non-target class from PTC and FTC capsules. Low PIA scores for μ_1 and σ in the capsule-related approach and higher PIA scores in other approaches support this conclusion. PIA performed for the best C-SVC classifiers in all three approaches, suggest that SOS parameters calculated based on GLCM with steps d = 3-12 px cover

the main differences between the PTC and FTC capsules, although they are not completely discriminative.

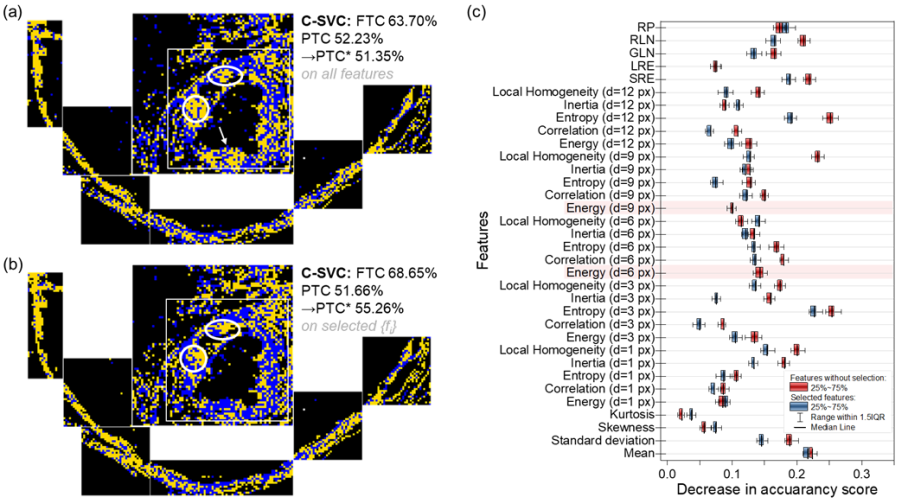


Figure 4.30 Performance of C-SVC classifier optimized for datasets considering the ratio of clusters in PTC and FTC samples: (a) classification of a global test set based on the complete set of features; (b) classification of a global test set based on the reduced set of features; (c) PIA. Blue coloured pixels in (a) and (b) mark images classified as PTC, yellow – classified as FTC. White circles and arrows mark the areas of calcifications. Training/validation data split was 70/30. Reprinted from [***Paper D***].

While multi-class classification did not significantly improve test accuracy for the global test data, unsupervised ML segmentation highlighted differences in PTC and FTC capsules and adjacent tissue, explaining classifier performance variations. Adjacent tissue lacks detectable signatures of PTC or FTC progression and can either be removed from the analysis by binary *k*-means (II approach) or considered as a separate class in multi-class classification (III approach). Similar heterogenous collagen patterns in PTC and FTC capsules complicate classification, while calcifications in PTC, which differ in texture features from the PTC capsule are misclassified as FTC by all classifiers. The lower heterogeneity of FTC as compared to PTC capsules allows C-SVC to distinguish between PTC and FTC, while other classifiers struggle with overfitting or data size reduction (MLP). PIA shows PTC and FTC capsule differences are mainly described by SOS (GLCM, d = 3-12 px) and HOS (excluding *LRE*), while FOS features only distinguish capsules from adjacent tissue.

This was summarized in the fourth statement of the thesis: *A supervised machine learning model C-SVC enables differential diagnosis of papillary and follicular thyroid carcinomas based on SHG imaging, with an accuracy of 84.73%.*

5 CONCLUSIONS

1. Wide-field SHG microscopy combined with quantitative image analysis is a robust and label-free method to assess fibrotic remodeling in PAH. Time-dependent changes in collagen content and morphology were detected in lung tissue from MCT-treated rats. FOS analysis showed progressive collagen accumulation, while SOS analysis indicated a densification of the perivascular collagen network with subsequent expansion into the alveolar region. FFT analysis also showed a dynamic modulation of fiber orientation characterized by initial alignment followed by disorganization late in the disease. These results emphasize the potential of SHG-based approaches for non-destructive assessment of fibrosis in PAH and related lung diseases.

2. The application of wide-field PSHG microscopy to images of whole sections of thyroid nodules on histologic slides allowed the extraction of quantitative parameters describing collagen orientation and ultrastructure within the nodule capsule. Using a cylindrical collagen model, regions of capsular invasion can be effectively distinguished from non-invasive areas by statistical analysis and unsupervised ML. This method allows objective and reproducible assessment of collagen ultrastructure alterations in capsular invasion in thyroid neoplasms and can serve as a basis for the development of automated diagnostic tools in thyroid pathology.

3. A method combining wide-field SHG microscopy, texture analysis and unsupervised machine learning is developed to quantitatively assess collagen capsule structure in PTC. The two-step *k*-means clustering revealed pronounced heterogeneity within the capsule and delineated regions with distinct structural features corresponding to intact, invaded and potentially pre-invasive sites. In particular, the approach identified areas of subtle microinvasion that were not detected on initial histopathologic examination. The ability of the proposed unsupervised ML method to detect such regions highlights its potential as a complementary diagnostic tool to improve accuracy, reduce observer variability and support the development of automated classification systems.

4. The supervised ML algorithms applied to SHG-derived intensity and texture features enable efficient differential diagnosis of PTC and FTC. Although the classification task is complicated by feature redundancy and labeling inaccuracies, caused by the inclusion of adjacent tissue, calcifications, and intertumoral capsule similarity, significant improvement of the classifier performance was achieved through specific data processing and application of unsupervised segmentation to exclude non-informative regions. The C-SVC classifier achieved the highest validation accuracy, indicating its robustness to noise and bias.

6 BIBLIOGRAPHY

[1] R.L. Siegel, T.B. Kratzer, A.N. Giaquinto, H. Sung, A. Jemal, CA A Cancer J Clinicians 75 (2025) 10–45.
[2] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R.L. Siegel, I. Soerjomataram, A. Jemal, CA A Cancer J Clinicians 74 (2024) 229–263.
[3] S. Prabhakaran, C. Yapp, G.J. Baker, J. Beyer, Y.H. Chang, A.L. Creason, R. Krueger, J. Muhlich, N.H. Patterson, K. Sidak, D. Sudar, A.J. Taylor, L. Ternes, J. Troidl, X. Yubin, A. Sokolov, D.R. Tyson, Molecular Oncology (2025) 1878–0261.13783.
[4] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, CA A Cancer J Clinicians 71 (2021) 209–249.
[5] L. Boucai, M. Zafereo, M.E. Cabanillas, JAMA 331 (2024) 425.
[6] M. Nishino, J. Jacob, Seminars in Diagnostic Pathology 37 (2020) 219–227.
[7] Y. Zhu, Y. Li, C.K. Jung, D.E. Song, J.-F. Hang, Z. Liu, D. Jain, C.-R. Lai, M. Hirokawa, K. Kakudo, A. Bychkov, Endocr Pathol 31 (2020) 132–140.
[8] F. Basolo, E. Macerola, A.M. Poma, L. Torregrossa, Endocrine 80 (2023) 470–476.
[9] Z. Liu, W. Zeng, L. Huang, Z. Wang, M. Wang, L. Zhou, D. Chen, H. Feng, W. Zhou, L. Guo, Am J Cancer Res 8 (2018) 1440–1448.
[10] E.M. Aboelnaga, R.A. Ahmed, Cancer Biol Med 12 (2015) 53–59.
[11] O. Gimm, H. Dralle, in: Surgical Treatment: Evidence-Based and Problem-Oriented, Zuckschwerdt, 2001.
[12] P.H. Dedhia, M.C. Saucke, K.L. Long, G.M. Doherty, S.C. Pitt, JAMA Netw Open 5 (2022) e228722.
[13] P.A. Singer, Arch Intern Med 156 (1996) 2165.
[14] G. Tini, M. Sarocchi, G. Tocci, E. Arboscello, G. Ghigliotti, G. Novo, C. Brunelli, D. Lenihan, M. Volpe, P. Spallarossa, International Journal of Cardiology 281 (2019) 133–139.
[15] M. Gürdoğan, Anatol J Cardiol (2023) 299–307.
[16] H.W. Farber, J. Loscalzo, New England Journal of Medicine 351 (2004) 1655–1665.
[17] O. Boucherat, G. Vitry, I. Trinh, R. Paulin, S. Provencher, S. Bonnet, Pulm. Circ. 7 (2017) 285–299.
[18] P.J. Campagnola, C. -Y. Dong, Laser & Photonics Reviews 5 (2011) 13–26.
[19] X. Chen, O. Nadiarynkh, S. Plotnikov, P.J. Campagnola, Nat Protoc 7 (2012) 654–669.
[20] L.G. Eftimie, R.R. Glogojeanu, A. Tejaswee, P. Gheorghita, S.G. Stanciu, A. Chirila, G.A. Stanciu, A. Paul, R. Hristu, Sci Rep 12 (2022) 21636.

[21] R. Hristu, L.G. Eftimie, S.G. Stanciu, D.E. Tranca, B. Paun, M. Sajin, G.A. Stanciu, Biomed. Opt. Express 9 (2018) 3923.
[22] A. Dementjev, R. Rudys, R. Karpicz, D. Rutkauskas, Lith. J. Phys. 60 (2020).
[23] C. Macias-Romero, M.E.P. Didier, P. Jourdain, P. Marquet, P. Magistretti, O.B. Tarun, V. Zubkovs, A. Radenovic, S. Roke, Opt. Express 22 (2014) 31102.
[24] F. Radaelli, L. D'Alfonso, M. Collini, F. Mingozzi, L. Marongiu, F. Granucci, I. Zanoni, G. Chirico, L. Sironi, Sci Rep 7 (2017) 17468.
[25] R. Cicchi, N. Vogler, D. Kapsokalyvas, B. Dietzek, J. Popp, F.S. Pavone, J. Biophoton. 6 (2013) 129–142.
[26] R. Hristu, S.G. Stanciu, D.E. Tranca, G.A. Stanciu, Journal of Biophotonics 10 (2017) 1171–1179.
[27] X. Chen, Z. Huang, G. Xi, Y. Chen, D. Lin, J. Wang, Z. Li, L. Sun, J. Chen, R. Chen, in: Wuhan, China, 2012, p. 83290H.
[28] D. Tokarz, R. Cisek, A. Joseph, S.L. Asa, B.C. Wilson, V. Barzda, Laboratory Investigation 100 (2020) 1280–1287.
[29] S.G. Stanciu, R. Hristu, G.A. Stanciu, D.E. Tranca, L. Eftimie, A. Dumitru, M. Costache, H.A. Stenmark, H. Manders, A. Cherian, M. Tark-Dame, E.M.M. Manders, Proc. Natl. Acad. Sci. U.S.A. 119 (2022) e2214662119.
[30] I. Gregor, M. Spiecker, R. Petrovsky, J. Großhans, R. Ros, J. Enderlein, Nat Methods 14 (2017) 1087–1089.
[31] R. Amor, A. McDonald, J. Trägårdh, G. Robb, L. Wilson, N.Z. Abdul Rahman, J. Dempster, W.B. Amos, T.J. Bushell, G. McConnell, PLoS ONE 11 (2016) e0147115.
[32] A. Dementjev, R. Rudys, R. Karpicz, D. Rutkauskas, Lithuanian Journal of Physics 60 (2020).
[33] K. Mirsanaye, L. Uribe Castaño, Y. Kamaliddin, A. Golaraei, R. Augulis, L. Kontenis, S.J. Done, E. Žurauskas, V. Stambolic, B.C. Wilson, V. Barzda, Sci Rep 12 (2022) 10290.
[34] H. Zhao, R. Cisek, A. Karunendiran, D. Tokarz, B.A. Stewart, V. Barzda, Biomed. Opt. Express 10 (2019) 5130.
[35] L.B. Mostaço-Guidolin, A.C.-T. Ko, F. Wang, B. Xiang, M. Hewko, G. Tian, A. Major, M. Shiomi, M.G. Sowa, Sci Rep 3 (2013) 2190.
[36] R.M. Haralick, K. Shanmugam, I. Dinstein, IEEE Transactions on Systems, Man, and Cybernetics SMC-3 (1973) 610–621.
[37] M. Pomeroy, H. Lu, P.J. Pickhardt, Z. Liang, in: Medical Imaging 2018: Computer-Aided Diagnosis, SPIE, 2018, pp. 507–513.
[38] W. Chen, M.L. Giger, H. Li, U. Bick, G.M. Newstead, Magnetic Resonance in Medicine 58 (2007) 562–571.
[39] I. Aganj, M.G. Harisinghani, R. Weissleder, B. Fischl, Sci Rep 8 (2018) 13012.
[40] L. Liu, A.I. Aviles-Rivero, C.-B. Schönlieb, IEEE Trans. Neural Netw. Learning Syst. 36 (2025) 147–159.

[41] B. Zhang, H. Shi, H. Wang, JMDH Volume 16 (2023) 1779–1791.
[42] A. Nagpal, V. Singh, Procedia Computer Science 132 (2018) 244–252.
[43] O.O. Oladimeji, H. Ayaz, I. McLoughlin, S. Unnikrishnan, Results in Engineering 24 (2024) 103071.
[44] E.E. Montelongo González, J.A. Reyes Ortiz, B.A. González Beltrán, CyS 24 (2020).
[45] J. Tang, S. Alelyani, H. Liu, in: Data Classification, Chapman and Hall/CRC, 2014.
[46] I. Freund, M. Deutsch, A. Sprecher, Biophysical Journal 50 (1986) 693–712.
[47] C. Sheppard, J. Gannaway, R. Kompfner, D. Walsh, IEEE J. Quantum Electron. 13 (1977) 912–912.
[48] F. Helmchen, W. Denk, Nat Methods 2 (2005) 932–940.
[49] W.E. Lamb, Phys. Rev. 134 (1964) A1429–A1450.
[50] A.C. Millard, P.J. Campagnola, W. Mohler, A. Lewis, L.M. Loew, in: Methods in Enzymology, Elsevier, 2003, pp. 47–69.
[51] J.N. Gannaway, C.J.R. Sheppard, Opt Quant Electron 10 (1978) 435–439.
[52] P. Campagnola, Anal. Chem. 83 (2011) 3224–3231.
[53] P.J. Campagnola, L.M. Loew, Nat Biotechnol 21 (2003) 1356–1360.
[54] S.V. Plotnikov, A.C. Millard, P.J. Campagnola, W.A. Mohler, Biophysical Journal 90 (2006) 693–703.
[55] A. Aghigh, S. Bancelin, M. Rivard, M. Pinsard, H. Ibrahim, F. Légaré, Biophys Rev 15 (2023) 43–70.
[56] E.E. Hoover, J.A. Squier, Nature Photon 7 (2013) 93–101.
[57] E. Hemmer, A. Benayas, F. Légaré, F. Vetrone, Nanoscale Horiz. 1 (2016) 168–184.
[58] J. Squier, M. Müller, Review of Scientific Instruments 72 (2001) 2855–2867.
[59] W. Denk, J.H. Strickler, W.W. Webb, Science 248 (1990) 73–76.
[60] J. Jiang, R. Yuste, Microsc Microanal 14 (2008) 526–531.
[61] D.S. James, P.J. Campagnola, BME Front 2021 (2021) 3973857.
[62] A. Pena, A. Fabre, D. Débarre, J. Marchal-Somme, B. Crestani, J. Martin, E. Beaurepaire, M. Schanne-Klein, Microscopy Res & Technique 70 (2007) 162–170.
[63] S.-Y. Chen, Z.-T. Su, D.-J. Lin, M.-X. Lee, M.-C. Chan, S. Das, F.-J. Kao, G.-Y. Zhuo, Results in Physics 28 (2021) 104653.
[64] H. Reis, The Journal of Chemical Physics 125 (2006).
[65] Y.R. Shen, Nature 337 (1989) 519–525.
[66] J.L. Oudar, D.S. Chemla, The Journal of Chemical Physics 66 (1977) 2664–2668.
[67] I.D.L. Albert, T.J. Marks, M.A. Ratner, J. Am. Chem. Soc. 120 (1998) 11174–11181.

[68] B. Li, J. Li, W. Gan, Y. Tan, Q. Yuan, Anal. Chem. 93 (2021) 14146–14152.
[69] P.N. Butcher, Elements of Nonlinear Optics, 1st ed, Cambridge University Press, West Nyack, 1990.
[70] F. Tiaho, G. Recher, D. Rouède, Opt. Express 15 (2007) 12286.
[71] T.L. Mazely, W.M. Hetherington, The Journal of Chemical Physics 86 (1987) 3640–3647.
[72] D. Sharoukhov, H. Lim, Journal of Modern Optics 63 (2016) 71–75.
[73] S. Ricard-Blum, Cold Spring Harbor Perspectives in Biology 3 (2011) a004978–a004978.
[74] A. Zoumi, A. Yeh, B.J. Tromberg, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 11014–11019.
[75] R.M. Williams, W.R. Zipfel, W.W. Webb, Biophysical Journal 88 (2005) 1377–1386.
[76] K. Tilbury, P.J. Campagnola, Perspect Medicin Chem 7 (2015) PMC.S13214.
[77] E. Brown, T. McKee, E. diTomaso, A. Pluen, B. Seed, Y. Boucher, R.K. Jain, Nat Med 9 (2003) 796–800.
[78] K. Tilbury, J. Hocker, B.L. Wen, N. Sandbo, V. Singh, P.J. Campagnola, J. Biomed. Opt 19 (2014) 086014.
[79] P.P. Provenzano, K.W. Eliceiri, J.M. Campbell, D.R. Inman, J.G. White, P.J. Keely, BMC Med 4 (2006) 38.
[80] P. Lu, V.M. Weaver, Z. Werb, Journal of Cell Biology 196 (2012) 395–406.
[81] K.R. Levental, H. Yu, L. Kass, J.N. Lakins, M. Egeblad, J.T. Erler, S.F.T. Fong, K. Csiszar, A. Giaccia, W. Weninger, M. Yamauchi, D.L. Gasser, V.M. Weaver, Cell 139 (2009) 891–906.
[82] M.W. Pickup, J.K. Mouw, V.M. Weaver, EMBO Reports 15 (2014) 1243–1253.
[83] K. Kessenbrock, V. Plaks, Z. Werb, Cell 141 (2010) 52–67.
[84] J. Chen, T.G. Diacovo, D.G. Grenache, S.A. Santoro, M.M. Zutter, The American Journal of Pathology 161 (2002) 337–344.
[85] E.A. Brett, M.A. Sauter, H.-G. Machens, D. Duscher, Cancer Metab 8 (2020) 14.
[86] G. Xi, W. Guo, D. Kang, J. Ma, F. Fu, L. Qiu, L. Zheng, J. He, N. Fang, J. Chen, J. Li, S. Zhuo, X. Liao, H. Tu, L. Li, Q. Zhang, C. Wang, S.A. Boppart, J. Chen, Theranostics 11 (2021) 3229–3243.
[87] K. Wolf, M. Te Lindert, M. Krause, S. Alexander, J. Te Riet, A.L. Willis, R.M. Hoffman, C.G. Figdor, S.J. Weiss, P. Friedl, Journal of Cell Biology 201 (2013) 1069–1084.
[88] M.M. Yallapu, K.S. Katti, D.R. Katti, S.R. Mishra, S. Khan, M. Jaggi, S.C. Chauhan, Medicinal Research Reviews 35 (2015) 198–223.
[89] J.T. Erler, K.L. Bennewith, T.R. Cox, G. Lang, D. Bird, A. Koong, Q.-T. Le, A.J. Giaccia, Cancer Cell 15 (2009) 35–44.

[90] I. Bourgot, I. Primac, T. Louis, A. Noël, E. Maquoi, Front. Oncol. 10 (2020) 1488.
[91] R.L. Benza, D.P. Miller, M. Gomberg-Maitland, R.P. Frantz, A.J. Foreman, C.S. Coffey, A. Frost, R.J. Barst, D.B. Badesch, C.G. Elliott, T.G. Liou, M.D. McGoon, Circulation 122 (2010) 164–172.
[92] L.A. Shimoda, S.S. Laurie, J Mol Med 91 (2013) 297–309.
[93] R. Xiao, Y. Su, T. Feng, M. Sun, B. Liu, J. Zhang, Y. Lu, J. Li, T. Wang, L. Zhu, Q. Hu, JAHA 6 (2017) e004865.
[94] C.D. Seib, J.A. Sosa, Endocrinology and Metabolism Clinics of North America 48 (2019) 23–35.
[95] A. Prete, P. Borges De Souza, S. Censi, M. Muzza, N. Nucci, M. Sponziello, Front. Endocrinol. 11 (2020) 102.
[96] J. Xia, Y. Shi, X. Chen, Sci Rep 14 (2024) 20977.
[97] O. Koperek, R. Asari, B. Niederle, K. Kaserer, Histopathology 58 (2011) 919–924.
[98] A. Avagliano, G. Fiume, C. Bellevicine, G. Troncone, A. Venuta, V. Acampora, S. De Lella, M.R. Ruocco, S. Masone, N. Velotti, P. Carotenuto, M. Mallardo, C. Caiazza, S. Montagnani, A. Arcucci, Cancers 14 (2022) 4172.
[99] Z. Yuan, B. Lin, C. Wang, Z. Yan, F. Yang, H. Su, Journal of Biological Chemistry 301 (2025) 108330.
[100] A. Golaraei, L. Kontenis, R. Cisek, D. Tokarz, S.J. Done, B.C. Wilson, V. Barzda, Biomed. Opt. Express 7 (2016) 4054.
[101] J. Adur, V.B. Pelegati, A.A. De Thomaz, M.O. Baratti, L.A.L.A. Andrade, H.F. Carvalho, F. Bottcher-Luiz, C.L. Cesar, Journal of Biophotonics 7 (2014) 37–48.
[102] R. Cicchi, D. Kapsokalyvas, V. De Giorgi, V. Maio, A. Van Wiechen, D. Massi, T. Lotti, F.S. Pavone, Journal of Biophotonics 3 (2010) 34–43.
[103] A. Golaraei, L.B. Mostaço-Guidolin, V. Raja, R. Navab, T. Wang, S. Sakashita, K. Yasufuku, M.-S. Tsao, B.C. Wilson, V. Barzda, Biomed. Opt. Express 11 (2020) 1851.
[104] J.W. Birk, M. Tadros, K. Moezardalan, O. Nadyarnykh, F. Forouhar, J. Anderson, P. Campagnola, Dig Dis Sci 59 (2014) 1529–1534.
[105] K.R. Campbell, R. Chaudhary, M. Montano, R.V. Iozzo, W.A. Bushman, P.J. Campagnola, J. Biomed. Opt. 24 (2019) 1.
[106] D. Tokarz, R. Cisek, A. Golaraei, S.L. Asa, V. Barzda, B.C. Wilson, Biomed. Opt. Express 6 (2015) 3475.
[107] A.-M. Pena, T. Baldeweck, E. Decencière, S. Koudoro, S. Victorin, E. Raynaud, B. Ngo, P. Bastien, S. Brizion, E. Tancrède-Bohin, Sci Rep 12 (2022) 14863.
[108] J. Li, P. Jiang, Q. An, G.-G. Wang, H.-F. Kong, Computers in Biology and Medicine 169 (2024) 107777.
[109] D. Ganguly, S. Chakraborty, M. Balitanas, T. Kim, in: T. Kim, A. Stoica, R.-S. Chang (Eds.), Security-Enriched Urban Computing and

Smart Grid, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 504–516.
[110] Xiaoou Tang, IEEE Trans. on Image Process. 7 (1998) 1602–1609.
[111] M.K. Ghalati, A. Nunes, H. Ferreira, P. Serranho, R. Bernardes, IEEE Rev. Biomed. Eng. 15 (2022) 222–246.
[112] A.K. Jain, Pattern Recognition Letters 31 (2010) 651–666.
[113] A. Anaya-Isaza, L. Mera-Jiménez, M. Zequera-Diaz, Informatics in Medicine Unlocked 26 (2021) 100723.
[114] L. Breiman, Machine Learning 45 (2001) 5–32.
[115] T. Chen, C. Guestrin, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, 2016, pp. 785–794.
[116] G. Ke, Q. Meng, T. Finely, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, in: Advances in Neural Information Processing Systems 30 (NIP 2017), 2017.
[117] C.-C. Chang, C.-J. Lin, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27.
[118] D.P. Kingma, J. Ba, (2014).
[119] D. Shen, G. Wu, H.-I. Suk, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.
[120] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Int J Comput Vis 128 (2020) 336–359.
[121] D.W. Wilson, H.J. Segall, L.C. Pan, M.W. Lamé, J.E. Estep, D. Morin, Critical Reviews in Toxicology 22 (1992) 307–325.
[122] A.G. Sapozhnikov, A.E. Dorosevich, Histological and Microscopic Technique: Manual (2000).
[123] R. Hristu, S.G. Stanciu, A. Dumitru, L.G. Eftimie, B. Paun, D.E. Tranca, P. Gheorghita, M. Costache, G.A. Stanciu, Sci Data 9 (2022) 376.
[124] R. Chetty, J Clin Pathol 57 (2004) 672.2-672.
[125] R. Hristu, S.G. Stanciu, A. Dumitru, B. Paun, I. Floroiu, M. Costache, G.A. Stanciu, Biomed. Opt. Express 12 (2021) 5829.
[126] IHC Antigen Retrieval Protocol | Abcam.Com (2023).
[127] E.I. Romijn, A. Finnøy, R. Kumar, M.B. Lilledahl, PLoS ONE 13 (2018) e0195027.
[128] M. Kociołek, M. Strzelecki, R. Obuchowicz, Computerized Medical Imaging and Graphics 81 (2020) 101716.
[129] N. Otsu, IEEE Trans. Syst., Man, Cybern. 9 (1979) 62–66.
[130] A. Fitzgibbon, M. Pilu, R.B. Fisher, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (1999) 476–480.
[131] P.P. van Zuijlen, H.J. de Vries, E.N. Lamme, J.E. Coppens, J. van Marle, R.W. Kreis, E. Middelkoop, The Journal of Pathology 198 (2002) 284–291.
[132] M. Thursby, Kisuck Yoo, B. Grossman, IEEE Trans. Aerosp. Electron. Syst. 31 (1995) 1341–1347.

[133] M. Lombardo, D. Merino, P. Loza-Alvarez, G. Lombardo, Biomed. Opt. Express 6 (2015) 2803.
[134] S. Psilodimitrakopoulos, E. Gavgiotaki, K. Melessanaki, V. Tsafas, G. Filippidis, Microsc Microanal 22 (2016) 1072–1083.
[135] L.B. Mostaço-Guidolin, A.C.-T. Ko, F. Wang, B. Xiang, M. Hewko, G. Tian, A. Major, M. Shiomi, M.G. Sowa, Sci Rep 3 (2013) 2190.
[136] M. Kröger, J. Schleusener, S. Jung, M.E. Darvin, Photonics 8 (2021) 404.
[137] L. Mostaço-Guidolin, N. Rosin, T.-L. Hackett, IJMS 18 (2017) 1772.
[138] R.M. Haralick, K. Shanmugam, I. Dinstein, IEEE Transactions on Systems, Man, and Cybernetics SMC-3 (1973) 610–621.
[139] R.W. Conners, C.A. Harlow, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-2 (1980) 204–222.
[140] N. Aggarwal, R. K. Agrawal, JSIP 03 (2012) 146–153.
[141] Yu.V. Kistenev, D.A. Vrazhnov, V.V. Nikolaev, E.A. Sandykova, N.A. Krivova, Biochemistry Moscow 84 (2019) 108–123.
[142] Z. Nejim, L. Navarro, C. Morin, P. Badel, Res. Biomed. Eng. 39 (2022) 273–295.
[143] A.A. Zeitoune, J.S. Luna, K. Sanchez Salas, L. Erbes, C.L. Cesar, L.A. Andrade, H.F. Carvahlo, F. Bottcher-Luiz, V.H. Casco, J. Adur, Cancer Inform 16 (2017) 117693511769016.
[144] J. Poole, Characterization and Classification of Fibrillar Collagen Networks Using Gray-Level Texture Analysis, Master of Applied Science, Carleton University, 2022.
[145] L.B. Mostaço-Guidolin, E.T. Osei, J. Ullah, S. Hajimohammadi, M. Fouadi, X. Li, V. Li, F. Shaheen, C.X. Yang, F. Chu, D.J. Cole, C.-A. Brandsma, I.H. Heijink, G.N. Maksym, D. Walker, T.-L. Hackett, Am J Respir Crit Care Med 200 (2019) 431–443.
[146] M.M. Galloway, Computer Graphics and Image Processing 4 (1975) 172–179.
[147] D.-H. Xu, A.S. Kurani, J.D. Furst, D.S. Raicu, in: ACTA Press, Marbella, Spain, 2004, p. 131.
[148] A. Humeau-Heurtier, IEEE Access 7 (2019) 8975–9000.
[149] C. Li, S. Wang, C. Li, Y. Yin, F. Feng, H. Fu, H. Wang, S. Chen, Front. Oncol. 12 (2022) 896593.
[150] R. Taylor, Journal of Diagnostic Medical Sonography 6 (1990) 35–39.
[151] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Mach Learn Res 12 (2011) 2825–2830.
[152] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, 2nd ed., Springer New York, New York, NY, 2021.

[153] D.S. Wilks, in: D.S. Wilks (Ed.), Statistical Methods in the Atmospheric Sciences (Fourth Edition), Fourth Edition, Elsevier, 2019, pp. 617–668.
[154] W.K. Härdle, L. Simar, Applied Multivariate Statistical Analysis, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.
[155] S. Wold, K. Esbensen, P. Geladi, Chemom Intell Lab Syst 2 (1987) 37–52.
[156] M. Berger, Geometry I, Springer Science & Business Media, 2009.
[157] S. Lloyd, IEEE Trans. Inform. Theory 28 (1982) 129–137.
[158] D. Arthur, S. Vassilvitskii, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, USA, 2007, pp. 1027–1035.
[159] C. Fernandez-Lozano, J.A. Seoane, M. Gestal, T.R. Gaunt, J. Dorado, C. Campbell, Soft Comput 19 (2015) 2469–2480.
[160] E.W. Steyerberg, in: Clinical Prediction Models, Springer New York, New York, NY, 2009, pp. 11–31.
[161] R. Battiti, IEEE Trans. Neural Netw. 5 (1994) 537–550.
[162] A. Kraskov, H. Stögbauer, P. Grassberger, Phys. Rev. E 69 (2004) 066138.
[163] D.R. Cox, Journal of the Royal Statistical Society Series B: Statistical Methodology 20 (1958) 215–232.
[164] L. Yang, A. Shami, Neurocomputing 415 (2020) 295–316.
[165] H.-L. Le, T.-T. Le, T.-T.-H. Vu, T. Doan-Hieu, C. Dinh Van, C. Minh, T.-T.-T. Ngo, International Journal of Computers and Their Applications 30 (2023) 351–361.
[166] P. Contreras, J. Orellana-Alvear, P. Muñoz, J. Bendix, R. Célleri, Atmosphere 12 (2021) 238.
[167] A. Maleki, M. Raahemi, H. Nasiri, Biomedical Signal Processing and Control 86 (2023) 105152.
[168] X.Y. Liew, N. Hameed, J. Clos, Machine Learning with Applications 6 (2021) 100154.
[169] A. Defazio, F. Bach, S. Lacoste-Julien, in: Proceedings of the 28th International Conference on Neural Information Processing Systems, MIT Press, Montreal, Canada, 2014, pp. 1646–1654.
[170] J. Novakovic, A. Veljovic, in: 2011 IEEE 9th International Symposium on Intelligent Systems and Informatics, IEEE, Subotica, Serbia, 2011, pp. 465–470.
[171] S. Trenn, IEEE Trans. Neural Netw. 19 (2008) 836–844.
[172] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, A. Talwalkar, Journal of Machine Learning Research 18 (2018) 1–52.
[173] D.S. Soper, Algorithms 16 (2022) 17.
[174] L.A. Yates, Z. Aandahl, S.A. Richards, B.W. Brook, Ecological Monographs 93 (2023) e1557.
[175] A.C. Müller, S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists, O'Reilly Media, Inc., 2016.

[176] D. Berrar, P. Flach, Briefings in Bioinformatics 13 (2012) 83–97.
[177] Y. Padrez, L. Golubewa, I. Timoshchenko, A. Enache, L.G. Eftimie, R. Hristu, D. Rutkauskas, Computerized Medical Imaging and Graphics (2024) 102440.
[178] S.E. Dunsmore, D.E. Rannels, American Journal of Physiology-Lung Cellular and Molecular Physiology 270 (1996) L3–L27.
[179] L.A. Organ, A.-M.R. Duggan, E. Oballa, S.C. Taggart, J.K. Simpson, A.R. Kang'ombe, R. Braybrooke, P.L. Molyneaux, B. North, Y. Karkera, D.J. Leeming, M.A. Karsdal, C.B. Nanthakumar, W.A. Fahy, R.P. Marshall, R.G. Jenkins, T.M. Maher, Respir Res 20 (2019) 148.
[180] H. Jessen, N. Hoyer, T.S. Prior, P. Frederiksen, M.A. Karsdal, D.J. Leeming, E. Bendstrup, J.M.B. Sand, S.B. Shaker, Respir Res 22 (2021) 205.
[181] H. Birkedal-Hansen, W.G.I. Moore, M.K. Bodden, L.J. Windsor, B. Birkedal-Hansen, A. DeCarlo, J.A. Engler, Critical Reviews in Oral Biology & Medicine 4 (1993) 197–250.
[182] A. Pardo, S. Cabrera, M. Maldonado, M. Selman, Respir Res 17 (2016) 23.
[183] N. Vuillemin, P. Mahou, D. Débarre, T. Gacoin, P.-L. Tharaux, M.-C. Schanne-Klein, W. Supatto, E. Beaurepaire, Sci Rep 6 (2016) 29863.
[184] H. Mehidine, A. Chalumeau, F. Poulon, F. Jamme, P. Varlet, B. Devaux, M. Refregiers, D. Abi Haidar, Sci Rep 9 (2019).
[185] P.T.C. So, C.Y. Dong, B.R. Masters, K.M. Berland, Annu. Rev. Biomed. Eng. 2 (2000) 399–429.
[186] W. Draxinger, J.P. Kolb, D. Weng, H. Hakert, S.N. Karpf, R. Huber, J. Popp, T. Meyer, J. Limpert, T. Gottschall, M. Eibl, R. Brinkmann, R. Birngruber, in: A. Periasamy, P.T. So, K. König (Eds.), Multiphoton Microscopy in the Biomedical Sciences XIX, SPIE, San Francisco, United States, 2019, p. 83.
[187] I. Adzerikho, O. Yatsevich, T. Vladimirskaja, G. Semenkova, N. Amaegberi, European Heart Journal 42 (2021).
[188] S. Wu, H. Li, H. Yang, X. Zhang, Z. Li, S. Xu, JBO 16 (2011) 040502.
[189] E.T. Osei, L. B. Mostaço-Guidolin, A. Hsieh, S.M. Warner, M. AL-Fouadi, M. Wang, D.J. Cole, G.N. Maksym, T. S. Hallstrand, W. Timens, C.-A. Brandsma, I.H. Heijink, T.-Louise. Hackett, Sci Rep 10 (2020) 8721.
[190] T. Zhang, N. Kawaguchi, K. Yoshihara, E. Hayama, Y. Furutani, K. Kawaguchi, T. Tanaka, T. Nakanishi, Respir Res 20 (2019) 79.
[191] M. Gharaee-Kermani, E.M. Denholm, S.H. Phan, Journal of Biological Chemistry 271 (1996) 17779–17784.
[192] S. O'Reilly, M. Ciechomska, R. Cant, J.M. van Laar, Journal of Biological Chemistry 289 (2014) 9952–9960.
[193] S. O'Reilly, R. Cant, M. Ciechomska, J.M. van Laar, Clin Trans Immunol 2 (2013) e4.

[194] K. Mirsanaye, L. Uribe Castaño, Y. Kamaliddin, A. Golaraei, L. Kontenis, E. Žurauskas, R. Navab, K. Yasufuku, M.-S. Tsao, B.C. Wilson, V. Barzda, Sci Rep 12 (2022) 20713.
[195] J. Winkler, A. Abisoye-Ogunniyan, K.J. Metcalf, Z. Werb, Nat Commun 11 (2020) 5120.
[196] D. Kang, K. Li, L. Zuo, H. Wu, S. Huang, J. Zhang, B. Wei, C. Xu, H. Wang, New J. Chem. 48 (2024) 7990–7996.
[197] D. Xydias, G. Ziakas, S. Psilodimitrakopoulos, A. Lemonis, E. Bagli, T. Fotsis, A. Gravanis, D.S. Tzeranis, E. Stratakis, Biomed. Opt. Express 12 (2021) 1136.
[198] S. Brasselet, Adv. Opt. Photon. 3 (2011) 205.
[199] A.E. Tuer, S. Krouglov, N. Prent, R. Cisek, D. Sandkuijl, K. Yasufuku, B.C. Wilson, V. Barzda, J. Phys. Chem. B 115 (2011) 12759–12769.
[200] V. Triggiani, E. Guastamacchia, B. Licchelli, E. Tafaro, Thyroid 18 (2008) 1017–1018.
[201] L. Cardisciani, F. Policardo, P. Tralongo, V. Fiorentino, E.D. Rossi, Pathologica 114 (2022) 373–375.
[202] L.B. Ferreira, E. Gimba, J. Vinagre, M. Sobrinho-Simões, P. Soares, IJMS 21 (2020) 7718.
[203] I.T. Jolliffe, Applied Statistics 21 (1972) 160.
[204] A. Kallner, in: Laboratory Statistics, Elsevier, 2018, pp. 1–140.
[205] R. Ghossein, Head and Neck Pathol 3 (2009) 86–93.
[206] Y.-Q. Chen, J.-C. Kuo, M.-T. Wei, Y.-C. Chen, M.-H. Yang, A. Chiou, Acta Biomaterialia 84 (2019) 280–292.
[207] J.R. Giles, A.-M. Globig, S.M. Kaech, E.J. Wherry, Immunity 56 (2023) 2231–2253.
[208] H.C. Pruitt, D. Lewis, M. Ciccaglione, S. Connor, Q. Smith, J.W. Hickey, J.P. Schneck, S. Gerecht, Matrix Biology 85–86 (2020) 147–159.
[209] M. Papanicolaou, A.L. Parker, M. Yam, E.C. Filipe, S.Z. Wu, J.L. Chitty, K. Wyllie, E. Tran, E. Mok, A. Nadalini, J.N. Skhinas, M.C. Lucas, D. Herrmann, M. Nobis, B.A. Pereira, A.M.K. Law, L. Castillo, K.J. Murphy, A. Zaratzian, J.F. Hastings, D.R. Croucher, E. Lim, B.G. Oliver, F.V. Mora, B.L. Parker, D. Gallego-Ortega, A. Swarbrick, S. O'Toole, P. Timpson, T.R. Cox, Nat Commun 13 (2022) 4587.
[210] Z. Yuan, Y. Li, S. Zhang, X. Wang, H. Dou, X. Yu, Z. Zhang, S. Yang, M. Xiao, Mol Cancer 22 (2023) 48.
[211] L.S.M. Lecker, C. Berlato, E. Maniati, R. Delaine-Smith, O.M.T. Pearce, O. Heath, S.J. Nichols, C. Trevisan, M. Novak, J. McDermott, J.D. Brenton, P.R. Cutillas, V. Rajeeve, A. Hennino, R. Drapkin, D. Loessner, F.R. Balkwill, Cancer Research 81 (2021) 5706–5719.
[212] R. Hristu, L.G. Eftimie, B. Paun, S.G. Stanciu, G.A. Stanciu, Journal of Biophotonics 13 (2020) e202000262.

[213] D. Tokarz, R. Cisek, A. Joseph, S.L. Asa, B.C. Wilson, V. Barzda, Laboratory Investigation 100 (2020) 1280–1287.
[214] B.R. Haugen, E.K. Alexander, K.C. Bible, G.M. Doherty, S.J. Mandel, Y.E. Nikiforov, F. Pacini, G.W. Randolph, A.M. Sawka, M. Schlumberger, K.G. Schuff, S.I. Sherman, J.A. Sosa, D.L. Steward, R.M. Tuttle, L. Wartofsky, Thyroid 26 (2016) 1–133.
[215] J.A. Sipos, E.L. Mazzaferri, Clinical Oncology 22 (2010) 395–404.
[216] M.E. Cabanillas, D.G. McFadden, C. Durante, The Lancet 388 (2016) 2783–2795.
[217] N. A. Shah, J. Suthar, T. A., A. Enache, L. G. Eftimie, R. Hristu, A. Paul, in: S.M. Astley, W. Chen (Eds.), Medical Imaging 2024: Computer-Aided Diagnosis, SPIE, San Diego, United States, 2024, p. 79.
[218] S.D. Kok, P.M.R. Schaap, L. Van Dommelen, L.M.G. Van Huizen, C. Dickhoff, E.M.N. Dijkum, A.F. Engelsman, P. Van Der Valk, M.L. Groot, Journal of Biophotonics 17 (2024) e202300079.
[219] S. Petscharnig, M. Lux, S. Chatzichristofis, in: Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, ACM, Florence Italy, 2017, pp. 1–6.
[220] L. Jing, L. Li, Z. Sun, Z. Bao, C. Shao, J. Yan, Q. Pang, Y. Geng, L. Zhang, X. Wang, Z. Wang, Journal of Cardiovascular Pharmacology 74 (2019) 372–378.
[221] S. De Santis, G. Sotgiu, A. Crescenzi, C. Taffon, A.C. Felici, M. Orsini, Journal of Pharmaceutical and Biomedical Analysis 190 (2020) 113534.
[222] A.D. Hofemeier, H. Hachmeister, C. Pilger, M. Schürmann, J.F.W. Greiner, L. Nolte, H. Sudhoff, C. Kaltschmidt, T. Huser, B. Kaltschmidt, Sci Rep 6 (2016) 26716.
[223] K.S. Shin, M. Laohajaratsang, S. Men, B. Figueroa, S.M. Dintzis, D. Fu, Theranostics 10 (2020) 5865–5878.
[224] Z. Li, J. Wei, B. Chen, Y. Wang, S. Yang, K. Wu, X. Meng, Molecules 28 (2023) 3705.
[225] T.R. Cox, Nat Rev Cancer 21 (2021) 217–238.
[226] A. Bailly, C. Blanc, É. Francis, T. Guillotin, F. Jamal, B. Wakim, P. Roy, Computer Methods and Programs in Biomedicine 213 (2022) 106504.
[227] T.E. Luvhengo, I. Bombil, A. Mokhtari, M.S. Moeng, D. Demetriou, C. Sanders, Z. Dlamini, Biomedicines 11 (2023) 1217.
[228] H.G. Lee, Y.J. Choi, E.C. Chung, J Korean Soc Radiol 64 (2011) 17.

7 SANTRAUKA

7.1 Įvadas

Vėžys yra viena pagrindinių žmonių mirtingumo priežasčių visame pasaulyje [1]. Remiantis 2024 m. paskelbta Tarptautinės vėžio tyrimų agentūros (IARC) ataskaita [2] , 2022 m. visame pasaulyje užregistruota apie 20 milijonų naujų vėžio atvejų ir apie 9,7 milijono mirčių nuo vėžio. Be dabartinės statistikos, IARC taip pat pateikė demografinę prognozę, kad iki 2050 m. naujų vėžio atvejų skaičius pasieks 35 milijonus, o tai padidins mirtingumą. Ankstyva diagnozė ir naujų vėžio progresavimo žymenų bei būdingų požymių atradimas negali užkirsti kelio dideliam vėžio atvejų augimui, tačiau tai gali žymiai sumažinti mirtingumą, užkirsti kelią vėžio atsinaujinimui ir metastazėms, pagerinti pacientų atsigavimą ir gyvenimą po gydymo. Tobulėjant technologinei įrangai ir dirbtinio intelekto įtraukimui į vėžio duomenų analizę, kompiuterizuotos vėžio audinių vaizdų analizės pažanga tampa galingu įrankiu, galinčiu efektyviai papildyti įprastinę vėžio audinių mėginių analizę, kurią tradiciškai atlieka patyrę patologai vizualinės apžiūros metu, ir suteikti naujų vėžio diagnostikos įžvalgų [3].

Skydliaukės vėžys yra viena labiausiai paplitusių piktybinių endokrininės sistemos ligų, kuriai būdingas nekontroliuojamas skydliaukės ląstelių dauginimasis. Pastaraisiais dešimtmečiais šios ligos atvejų skaičius nuolat didėjo, o 2020 m. visame pasaulyje užregistruota apie 586 000 naujų atvejų [4]. Papilinė skydliaukės karcinoma (PTC) ir folikulinė skydliaukės karcinoma (FTC) yra dažniausios gerai diferencijuojamos karcinomos, kurios kartu sudaro apie 88% visų skydliaukės navikų [5].

Skydliaukės mazgų kapsulės būsena, t. y., faktas, ar juos supa skaidulinė kolageno kapsulė, yra pagrindinis histopatologinis požymis. Kapsulės buvimas ir vientisumas turi įtakos naviko invazyvumo, piktybiškumo potencialo ir prognozės vertinimui. PTC atveju kapsulė gali būti nepilna arba infiltracinė, o FTC atveju ji paprastai būna geriau išsivysčiusi. Kapsulinė invazija, kai naviko ląstelės prasiskverbia per visą kapsulės storį, yra pagrindinis kriterijus, leidžiantis atskirti gerybinius ir piktybinius folikulinius navikus [6]. Tačiau navikų klasifikavimas pagal kapsulės invaziją yra sudėtingas dėl stebėtojo kintamumo ir sudėtingų histologinių kriterijų. Be to, standartiniai histopatologiniai metodai, kuriais dažnai tiriami tik riboti audinių pjūviai, sukelia mikroinvazijų nepastebėjimo riziką.

Nors FTC prognozė dažnai yra blogesnė nei PTC [9,10] ir jai yra reikalinga visiška tiroidektomija [11], mažos rizikos PTC atvejai dažnai yra per dažnai diagnozuojami ir gydomi [12]. Todėl tikslus FTC ir PTC diferencijavimas yra būtinas, siekiant išvengti nereikalingo agresyvaus gydymo ir sumažinti pooperacines komplikacijas [13].

Arterinė hipertenzija yra viena dažniausių gretutinių vėžiu sergančių pacientų ligų [14]] ir yra dažnas vėžio gydymo šalutinis poveikis. Dasatinibu

ar kitais tirozino kinazės inhibitoriais gydomiems pacientams buvo pastebėta su vėžiu susijusi plaučių arterinė hipertenzija (PAH) [15]. PAH yra sunki kraujagyslių liga, kuriai būdingas padidėjęs plaučių arterinis slėgis, dėl kurio atsiranda kraujagyslių persitvarkymas ir per didelis fibrilinio kolageno kaupimasis plaučių arterijose. Šie pokyčiai prisideda prie kraujagyslių sustingimo ir ligos progresavimo [16].

Kolageno kiekio ir jo pasiskirstymo struktūriniai pokyčiai yra pagrindiniai tarpląstelinės matricos (ECM) remodeliacijos, susijusios su tokiomis patologinėmis būklėmis kaip PAH požymiai. Skydliaukės vėžio atveju kolageno remodeliacija taip pat vaidina svarbų vaidmenį naviko progresavime, o kapsulės pokyčiai koreliuoja su piktybiškumu ir invazyvumu. Šių įprastų patologinių mechanizmų supratimas pabrėžia kolageno nustatymo svarbą diagnozuojant ir stebint ligos progresavimą. Dėl pagrindinių patologinių pokyčių sudėtingumo ir didelio persidengimo PAH laikoma į vėžį panašia liga [17]. Supratimas apie ryšį tarp ECM remodeliacijos sergant vėžiu ir PAH su šių ligų progresavimu, sunkumu ir laipsniu bei šių duomenų panaudojimas tiksliai diagnozei nustatyti galėtų padėti išsamiai interpretuoti patologinę būklę ir veiksmingiau gydyti tiek vėžį, tiek PAH.

Antrosios harmonikos generacijos (SHG) mikroskopija suteikia metodą be žymeklių fibriliniam kolagenui - pagrindiniam skydliaukės mazgų kapsulių struktūriniam komponentui, vizualizuoti. SHG biovaizdavimas yra ypač efektyvus vertinant kolageno turtingų audinių pokyčius [18], nes kolagenas sukuria stiprų SHG signalą dėl savo necentrosimetrinės struktūros [19]. Ankstesni tyrimai [20,21] parodė, kad SHG mikroskopija kartu su kiekybine vaizdų analize gali atskirti gerybinius ir piktybinius skydliaukės mazgus. Nors įprastas SHG vaizdavimas yra grindžiamas skenuojančiais lazerio spinduliais, plataus lauko SHG mikroskopija leidžia vizualizuoti ištisas histologines plokšteles [22,23], leidžiant atlikti išsamią kolageno architektūros analizę.

SHG mikroskopija yra ypač vertinga vertinant audinių struktūrinę anizotropiją naudojant šviesos poliarizaciją. Poliarizacinė SHG mikroskopija (PSHG) buvo naudojama kiekybiškai analizuojant audinių kolageno mikrostruktūrą [24–26], įskaitant skydliaukę [21,27,28]. Nors SHG mikroskopijos variantai su lazerinio spindulio skenavimu pasirodė esą sėkmingi biomedicininiame vaizdavime, plataus lauko SHG mikroskopija sulaukia pripažinimo [23] dėl visų histologinių plokštelių atvaizdavimo. Panašiai kaip ir susijusi vaizdavimo technika - dviejų fotonų sužadinta fluorescencinė mikroskopija, kuriai taip pat galimi plataus lauko variantai [31], plataus lauko SHG mikroskopija išsivystė iš intensyvumu pagrįstų taikymo sričių [32] iki kiekybinės analizės [33] ir net tiesioginio vaizdavimo taikymo sričių [34].

Be SHG vaizdavimo, tekstūros analizės metodai, įskaitant pirmos eilės statistiką (FOS), antros eilės statistiką (SOS) ir aukštesnės eilės statistiką (HOS), pateikia kiekybinius kolageno tinklų savybių aprašymus [35,36]. Šie metodai plačiai naudojami tokioje medicininėje diagnostikoje kaip

kompiuterinė tomografija [37] ir magnetinio rezonanso tomografija (MRI) [38].

Vaizdo duomenų interpretavimas yra ribotas, kai jie yra analizuojami rankiniu būdu naudojant tradicinį vaizdų apdorojimo kanalą. Naudojantis dideliais nuskaitytais plotais ir didele erdvine plataus lauko SHG mikroskopijos skiriamąja geba, galimybė derinti jį su dviejų fotonų sužadinta fluorescencine mikroskopija ir pridėti dirbtinį intelektą prie įvairių iš vaizdo duomenų išgautų savybių atveria kelią dideliu tikslumu ir našumu pasižyminčiai automatizuotai kompiuterizuotai vėžio audinių mėginių vaizdų analizei.

Mašininio mokymosi (ML) metodai vis dažniau naudojami vėžio audinių analizei ir klasifikavimui automatizuoti ir tobulinti. Neprižiūrimos mašininio mokymosi (ML) technikos yra daugiausia skirtos vaizdų segmentavimui ir specifinių vaizdų ypatybių modelių aptikimui, remiantis būdingais ryšiais [39,40]. Prižiūrimi ML algoritmai daugiausia taikomi klasifikavimo problemoms spręsti ir yra rekomenduojami ligų diagnostikai. ML klasifikatoriai, įskaitant gilaus mokymosi (DL) modelius, parodė daug žadančius rezultatus atskiriant skirtingus vėžio tipus pagal MRT, kompiuterinę tomografiją ir SHG vaizdų analizę [41–44]. Tačiau efektyviam ML pagrindu veikiančiam klasifikavimui reikalingi aukštos kokybės duomenys, nes žymeklių triukšmas (neteisingas pavyzdžių žymėjimas) ir požymių triukšmas (nesvarbūs arba pertekliniai parametrai) gali reikšmingai paveikti modelio našumą [45]. Realaus pasaulio vaizdų duomenys retai atitinka šį kriterijų, todėl lieka vietos tolesniam ML algoritmų, architektūrų ir strategijų pritaikymui konkrečiam duomenų tipui.

Kiekybinės įžvalgos apie skydliaukės mazgų kapsulės struktūras ir ECM remodeliavimą yra būtinos norint pasiekti pažangos ir tikslumo diagnozuojant skydliaukės vėžį bei suprasti patologinių būklių evoliuciją. Didelės apimties plataus lauko SHG mikroskopija, kaip žymeklių nereikalaujanti ir kolagenui specifinė vaizdavimo technika, gali supaprastinti audinių mėginių paruošimą, pašalindama audinių fiksavimo ir dažymo etapus bei leisdama mėginius išmatuoti iš karto po chirurginio pašalinimo. Savo ruožtu, didelio masto SHG vaizdams analizuoti iš vėžio audinių skirtų ML pagrįstų metodų kūrimas rodo automatizuotą diagnostikos metodą, kuris gali papildyti tradicinį vizualinį vėžio audinių mėginių patikrinimą, sumažinant klaidingos diagnozės tikimybę ir palaikant optimalų klinikinių sprendimų priėmimą. Todėl išsamūs tyrimai šia kryptimi yra labai svarbūs, nes SHG plataus lauko mikroskopijos paprastumas, greitis ir specifiškumas kartu su efektyviais vėžio analizės ML metodais gali diagnozuoti tiksliai ir laiku, taip sumažinant mirčių nuo vėžio skaičių, nors tikimasi, kad atvejų skaičius per ateinančius dešimtmečius sparčiai didės.

7.2 Disertacijos tikslai

Disertacijos tikslas yra sukurti mašininio mokymosi modelius, kurie naudoja SHG didelio masto audinių pjūvių skenavimą patologinėms būklėms interpretuoti ir ligoms diagnozuoti.

7.3 Disertacijos uždaviniai

Siekiant aukščiau iškelto tikslo, disertacijos rėmuose buvo suformuluoti ir išspręsti šie uždaviniai:

1. Atlikti su monokrotalinu sukeltu PAH sergančiomis žiurkėmis skirtingose ligos stadijose atliktų SHG skenavimų ir kontrolinės grupės žiurkių plaučių audinių mėginių plataus lauko SHG intensyvumo charakteristikų ir tekstūros ypatybių statistinę analizę, atskleisti SHG vaizdų ypatybių modelius ir palyginti juos su imunohistocheminės analizės rezultatais.
2. Taikyti neprižiūrimą mašininio mokymosi algoritmą *k*-vidurkių klasterizavimą iš viso skydliaukės mazgo pjūvių plataus lauko gautų PSHG vaizdų 2D parametrų žemėlapių analizei, remiantis cilindriniu kolageno skaidulų hiperpoliarizuotumo modeliu, atskleisti kolageno skaidulų ultrastruktūros modelius nepažeistoje kapsulėje ir invazinėse srityse.
3. Atlikti neprižiūrimą mašininio mokymosi plataus lauko SHG skenavimų, skirtų kolageno pasiskirstymui papilinės skydliaukės karcinomos pjūviuose, intensyvumo ir tekstūros ypatybių analizę (pagrindinių komponenčių analizę (PCA) ir *k*-vidurkių klasterizaciją), siekiant nustatyti kapsulės invazines sritis ir pasiūlyti kiekybinį nepažeistos kapsulės ir invazinių sričių aprašymą.
4. Sukurti prižiūrimus ML modelius papilinių ir folikulinių skydliaukės karcinomų automatinei diferencinei diagnostikai, naudojant plataus lauko SHG vaizdavimą, atsižvelgiant į žymėjimo ir požymių triukšmo poveikį šių modelių prognozavimo našumui.

7.4 Ginamieji teiginiai

1. Plataus kampo SHG plaučių audinio pjūvių vaizdų statistinė analizė atskleidžia ir kokybiškai bei kiekybiškai apibūdina būdingus kolageno organizacijos, morfologijos ir kolageno kiekio pokyčius, susijusius su skirtingais PAH etapais.
2. Iš plataus lauko poliarizacijos būdu išskaidytų SHG vaizdų iš viso skydliaukės mazgų pjūvių išgautų cilindrinių modelių parametrų *k*-vidurkių klasterizacija leidžia atskirti kapsulės invazines sritis nuo nepažeistų vėžio ląsteles supančių kapsulės sričių, atskleidžiant kolageno ultrastruktūros modelius.
3. Neprižiūrimas mašininis mokymasis pagerina SHG vaizdų analizę, atskleidžia papilinės skydliaukės karcinomos kapsulės tekstūrinį heterogeniškumą ir leidžia nustatyti kapsulės invaziją, papildomo tyrimo

reikalaujančias sunkiai atskiriamas mikroinvazijas ir sritis, remiantis konkrečiais vaizdo parametrų rinkiniais.

4. Prižiūrimas mašininio mokymosi modelis C-SVC leidžia 84,73% tikslumu diferencijuoti papilines ir folikulines skydliaukės karcinomas, remiantis SHG vaizdavimu.

7.5 Darbo naujumas ir aktualumas

1. Remiantis iš SHG kolageno tinklo organizacijos vaizdų gautais statistiniais parametrais kartu su imunohistocheminės analizės rezultatais, buvo nustatytos specifinės PAH progresavimo fazės ir pasiūlytas jų kiekybinis aprašymas. Šios fazės galėtų būti įvertintos kaip PAH patogenezės kontroliniai taškai.

2. Neprižiūrima kolageno ultrastruktūros ir su orientacija susijusių parametrų, išgautų iš viso skydliaukės mazgų pjūvio PSHG vaizdų rinkinių, ML analizė leido sukurti būdingus skydliaukės mazgų žemėlapius ir jų kiekybinį aprašymą, palengvinant skirtingų mazgų kapsulės sričių palyginimą ir išryškinant invazines sritis kapsulėje.

3. Neprižiūrima plataus lauko SHG vaizdų su visais skydliaukės mazgais ML analizė leido atskleisti PTC supančios kolageno kapsulės tekstūrinį heterogeniškumą. Kiekybinis šio heterogeniškumo aprašymas, atsispindintis tekstūros ypatybėse ir kolageno kapsulės specifiniame erdviniame pasiskirstyme, atskleidžia su vėžio plitimu susijusius kolageno tinklo pokyčius.

4. Sukurtas prižiūrimas ML pagrįstas metodas leidžia efektyviai atskirti FTC ir PTC naudojant didelius skydliaukės pjūvių SHG vaizdų duomenų rinkinius. Siūloma duomenų triukšmo valdymo strategija pagerina diagnostikos tikslumą ir parodo SHG mikroskopija pagrįstos automatizuotos PTC ir FTC diagnostikos įgyvendinamumą.

Disertacijos rezultatai atveria kelią automatizuotai ir kiekybinei SHG vaizdavimu pagrįstai PAH diagnozei ir turi potencialo tapti papildomu objektyviu PAH sukeltos fibrozės gydymo metodu, apsaugotu nuo klaidingų sprendimų priėmimo. Be to, tiek PSHG, tiek plataus lauko SHG mikroskopijos įgalinta kiekybinė analizė gali būti naudinga automatizuotam kapsulės invazinių vietų vertinimui skydliaukės patologijoje, padedant atskleisti sunkiai atskiriamas invazijas ir išryškinant PTC kapsulės sritis, kurias reikia ištirti atidžiau ir kruopščiau. Kolageno ultrastruktūros duomenys gali suteikti įžvalgų apie vėžio progresavimo, plitimo ir metastazių molekulinį pagrindą. Visa tai suteikia patikimą pagrindą laikyti ML pagalba atliekamą SHG mikroskopiją nauju efektyviu skydliaukės vėžio diagnostikos metodu.

7.6 Metodika

Buvo tiriami dviejų tipų audinių mėginiai: (i) Plaučių audinių mėginiai iš žiurkių, kurių monokrotalino (MCT) sukelta PAH buvo skirtingose PAH stadijose, ir iš sveikų žiurkių (kontrolinė grupė) ir (ii) žmogaus skydliaukės audinių mėginiai su PTC ir FTC pjūviais.

Visi eksperimentai buvo atlikti pagal atitinkamas gaires ir reglamentus, vadovaujantis Helsinkio deklaracija ir gavus raštišką pacientų sutikimą (skydliaukės mėginiai). Visi mėginiai buvo formalinu fiksuoti, parafinu įlieti 3–5 μm storio pjūviai, padėti tarp dviejų stiklinių plokštelių. Visi audinių mėginiai buvo nudažyti hematoksilinu ir eozinu (H&E).

Žiurkių plaučių audinių pjūviai buvo eksperimentiškai analizuojami imunohistochemijos (IHC) vaizdavimo metodu siekiant atskleisti I ir III kolageno bei metaloptotazės raiškos lygių pokyčius (TIMP)-1, kurie yra su PAH progresavimu susiję molekuliniai fibrozės vystymosi žymenys. Siekiant kiekybiškai įvertinti biomolekulinių žymenų raišką, *teigiamų pikselių skaičiavimo* algoritmas buvo pritaikytas aiškiai apibrėžtų branduolių, ląstelių ir kraujagyslių nepersidengiančių sričių plaučių ekspresijos indekso (IE) apskaičiavimui. Be to, audinių kolageno tinklai buvo vaizduojami naudojant specialiai sukurtą plataus lauko SHG mikroskopijos sistemą. Taip pat buvo atliktas dviejų fotonų sužadinimo fluorescencijos (TPEF) vaizdavimas, siekiant vizualizuoti plaučių audinio pjūvių nekolagenines sritis.

Skydliaukės pjūviai buvo vaizduojami naudojant tą pačią specialiai pagamintą plataus lauko SHG mikroskopijos įrangą su lazeriniu sužadinimu ir apskritimine poliarizacija arba tiesine poliarizacija, jei SHG vaizdavimas buvo atliekamas poliarizacijos būdu. Iš viso buvo gauti atitinkamai 23652 ir 21708 150×150 μm² dydžio PTC ir FTC SHG vaizdai.

Vaizdų analizė buvo atlikta naudojant neprižiūrimus ir prižiūrimus ML metodus, pagrįstus intensyvumo ir tekstūros ypatybėmis, išgautomis iš plataus lauko SHG vaizdų. Intensyvumo ir tekstūros funkcijos apėmė greitąją Furjė transformaciją (FFT), FOS, SOS (apskaičiuotą pagal pilkojo lygio bendro pasireiškimo matricą (GLCM) 1, 3, 6, 9, 12 veiksmams) ir HOS (apskaičiuotą pagal pilkojo lygio bėgimo ilgio matricą (GLRLM)). Neprižiūrimas ML apėmė PCA ir *k*-vidurkių klasterizaciją bei buvo pritaikytas viso audinio SHG skenavimui segmentuoti, siekiant atskleisti kolageno tekstūrų heterogeniškumą PTC ir FTC kapsulėse. Prižiūrimas ML apėmė Random Forest (RF) [114], logistinę regresiją (LR), eXtreme gradient stiprinimą (XGBoost) [115], Šviesos gradientą stiprinantį aparatą (LightGBM) [116], C atramos vektorių klasifikaciją (C-SVC) [117] ir Daugiasluoksnį perceptroną (MLP) [118], kurie buvo pritaikyti PTC ir FTC klasifikavimui pagal iš SHG vaizdų tekstūros išgautas ypatybes. Klasifikatorių optimizavimas buvo atliktas naudojant tinklelio paiešką kartu su 10 kartų kryžminiu patvirtinimu. Siekiant įvertinti funkcijų pertekliaus ir nereikšmingumo poveikį klasifikavimui, funkcijų parinkimas buvo atliktas naudojant rekursinį funkcijų eliminavimą su kryžminiu patvirtinimu ir tiesiniu SVC įverčiu (RFECV-LinearSVC), po kurio sekė abipusės informacijos funkcijų parinkiklis (MIFS). Funkcijų svarbos analizė atlikta naudojant permutacijos svarbos analizę (PIA) arba funkcijų svarbos analizę, įterptą į klasifikatorių bibliotekas.

Skaičiavimai buvo atliekami su Python v3.9, Intel i7-13700KF CPU su 16 branduoliais ir 24 gijomis; 32 GB laisvosios prieigos atmintimi; Nvidia GeForce RTX 3060 Ti grafikos plokšte su 4864 branduoliais.

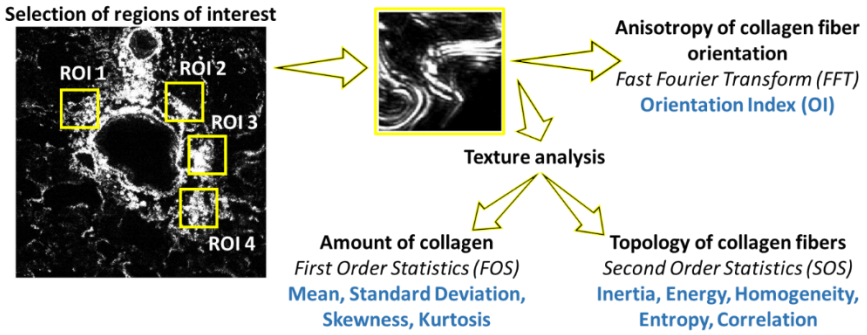
7.7 Rezultatai

7.7.1 Statistinė žiurkių PAH fibrozės progresavimo intensyvumo ir tekstūros analizė

Šis skyrius skirtas kiekybinei ir kokybinei fibrozės, lydinčios MCT sukeltos PAH progresavimą žiurkėms, vystymosi analizei, remiantis plataus lauko SHG plaučių audinio pjūvių vaizdais. Šiame skyriuje pateikiami rezultatai buvo skelbiami ***Paper A*** ir pateikiami **1** ir **2 konferencijose**.

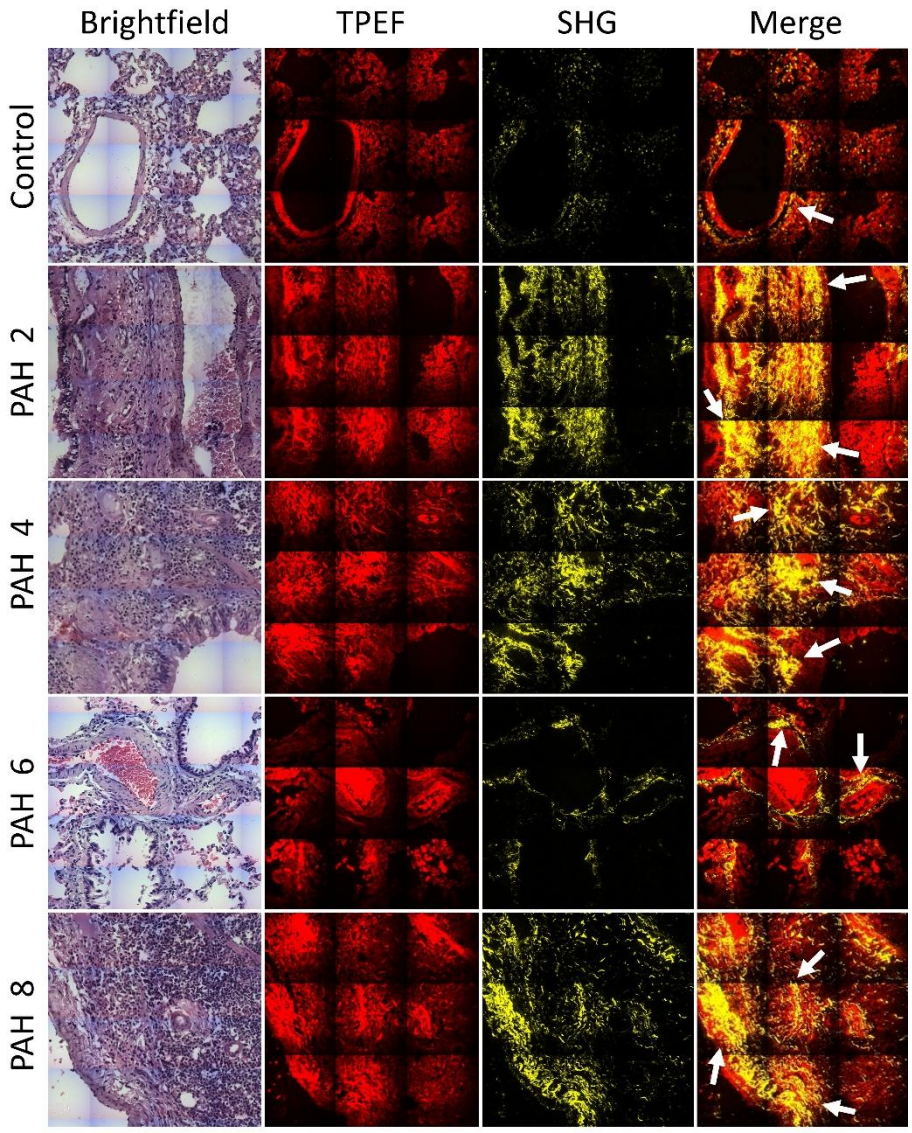
Jautrių, žymeklių nereikalaujančių ir greitų didele skiriamąja geba ir patikima statistine vaizdų analize pasižyminčių vaizdavimo metodų kūrimas yra nepaprastai svarbus ankstyvai PAH progresavimo diagnostikai. Plataus lauko plaučių audinio SHG mikroskopija kartu su kiekybine SHG vaizdų analize gali žymiai supaprastinti fibrozės progresavimo nustatymą PAH metu, leisti objektyviai įvertinti patologijos stadiją ir apsaugoti vertinimą nuo klaidingų sprendimų priėmimo, kaip histologinės analizės atveju.

Siekiant tai įrodyti, žiurkių plaučių audinių mėginių SHG vaizdai buvo nuskaityti ir analizuojami rankiniu būdu pasirinktose dominančiose srityse (ROIs). ROI vaizdai buvo kiekybiškai apibūdinti naudojant FFT ir tekstūros analizę, kur tekstūros analizę sudarė FOS ir SOS *7.1 paveikslas*.



7.1 paveikslas SHG vaizdų analizės procedūros. Perspausdinta iš [***Paper A***].

Be to, SHG vaizdavimas buvo derinamas su TPEF vaizdavimu, kad būtų galima vizualizuoti tiek fibrozines struktūras, tiek išorinį audinį ir taip gauti išsamų kolageno skaidulų išsišakojimo ir padidėjimo audinio viduje vaizdą. Endogeninis TPEF signalas dažnai kyla iš tokių metabolinių junginių kaip nikotinamido adenino dinukleotidas, flavinai, lipopigmentai, porfirinai ir t. t. Mūsų atveju vienalaikei šviesaus lauko mikroskopijai taikomas H&E dažymas taip pat prisideda prie TPEF signalo. H&E dažytų plaučių audinio pjūvių vaizdai pateikiami *7.2 paveiksle* su geltonos spalvos kolageno SHG ir raudonos spalvos TPEF.



7.2 paveikslas. Kontrolinės grupės žiurkių ir 2, 4, 6 ir 8 savaitės PAH progresijos žiurkių plaučių audinio šviesaus lauko vaizdai, TPEF vaizdai, SHG vaizdai ir kombinuoti TPEF bei SHG vaizdai. Atvaizdo dydis yra 450 µm × 450 µm. Paimta iš [***Paper A***].

Kontroliniame mėginyje kolagenas supa kraujagyslės sienelę (pažymėta rodyklėmis *7.2 paveiksle*), bet jo nėra plaučių audinyje. H&E dažytų PAH vaizduose matyti tik kraujagyslių sienelių sustorėjimas lyginant su kontroliniu mėginiu, o SHG/TPEF vaizduose – reikšmingas kolageno padaugėjimas. *7.2 paveiksle* galima matyti, kad kolageno kiekis laikui bėgant didėja, o labai susiformavusios ilgos skaidulos sudaro tankius tinklus tiek aplink kraujagysles, tiek aplinkiniuose audiniuose, taip giliai plisdamos alveolių

srityje (pažymėta rodyklėmis *7.2 paveiksle*).). Tai rodo reikšmingą kolageno ekspresijos padidėjimą po 4–8 savaičių PAH progresavimo.

Kiekybinė IHC analizė leidžia atskleisti I ir III kolageno bei TIMP-1, kurie yra plaučių arterinės hipertenzijos (PAH) progresavimo lydinčių fibrozės vystymosi molekuliniai žymenys, raiškos lygių pokyčius. Iš apskaičiuotų IE matyti, kad kolageno I raiška žiurkių plaučių audinyje reikšmingai padidėjo visose eksperimentinėse grupėse (2, 4, 6 ir 8 PAH progresavimo savaitėmis), lyginant su kontroline grupe. III kolageno raiška yra periodiška su reikšmingu padidėjimu po 2 savaičių nuo PAH progresavimo; vėliau kolageno III lygis normalizuojasi iki kontrolinių verčių, o po 6 savaičių nuo PAH progresavimo dar šiek tiek padidėja. TIMP-1 raiška priklauso nuo laiko; po 2 savaičių PAH progresavimo žiurkėms ji padidėja dvigubai, lyginant su sveikų gyvūnų kontroline grupe. Tačiau, priešingai nei vėlesnėse PAH stadijose padidėjęs I ir III kolageno kiekis, TIMP-1 raiška per likusį stebėjimo laikotarpį sumažėja iki kontrolinių verčių.

Remiantis SHG vaizdų analizės rezultatais ir atsižvelgiant į fiziologinį foną bei IHC analizės rezultatus, galima pasiūlyti šiuos PAH vystymosi / progresavimo etapus:

I. Pradžioje patologija vystosi greitai: 2 savaitę μ_1, σ, *OI*, *I* ir *H* padidėja, o *E* ir *L* sumažėja, rodant kolageno kaupimąsi, kolageno skaidulų tempimą ir jų plitimą plaučių audinyje.
II. Organizmas bando reguliuoti ligos progresavimą: 4 savaitę μ_1 ir σ bei aukštų *H* reikšmių sumažėjimas rodo galimą su uždegimu susijusį kolageno struktūros sutrikimą ir kolageno skaidymo sistemos aktyvaciją, kuri turėtų ištaisyti kolageno sintezės ir skaidymo disbalansą.
III. Organizmo fantominis atsigavimas 6 savaitę, nes μ_1 ir σ sumažėjimas ir *H* yra žemi; tačiau g_1 ir g_2 padidėjimas rodo reikšmingą kolageno persiskirstymą, reiškiantį kolageno skaidulų sustorėjimą ir gilų įsiskverbimą į plaučių audinį, todėl tai galėtų būti PAH patogenezės "negrįžimo taškas";
IV. Galiausiai, 8 patologijos savaitę įvyksta visiškas audinių nepakankamumas, o tai yra reikšmingo kolageno kiekio padidėjimo ir kolageno pluoštų sustorėjimo rezultatas, apibūdinamas μ_1, σ, *I*, *H* padidėjimu bei *E* ir *L* sumažėjimu.

Apskritai SHG vaizdavimas pateikia išsamų morfologinių kolageno pokyčių vaizdą PAH sukeltos fibrozės progresavimo metu. Skirtingų FOS ir SOS parametrų evoliucija rodo tuos pačius būdingus kolageno skaidulų struktūros ir tinklo organizacijos pokyčius, kurie taip pat atitinka IHC rezultatus. Tai nepatenka į šios disertacijos apimtį, tačiau surinkus daugiau duomenų iš didesnio mėginių skaičiaus, SHG vaizdų analizė taip pat galėtų pateikti patikimas skirtingų PAH patogenezės etapų žymas.

Tai buvo apibendrinta pirmajame disertacijos teiginyje: *Plataus lauko SHG plaučių audinio pjūvių vaizdų statistinė analizė atskleidžia ir kokybiškai bei kiekybiškai apibūdina būdingus kolageno organizacijos,*

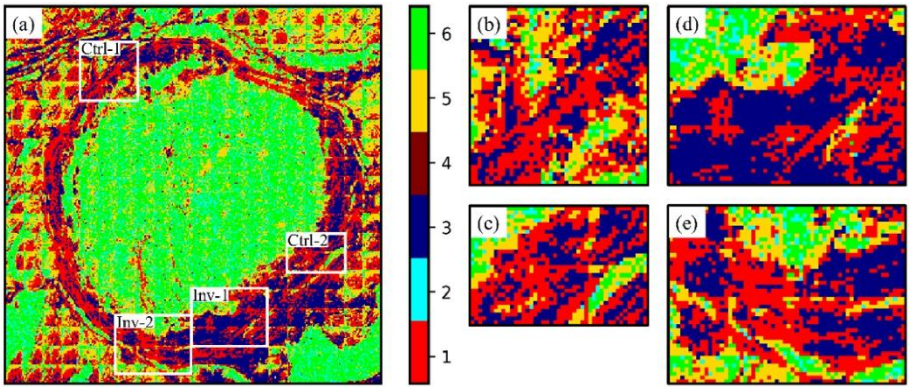
morfologijos ir kolageno kiekio pokyčius, susijusius su skirtingais PAH etapais.

7.7.2 ML pagrindu atlikta PTC kapsulės invazijos kolageno ultrastruktūros analizė, pagrįsta plataus lauko PSHG vaizdavimu

Šiame skyriuje aprašomas PSHG taikymas tiriant kolageno ultrastruktūros pokyčius kapsulės invazijos ir nepažeistos kolageno kapsulės aplink PTC srityse naudojant vienos ašies kolageno skaidulų molekulinį modelį ir įrodant neprižiūrimų ML algoritmų naudojimo PSHG vaizdų analizei privalumus. Šio skyriaus rezultatai buvo paskelbti ***Paper B*** ir pateikti **6 konferencijoje**.

Siekiant dar labiau vizualiai pagerinti galimų mikroinvazijos vietų identifikavimą, buvo atlikta analizė naudojant neprižiūrimus ML metodus. Duomenų klasifikavimas iš visų su poliarizacija susijusių parametrų žemėlapių, išskyrus φ, lėmė duomenų segmentavimą į $k = 6$ klasterius. Bendras k-vidurkiais klasifikuotų sumažinto atrankos vaizdų skaičius buvo 90000. Šių vidutinių parametrų žemėlapių duomenys buvo standartizuoti naudojant "Robust Scaler" algoritmą.

Gautas viso mazgo klasterio žemėlapis rodo skirtingų klasterių charakteristikų parametrų reikšmių derinių erdvinį pasiskirstymą *7.3a paveikslas*.



7.3 paveikslas Su poliarizacija susijusių parametrų duomenų klasterizavimo žemėlapis. (a) visas skydliaukės mazgas; (b) Ctrl-1; (c) Ctrl-2; (d) Inv-1; (e) Inv-2. Perspausdinta iš [***Paper B***].

Vizualiai analizuojant visą kapsulę, ją daugiausia sudaro 1 klasteris, kuris pasiskirstęs ir už jos ribų; 3 klasteris daugiausia yra vidinėje kapsulės dalyje ir aptinkamas tik kai kuriose specifinėse vietose; 5 klasteris daugiausia padengia vidinę ir išorinę kapsulės puses, o tokie kolagenai tikriausiai sudaro daugybę folikulų grupes supančių pertvarų. Tolesnis klasterių susidarymo nepažeistoje kapsulėje tyrimas ROIs *Ctrl-1* ir *Ctrl-2* *7.3b,c paveikslai* ir

mikroinvazijos vietose ROIs *Inv-1* ir *Inv-2* *7.3d,e paveikslai* atskleidžia aiškius kiekybinius klasterių formavimosi skirtumus *7-1 lentelė*. Svarbus invazinis bruožas yra padidėjęs 3 klasterio paplitimas ir sumažėjęs 1 bei 5 klasterių paplitimas. Todėl greita vizualinė klasterio žemėlapio apžiūra palengvina mazgo kapsulės sričių, kurios gali sukelti įtarimą dėl invazijos, nustatymą. 1, 3 ir 5 klasterių kolagenų vidutiniai spiraliniai žingsnio kampai yra 42.5°, 43.9° ir 37.9° atitinkamai; tai rodo didelius kolageno molekulinės struktūros pokyčius pažeistoje kapsulėje, lyginant su kontrolinėmis sritimis *7.3 paveikslas*. Tokie trigubos spiralės struktūros pokyčiai gali turėti įtakos kolageno sąveikai su normaliomis ir vėžio ląstelėmis [195]. Neseniai buvo įrodyta, kad kolagenai su sandariai uždaryta triguba spirale (su 43.9° spiraliniu žingsnio kampu [195] kaip 3 klasteryje) pasižymi didesniu jungimosi prie vėžinių ląstelių efektyvumu, lyginant su normaliomis ląstelėmis, ir todėl gali skatinti vėžio progresavimą ir metastazes [196].

7-1 lentelė Su kolagenu susijusių klasterių procentinė dalis nepažeistoje kapsulėje ir invazijos vietose. Paimta iš [***Paper B***].

Klasterio numeris	Kontrolė [%]	Invazija [%]
1	47,1	32,7
3	31,1	50,7
5	21,8	16,6

Didesnis kampinis išsiplėtimas pastebėtas įtartinų ir invazijos vietų aplinkoje. Priešingai, kiti čia tirti su kolageno struktūra susiję parametrai rodo potencialą išryškinti kolageno pokyčius pikselių lygmenyje (t. y., $\chi^{(2)}$ elementų ir spiralinio žingsnio kampo santykis). Jautrumo santykių biologinę reikšmę reikėtų interpretuoti atsargiai ir atsižvelgiant į kelis veiksnius, įskaitant teorinį kolageno modelį, vaizdo skiriamąją gebą ir kitus veiksnius. Jei visos kolageno molekulės fibrilėse yra išsidėsčiusios ta pačia kryptimi, antros eilės jautrumas turėtų atspindėti pirmos eilės hiperpoliarizuojamumą – tenzorių, apibūdinantį molekulės atsaką į veikiantį elektrinį lauką antros eilės netiesinių optinių efektų pavidalu. Todėl jautrumo santykis turėtų būti panašus į pirmos eilės hiperpoliarizuotumo santykį. Be to, kolageno molekulės žingsnio kampą galima įvertinti susiejant $\chi^{(2)}$ elementų santykį su emituojančio dipolio orientacijos kampu. Reikia atsargiai interpretuoti rezultatus ir tikrinti, ar yra tenkinamos pradinės prielaidos (pvz., Kleinmano simetrija). Nors šiame tyrime χ_{31}/χ_{15} skirstiniai yra artimi vienetui, yra didelių skirtumų ir absoliutūs rezultatai turi būti vertinami atsargiai. Vienas iš $\chi^{(2)}$ elemento santykių, tai yra, χ_{33}/χ_{31}, suteikia įžvalgų apie kolageno fibrilių anizotropiją židinio tūryje ir yra žinomas kaip anizotropijos parametras [197]. Literatūroje nurodomas χ_{33}/χ_{31}reikšmės nuo 1,2 ir 2,6, priklausomai nuo audinio tipo (mažesnės organizuoto kolageno vertės sausgyslėse su tiesiomis fibrilėmis židinio plokštumoje) ir naudojamos erdvinės skiriamosios gebos [198]. Priešingai, kolagenas parodė didesnes jautrumo vertes

audiniuose, kurių molekulė yra orientuota didesniu kampu fibrilėje [199]. Nors gautų jautrumo santykių interpretavimas biologiniame kontekste yra sudėtingas neatsižvelgiant į daugybę kintamųjų, jautrumo santykių skirtumai gali suteikti prasmingesnių įžvalgų apie patologijos pokyčius, vykstančius esant pastovioms techninėms sąlygoms.

Tai buvo apibendrinta antrajame disertacijos teiginyje: *iš plataus lauko poliarizacijos būdu išskaidytų SHG vaizdų iš viso skydliaukės mazgų pjūvių išgautų cilindrinių modelių parametrų k-vidurkių klasterizacija leidžia atskirti kapsulės invazines sritis nuo nepažeistų vėžio ląsteles supančių kapsulės sričių, atskleidžiant kolageno ultrastruktūros modelius*.

7.7.3 Kapsulinės invazijos skydliaukės mazgeliuose diagnostika ML pagrindu naudojant plataus lauko SHG mikroskopiją

Šiame skyriuje vaizduojamas neprižiūrimų ML algoritmų taikymas išsamiai PTC supančių kolageno kapsulių plataus lauko SHG vaizdų analizei, kuri leidžia interpretuoti kolageno struktūros pokyčius kapsulėje ir aptikti mikroinvazijos sritis arba sritis, kurioms reikalingas papildomas tyrimas. Rezultatai buvo paskelbti ***Paper C*** ir pateikti **3, 4, 5, 8** bei **10**. **konferencijose**.

Iš kiekvieno užfiksuoto SHG vaizdo buvo apskaičiuoti šie intensyvumo ir tekstūros parametrai: μ_1, σ, g_1, g_2 (FOS); *E*, *I*, *C*, *L*, *H* (SOS); *SRE*, *LRE*, *GLN*, *RLN*, *RP* (HOS).

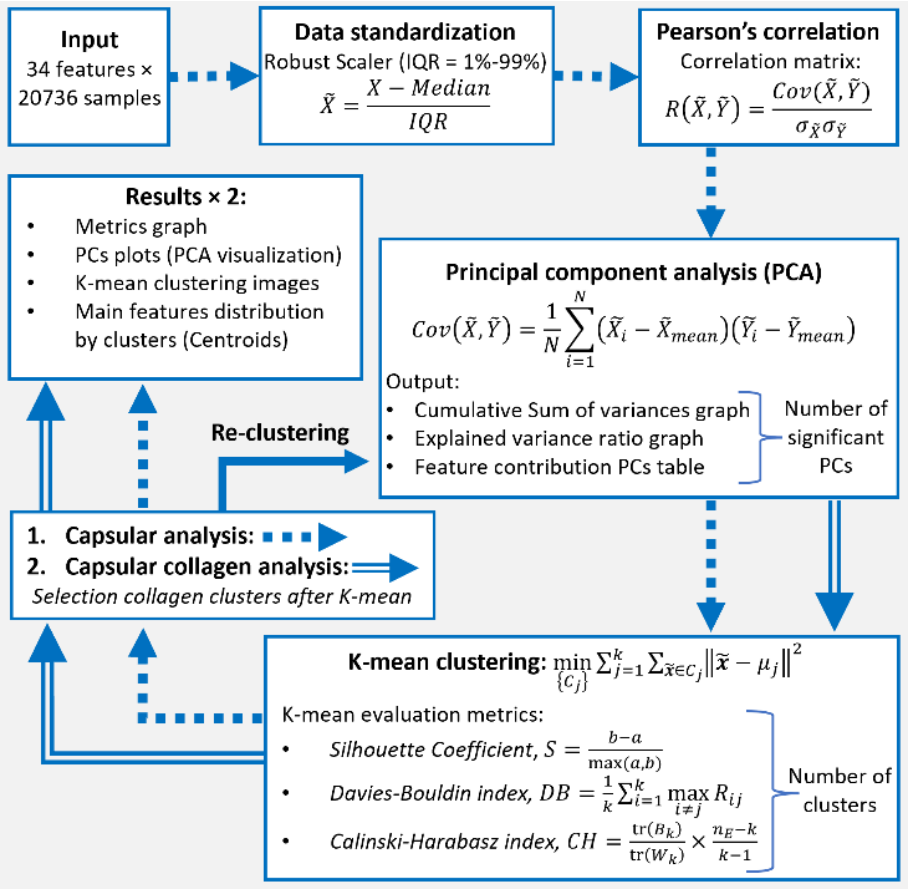
Pagrindiniai viso duomenų analizės ciklo etapai schematiškai pavaizduoti *7.4 paveiksle*. Iš viso vaizdo analizėje buvo atsižvelgta į 4 (FOS) + 5 × 5 (SOS) + 5 (HOS) = 34 intensyvumo ir tekstūros parametrus.

Norint klasifikuoti įrašytus SHG vaizdus į skirtingas kategorijas pagal jų intensyvumo ir tekstūros parametrus, buvo naudojamas neprižiūrimas *k*-vidurkių mašininio mokymosi metodas. Klasterinė analizė buvo atlikta dviem etapais: pirma, atskirti PTC kolageno kapsulę nuo aplinkinių audinių, o tada atskleisti galimus kolageno struktūros skirtumus tarp anotuotos kapsulės invazijos sričių ir nepažeistos kapsulės. Patologas taip pat patvirtino klasterizacijos rezultatus, kad būtų atmestas nepakankamas arba per didelis klasterizavimas. Iki *k*-vidurkių duomenys buvo apdoroti naudojant PCA, siekiant sumažinti turimų parametrų rinkinio sudėtingumą ir atsekti, kurie parametrai turi didžiausią reikšmę atskiriant vaizdus į skirtingas klases. Tokiu būdu klasterizavimas buvo atliekamas kompiuteriuose, o ne tiesiogiai su parametrais.

Išsamūs *k*-vidurkių klasterizavimo rezultatai, kai k = 8, PTC mazgų pjūviams pavaizduoti *7.5c, f, i, l paveiksle*.

Iš 8 klasterių du (6 ir 7 klasteriai) apima išskirtinius duomenis, kurie yra reti ir pasižymi kraštutinėmis atskirų parametrų vertėmis, lyginant su likusiu duomenų rinkiniu. Stiklinės plokštelės ir foninius SHG signalo lygius generuojančių vėžio ląstelių SHG vaizdai patenka į 5 klasterį.

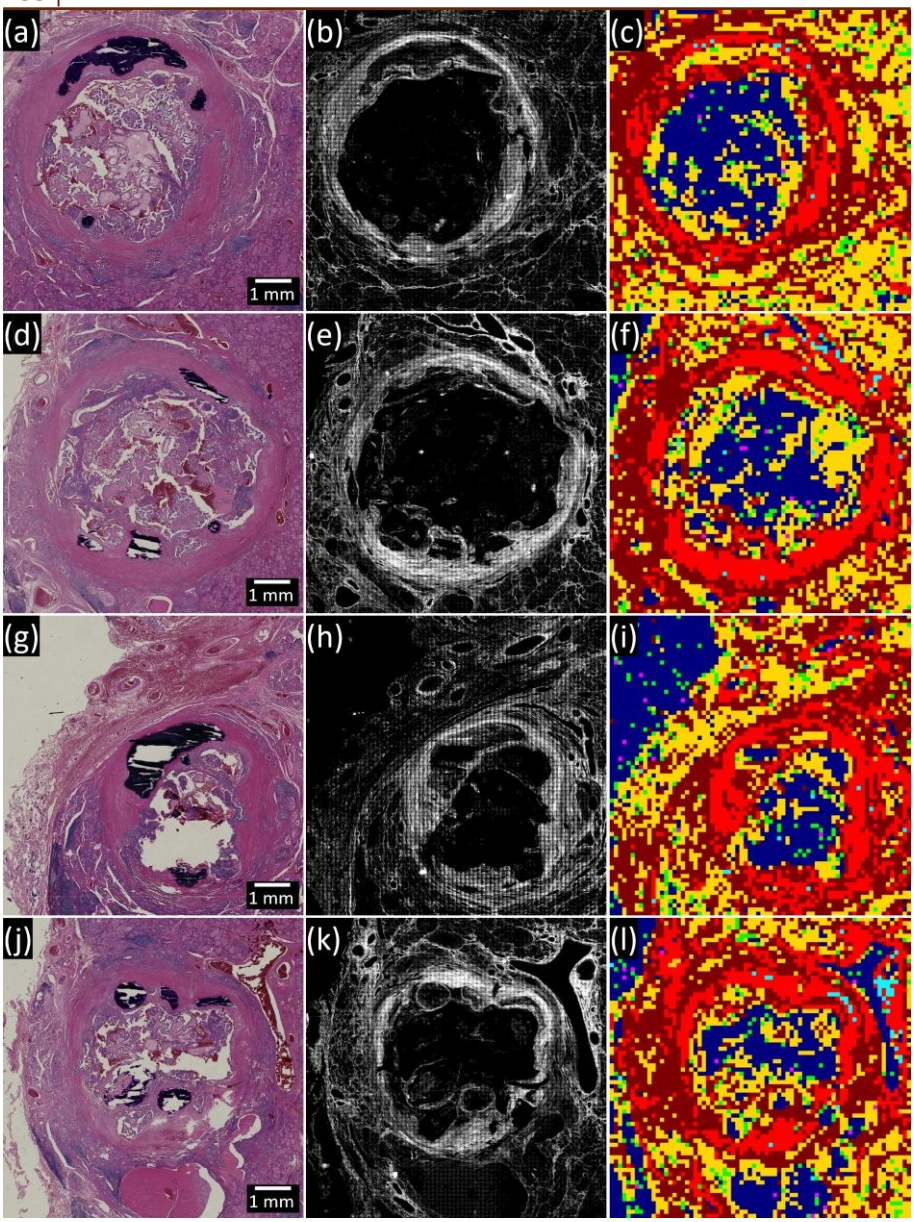
Tada, nagrinėjant skirtingų klasterių lokalizaciją mėginiuose (*7.5 paveikslas*), atrodo, kad likusius klasterius, nors ir šiek tiek sąlyginai, galima priskirti kolagenui aplink didelius indus šalia kapsulės (0 klasteris, 93 SHG vaizdai, *7.5 paveikslas*), PTC mazgą supančiai kolageno kapsulei ir kolageno plitimui į normalų audinį (1 ir 2 klasteriai), normaliems folikulams (3 klasteris) ir galimai uždegiminiam audiniui (4 klasteris).



7.4 paveikslas Duomenų apdorojimo seka ir ML pagrindu atlikta analizė. Perspausdinta iš [***Paper C***].

Siekiant sutelkti dėmesį į specifinius kolageno kapsulės aplink PTC pokyčius, buvo atrinkti du daugiausiai su kapsule susiję kolageno klasteriai (1 ir 2) ir atliktas jų papildomas klasterizavimas. Visi kiti duomenys (0, 3–7 klasteriai) nebuvo įtraukti į analizę ir gautuose klasterių žemėlapiuose pažymėti juoda spalva.

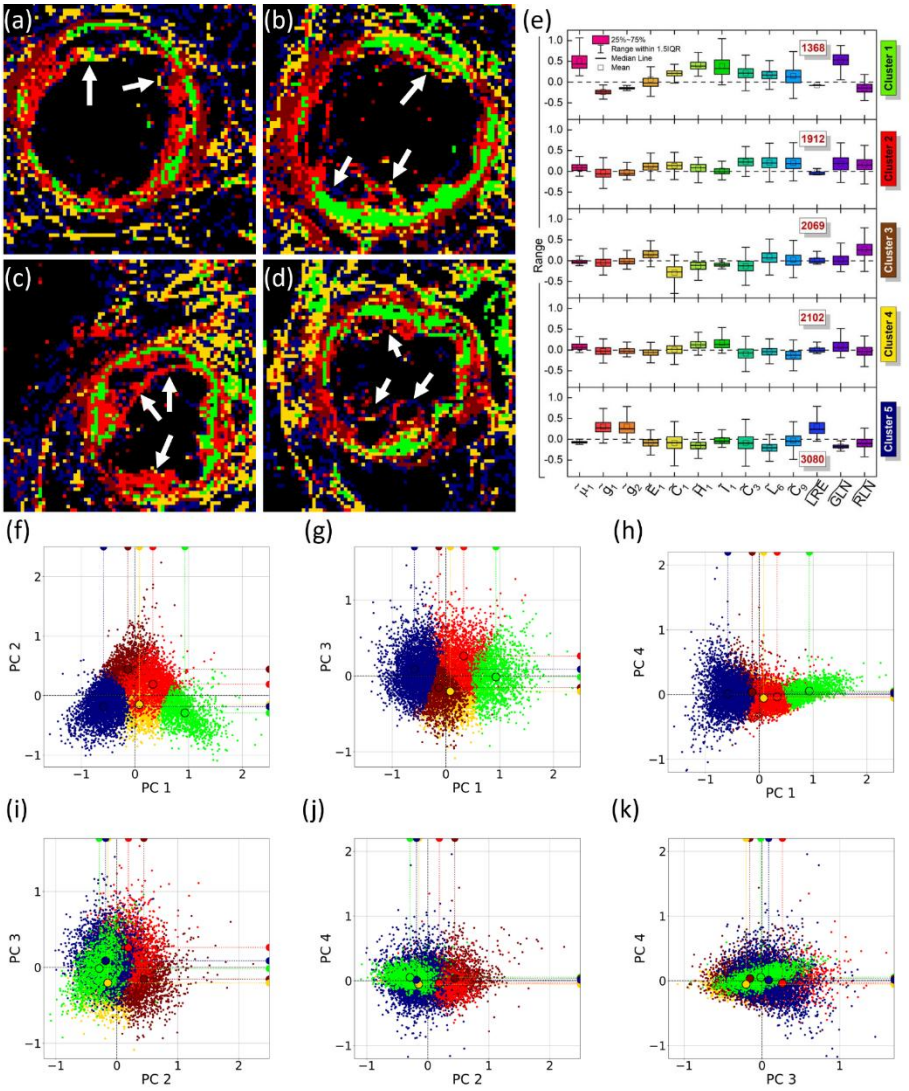
Pakartotiniai kapsulinio kolageno pasiskirstymo klasteriniai žemėlapiai, balų diagramos ir atitinkami centroidai pavaizduoti *7.6 paveiksle*.



7.5 paveikslas Analizuotų PTC mazgų pjūvių klasterizacijos *k*-vidurkiai: (a, d, g, j) – šviesaus lauko vaizdai; (b, e, h, k) – SHG vaizdai; (c, f, i, l) – klasterizacijos k-vidurkiai, k = 8. Paimta iš [***Paper C***].

Klasterių žemėlapiai atskleidžia reikšmingą kolageno struktūros heterogeniškumą kapsulėje aplink PTC (*7.6a-d paveikslas*). Kapsulę daugiausia sudaro trys klasteriai (1-3). 1 klasteris (žalias *7.6 paveiksle*) sudaro pagrindinę kapsulės dalį ir jo praktiškai nėra normaliuose audiniuose su folikulais. Nors jis paprastai sudaro ištisinį visos kapsulės branduolį, yra aiškiai atpažįstamų sričių, kur jį pakeičia 2 arba 3 klasteris (atitinkamai

raudonas ir rudas *7.6 paveiksle*) arba jų mišinys. 4 klasteris (geltonas *7.6 paveiksle*) yra šalia kapsulės iš išorės. 5 klasteris (tamsiai mėlynas *7.6 paveiksle*) beveik visas yra už kapsulės ribų ir tikriausiai sudaro normalių folikulų pertvaras. Kalcifikacijas rodo kolagenai, daugiausia priskirti 2, 4 ir 5 klasteriams (pažymėti rodyklėmis *7.6a-d paveiksle*).



7.6 paveikslas Tik su kolagenu susijusių klasterių klasterizacijos *k*-vidurkiai. (a-d) 4 analizuotų PTC mazgų pjūvių klasterių žemėlapiai. (e) Centroidai, sudaryti iš standartizuotų ir kiekvieną klasterį atitinkančių parametrų. (f-k) Taškų grafikai. Spalvoti apskritimai rodo klasterio centroido (klasterio masės centro) padėtį PC erdvėje; spalvoti puslankiai abscisių ir ordinačių ašyse žymi centroidų projekcijas. Klasterių žemėlapių dydis yra 8.4 mm × 8.4 mm. (f)-(k) punktuose pateikti duomenų taškai ir (a)-(d) punktuose pateikti klasteriai yra

tokios pat spalvos kaip (e) punktuose pateikti klasteriai bei yra identiškai suskirstyti į kategorijas. Perspausdinta iš [***Paper C***].

Kalbant apie gydymą, invazijos sričių buvimas ir mastas yra labai svarbūs chirurginiu požiūriu [216]. Pavyzdžiui, mazgas su lokalizuota invazija gali pareikalauti platesnės rezekcijos nei vien lobektomija ir tokiu atveju bus reikalinga visiška tiroidektomija. Išplitusios invazijos, pvz., išplitimo į skydliaukę ar metastazių limfmazgiuose, atvejais gali prireikti adjuvantinio gydymo, pvz., radioaktyviojo jodo abliacijos, siekiant paveikti likusį piktybinį audinį. Tai nebūtų taikoma visiškai kapsule apimtų mazgų be invazijos požymių atveju. Galiausiai, invazijos buvimas taip pat turi įtakos pacientų stebėjimui po operacijos ir galimiems adjuvantiniams gydymo būdams. Invazinių arba didelės rizikos navikų atveju reikalinga agresyvesnė stebėsena, kad būtų galima anksti nustatyti recidyvą ar metastazes. Taigi, bet kokia pažanga didinant invazijos nustatymo tikslumą pagerina diagnozės ir gydymo kokybę bei pacientų išgyvenamumą. Šiame tyrime pateikti rezultatai ir siūlomas ML pagrįstas metodas svariai prisideda prie šios pažangos.

Nors tradicinė histopatologija išlieka auksiniu kapsulės invazijos diagnostikos standartu, šiame tyrime naudojama SHG mikroskopija suteikia keletą techninių privalumų duomenų kaupimui. SHG mikroskopija suteikia informacijos apie kolageną skydliaukės mazgo kapsulėje.

Apibendrinant galima teigti, kad siūlomas automatinio ML pagrindu veikiančio su kolagenu susijusių SHG vaizdų atrankos metodas ir išvados apie PTC kapsulės heterogeniškumą labai tikėtinai gali tapti naujų veiksmingų automatinės diagnostikos modelių, pagrįstų neprižiūrimais ML algoritmais, kūrimo pagrindu, o vėliau juos išplėsti į prižiūrimus ML modelius. Dabartinio tyrimo rezultatai rodo, kad tradicinis metodas, kai ROI pasirenkami rankiniu būdu prižiūrimų ML modelių mokymui, yra nenuoseklus dėl PTC kapsulių heterogeniškumo. Jų heterogeniškumas, kurio negalima visiškai aptikti rankinio patikrinimo metu, gali sukelti didžiausią paklaidą pasirinktų "nepažeistos" (kontrolinės) kapsulės ROI rinkinyje, nes juose yra ir normalių, ir neidentifikuotų "įtartinų" kapsulės sričių. Dėl to prižiūrimo ML modelio negalima apmokyti teisingai, nes vienas iš mokymo duomenų rinkinių ("sveika" kapsulė) yra dviprasmiškas. Šiame tyrime siūlomas metodas, potencialiai papildytas THG, TPEF ir CARS "ląsteliniu kontekstu", gali būti efektyvus automatizuotas vėžio diagnostikos metodas

Tai buvo apibendrinta trečiajame disertacijos teiginyje: *Neprižiūrimas mašininis mokymasis pagerina SHG vaizdų analizę, atskleidžia papilinės skydliaukės karcinomos kapsulės tekstūrinį heterogeniškumą ir leidžia nustatyti kapsulės invaziją, papildomo tyrimo reikalaujančias sunkiai atskiriamas mikroinvazijas ir sritis, remiantis konkrečiais vaizdo parametrų rinkiniais.*

7.7.4 Prižiūrimas ML skydliaukės karcinomos diagnozei naudojant plataus lauko SHG mikroskopiją

Šios skyriaus rezultatai buvo pateikti ***Paper D*** ir **7** bei **9 konferencijose**.

Visi FTC ir PTC mėginiai buvo vaizduojami naudojant SHG mikroskopijos įrenginį. Siekiant užtikrinti įvairiapusį imties aprašymą, 34 intensyvumo ir tekstūros požymiai (4 FOS, 25 SOS ir 5 HOS), paimti iš kiekvieno 117 µm × 117 µm SHG vaizdo, buvo naudojami tolimesnei analizei. Tačiau ne visi požymiai yra labai gerai diferencijuojami ir svarbūs tikslui, o tai gali turėti įtakos klasifikavimo rezultatams. Norint įvertinti požymių pertekliaus ir nereikšmingumo įtaką klasifikacijai, požymių atranka buvo atlikta naudojant RFECV-LinearSVC, po kurio sekė MIFS. Kiekvieno žymeklio taisymo metodo funkcijų pasirinkimo rezultatai pavaizduoti *7-2 lentelėje*.

7-2 lentelė RFECV-LinearSVC ir MIFS parinktos funkcijos kiekviename etikečių taisymo metode. Paimta iš [***Paper D***].

Žymeklio taisymo metodas	Padalijimas	RFECV-LinearSVC neįtraukti; (neįtraukiant f_i, No.)	MIFS neįtraukti; (neįtraukiant f_i, No.)	Likusių f_i *skaičius*
I. Susijęs su audiniais	70/30	E_{12}, C_1; (2)	C_{12}, I_9, C_6, I_3, C_3; (5)	27
	80/20	*RLN*, I_9, H_1, C_1; (4)	C_{12}, C_6, I_3, C_3; (4)	26
	90/10	*RLN*, E_{12}, I_9, L_6, E_6, H_1, E_1; (7)	C_{12}, C_6, I_3; (3)	24
II. Susijęs su kapsule	70/30	*LRE*, L_{12}, E_{12}, E_9, H_6, E_6, I_3, E_1, μ_1; (9)	(0)	25
	80/20	*GLN*, *LRE*, L_{12}, C_{12}, E_{12}, E_9, H_6, E_6, I_3, E_3, E_1, μ_1; (12)	(0)	22
	90/10	*LRE*, E_{12}, E_6, μ_1; (4)	(0)	30
III. Daugiaklasis, atsižvelgia į kapsulės heterogeniškumą	70/30	E_9, E_6; (2)	(0)	32
	80/20	I_9; (1)	(0)	33
	90/10	I_9, E_9, E_6, E_3; (4)	(0)	30

Išnaša: Apatinis indeksas rodo žingsnį (px), naudojamą GLCM skaičiavimui. f_i – požymiai.

Su audiniais susiję SHG vaizdai buvo atskirti nuo su audiniais nesusijusių vaizdų naudojant iš SHG vaizdų išgautų tekstūros požymių vektorių PCA. Dvejetainis *k*-vidurkių klasterizavimas, pagrįstas pirmaisiais penkiais kompiuteriniais skaičiais, apimančiais daugiau nei 92 % duomenų dispersijos, leido segmentuoti audinio pjūvio vaizdą į su audiniais ir ne audiniais susijusius SHG vaizdus [177]. Toks klasterizavimas atskiria su audiniais ir ne audiniais susijusius taškus į dvi aiškiai apibrėžtas grupes. Klasifikatoriaus optimizavimui buvo naudojami pilni arba redukuoti audinių SHG vaizdų požymių vektoriai (per RFECV-LinearSVC/MIFS).

Didžiausios tikslumo vertės buvo pasiektos naudojant 90/10 (mokymo/patvirtinimo) duomenų padalijimą ir klasifikatoriaus optimizavimą naudojant pilnus požymių vektorius (*7-3 lentelė* "I. Susijęs su audiniais"). Visi ML modeliai (RF, XGBoost ir LightGBM) demonstruoja perteklinio pritaikymo požymius, rodančius duomenų nutekėjimą, dėl to mokymo duomenų rinkinyje gauti pernelyg optimistiniai rezultatai, tačiau patvirtinimo

duomenų rinkinyje gauti prasti rezultatai. Nors LR klasifikatoriaus tikslumas buvo gana patenkinamas, kiti jo rodikliai buvo gerokai blogesni nei kitų modelių. MLP ir C-SVC pademonstravo geriausius rezultatus patvirtinimo rinkinyje, o C-SVC pranoko MLP visuose rodikliuose. Mažesnės MLP atkūrimo ir F1 balo vertės rodo didesnį klaidingai neigiamų (PTC, prognozuojamas kaip FTC) ir klaidingai teigiamų (FTC, prognozuojamas kaip PTC) rezultatų dažnį, lyginant su C-SVC. RFECV-LinearSVC ir vėliau MIFS taikymas, siekiant atrinkti svarbias funkcijas ir neįtraukti nereikalingų funkcijų, pagerino MLP rodiklius ir sumažino C-SVC rodiklius.

7-3 lentelė Optimizuoto modelio našumo skaitmeninis įvertinimas (remiantis maksimaliu tikslumu), gautas atliekant 90/10 duomenų skaidymo mokymą / patvirtinimą MLP ir C-SVC modeliams. Perspausdinta iš [***Paper D***].

Žymeklio taisymo metodas	ML modelis	Tikslumas (patvirtinimas), %	Tikslumas (mokymas), %	Atkūrimas	Artumas	F1	AUC	Tikslumas (FTC testas), %	Tikslumas (PTC testas), %	Tikslumas (PTC* testas), %	Pastaba
I. Susijęs su audiniais	MLP	75.88	76.77	0.588	0.781	0.671	0.850	54.57	56.82	62.40	+++
	MLP*	78.00↑	79.00↑	0.687	0.763	0.723	0.862	63.94	44.69↓	49.91	↑++↓
	C-SVC	81.71	87.09	0.767	0.789	0.778	0.881	72.13	45.09	48.84	++
	C-SVC*	80.31↓	82.85↓	0.736	0.780	0.757	0.870	70.76↓	42.01↓	46.13	↓
II. Susijęs su kapsule	MLP	81.94	83.95	0.745	0.798	0.771	0.898	64.67	38.05	40.91	++
	MLP*	80.00	82.15	0.707	0.781	0.742	0.881	67.89	44.93	46.61	++↑
	C-SVC	82.07	88.36	0.770	0.785	0.778	0.901	71.35	43.28	46.19	++
	C-SVC*	82.20	87.61	0.748	0.802	0.774	0.886	65.69	52.74	56.16	+++↑

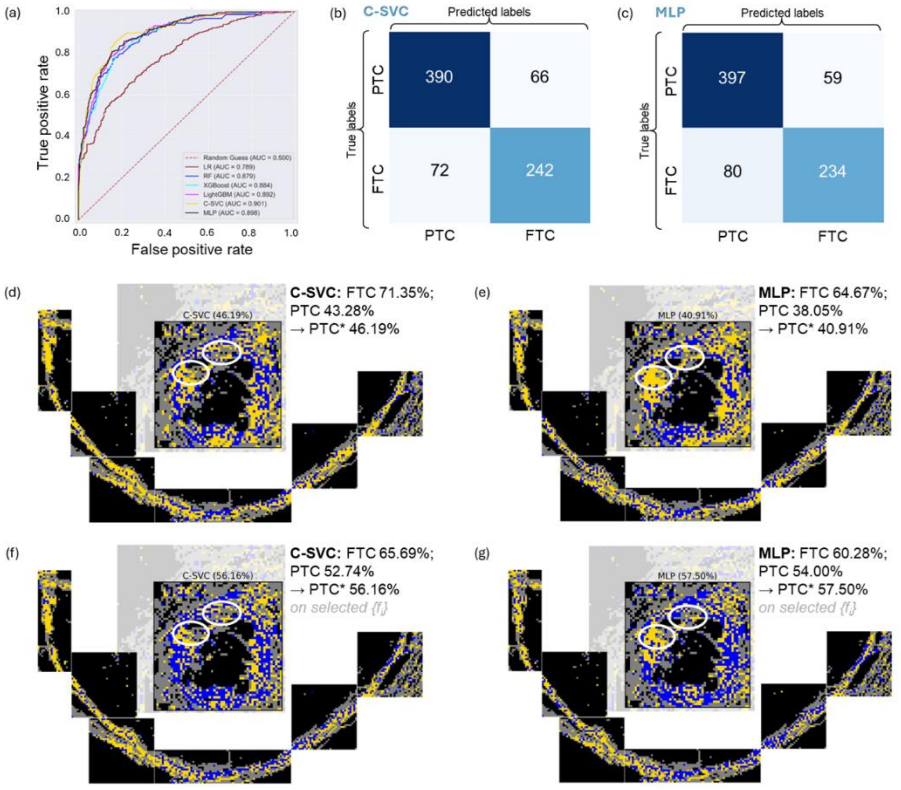
Išnaša: * – rodo, kad funkcijų parinkimas buvo atliktas prieš optimizuojant naudojamų klasifikatorių hiperparametrų konfigūracijas; geras tikslumo patvirtinimas / mokymas, geras atkūrimas / artumas / F1 / AUC, prastas realiam testų rinkiniui; +++ geras tikslumo patvirtinimas / mokymai, "klasifikuojamas" realiam testų rinkiniui. (↑) rodyklė rodo modelio našumo pagerėjimą, o (↓) rodyklė rodo modelio našumo sumažėjimą pašalinus nereikalingas ir nereikšmingas funkcijas.

Nežinomo testo rinkinyje MLP teisingo klasifikavimo rodiklis yra šiek tiek didesnis nei 50%, o C-SVC pasižymėjo maža PTC atskyrimo galia, tačiau gerai atskyrė FTC. Vizualinė klasifikuotų PTC vaizdų patikra parodė, kad normalus audinys aplink apskritą PTC kapsulę dažnai buvo klaidingai klasifikuojamas kaip FTC, todėl buvo gauti klaidingai neigiami rezultatai. Analizės srities apribojimas arčiau PTC kapsulės pagerino klasifikavimo našumą, padidindamas teisingų teigiamų rezultatų dalį iki 48,84% C-SVC atveju ir 62,40% MLP atveju.

Siekiant sumažinti aplinkinių audinių sukeltą PTC ir FTC mėginių duomenų persidengimą, buvo atrinkti tik mazgų kapsulių SHG vaizdai. Siekiant automatizuoti žymeklių triukšmo mažinimą, kapsulės vaizdų atskyrimas buvo atliktas naudojant neprižiūrimą ML metodą [177]: PCA buvo atliktas SHG duomenų rinkiniui, o vėliau buvo atliktas dvejetainis *k*-vidurkių klasterizavimas su gautais PC. Nors nei rankinis ženklinimas, nei šis metodas neužtikrina tobulo kapsulių atskyrimo, požymių dispersijos skirtumais pagrįstas *k*-vidurkių klasterizavimas suteikia objektyvesnį segmentavimą nei vizualinis patikrinimas. Tolesnė analizė atlikta su kapsulėmis susijusiais SHG vaizdais.

Nekapsulinių SHG vaizdų filtravimas žymiai sumažino mokymo / patvirtinimo duomenų rinkinius ir sukėlė nedidelį, bet valdomą klasės disbalansą. Lyginant su audiniais susijusiu SHG vaizdų duomenų rinkiniu, RFECV-LinearSVC ženkliai sumažino požymių skaičių, o MIFS nepašalino nė vieno. MIFS pašalintų požymių nebuvimas rodo, kad visos RFECV-LinearSVC parinktos savybės buvo svarbios PTC ir FTC kapsulių atskyrimui.

Skirtingai nuo visų audinių metodo, tikslumas išlieka pastovus esant 70/30, 80/20 arba 90/10 (mokymo/patvirtinimo) padalijimams, o du geriausi (MLP ir C-SVC) yra pavaizduoti *7-3* lentelė "II. Susijęs su kapsule". Apskritai modeliai veikia geriau nei tie, kurie buvo apmokyti naudojant visus su audiniais susijusius duomenis (*7.7a paveikslas*). Tačiau kiti modeliai (RF, XGBoost ir LightGBM) išlieka per daug pritaikyti.



7.7 paveikslas Kapsulėms skirtų duomenų rinkinių, kai santykis (90/10), ML modelių našumas: (a) Visų ML modelių ROC kreivės, (b) C-SVC painiavos matrica; (c) MLP painiavos matrica; (d) C-SVC klasifikavimas, atliktas su nauju duomenų rinkiniu (testo rinkiniu); (e) MLP klasifikavimas, atliktas su nauju duomenų rinkiniu (testo rinkiniu); (f) C-SCV (apmokytas su sumažintu požymių rinkiniu) klasifikavimas, atliktas su nauju duomenų rinkiniu (testo rinkiniu); (g) MLP (apmokytas su sumažintu požymių rinkiniu, 70/30) klasifikavimas, atliktas su nauju duomenų rinkiniu (testo rinkiniu). Mėlynos spalvos žymeklių vaizdai klasifikuojami kaip PTC, geltonos – kaip FTC. PTC

procentinė dalis rodo teisingai numatytų PTC plytelių dalį PTC mėginyje, įskaitant aplinkinius audinius. Žvaigždute žymima teisingai numatytų PTC plytelių dalis PTC mėginyje, išskyrus aplinkinius audinius. Balti apskritimai žymi kalcifikacijų vietas. Perspausdinta iš [***Paper D***].

MLP ir C-SVC atveju patvirtinimo rinkinio tikslumas atitinkamai pagerėja iki 81,94% ir 82,07% (*7-3 lentelė* "II. Susijęs su kapsule", *7.7 paveikslas*). Be to, MLP modelio atkūrimo ir F-1 balų rodikliai padidėja, o tai rodo geresnį našumą, kai jis apmokomas su kapsulėmis susijusiais duomenų rinkiniais.

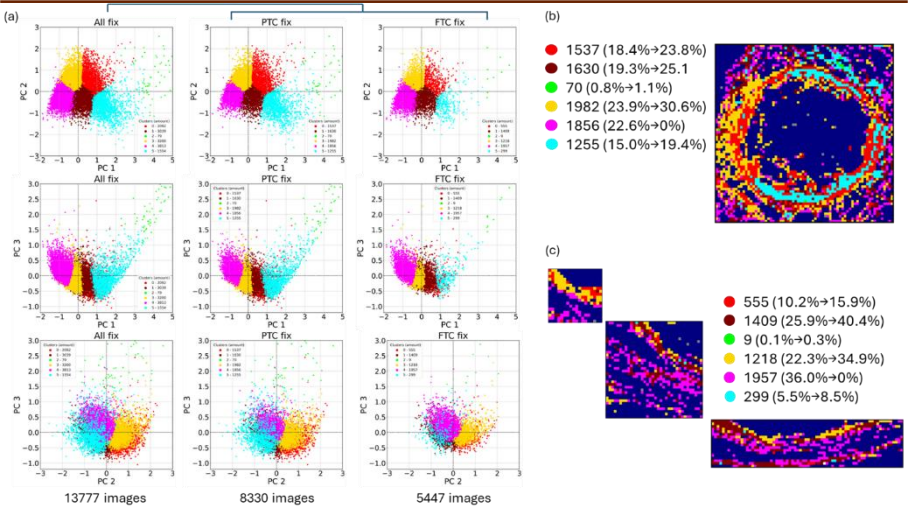
Tiek MLP, tiek C-SVC ne tik pagerino savo rezultatus patvirtinimo rinkinyje, bet ir teisingai klasifikavo FTC mėginius visuotiniame bandymų duomenų rinkinyje (*7.7d, e paveikslas*). MLP klasifikavimo tikslumas PTC atveju sumažėja, kai apmokoma naudojant su kapsulėmis susijusius duomenų rinkinius, lyginant su visais su audiniais susijusiais duomenų rinkiniais (*7.7d paveikslas*).

Funkcijų pasirinkimas naudojant RFECV-LinearSVC šiek tiek sumažino visų klasifikatorių tikslumą, tačiau žymiai pagerino našumą nežinomame visuotiniame bandymų duomenų rinkinyje visiems skaidymams (*7-3 lentelė* "II. Susijęs su kapsule", žvaigždutėmis pažymėtos eilutės). Medžių klasifikatoriai ir C-SVC gavo daugiausia naudos iš požymių atrankos, o C-SVC pasiekė atitinkamą 65,69% ir 56,16% tikslumą FTC ir PTC atveju mokymo / patvirtinimo padalijime, pranokdamas visus ankstesnius metodus (*7.7f paveikslas*). MLP, kuris buvo jautrus mokymo duomenų rinkinio dydžiui, pademonstravo geresnę klasifikaciją, pasiekdamas atitinkamai 57,50% ir 60,28% PTC ir FTC klasifikacijoms visuotiniame testų rinkinyje (*7.7g paveikslas*).

Tikslus PTC klasifikavimas išlieka sudėtingas, nepaisant su žymeklių ir požymių slopinimu susijusių patobulinimų. Tai gali būti dėl didelio kolageno ypatybių heterogeniškumo PTC kapsulėje ir tam tikrų PTC ir FTC kapsulės segmentų panašumo, dėl kurio FTC identifikavimas yra tikslesnis, o PTC – silpnesnis (*7.7d-g paveikslas*).

Abiejų tipų karcinomos mazgus supantis audinys nesuteikia svarbios klasifikacijos informacijos pagal SHG vaizdo tekstūros ypatybes. Tačiau tokie tekstūros požymiai kaip LRE, E_{12}, E_6, μ_1, pašalinti atliekant požymių atranką, kai buvo atsižvelgta tik į kapsulės kolageną, greičiausiai paaiškina skirtumus tarp gretimų audinių ir mazgų. Tai rodo, kad gretimi audiniai galėtų sudaryti papildomą klasę, padedančią spręsti klaidingo ženklinimo problemą taikant su audiniais susijusį metodą.

Siekiant įrodyti perinodulinio audinio panašumą, segmentavimas pagal intensyvumo ir tekstūros požymius buvo atliktas, kaip aprašyta [177]. PCA ir daugiaklasio *k*-vidurkių klasterizavimo rezultatai (*7.8 paveikslas*) rodo, kad nors kapsulinis kolagenas yra nevienalytis tiek PTC, tiek FTC, gretimi audiniai yra atskirti vienoje klasėje (rausvai raudona spalva, *7.8b-c paveikslas*).



7.8 paveikslas PTC ir FTC imčių požymių duomenų rinkinių PCA analizė ir jų klasterizavimas naudojant *k*-vidurkius: (a) PC1 ir PC2, PC3 ir PC1 bei PC3 ir PC2 taškų grafikai visiems duomenims ir atskirai PTC bei FTC; (b) atsitiktinių PTC ir FTC mėginių klasterių žemėlapis. Skaičiai rodo kiekvieno klasterio procentinę dalį atitinkamame duomenų rinkinyje. Antrasis skaičius skliausteliuose yra kiekvieno klasterio procentinė dalis kapsulėje. Rausvai raudonos spalvos klasteris buvo priskirtas normaliam kolagenui, supančiam normalius audinių folikulus, todėl buvo pašalintas iš ″kapsulinio kolageno″ klasės ir pridėtas prie atskiros klasės, kurioje derinami stiklelio ir normalaus audinio vaizdai, esantys tiek PTC, tiek FTC mėginiuose. Perspausdinta iš [***Paper D***].

Abi karcinomos kapsulės sudarytos iš tų pačių klasterių, todėl PTC ir FTC kapsulių klasifikavimas apsunkinamas net ir tada, kai gretimi audiniai neįtraukiami į analizę (pvz., *7.7d-g paveikslas*). Nepaisant bendros klasterių sudėties, klasterių santykiai skiriasi priklausomai nuo karcinomos tipo. Visus duomenis atspindintys PCA taškų grafikai (*7.8a paveikslas*) ir segmentuotų SHG skenavimų pavyzdžiai (*7.8b-c paveikslas*) rodo, kad FTC kapsulėse vyrauja rusvos ir geltonos spalvos sankaupos, o PTC kapsulėse – heterogeniškesnės. Pirmasis variantas greičiausiai paaiškina geresnį FTC klasifikavimą naudojant C-SVC klasifikatorių ankstesniuose metoduose, o antrasis variantas tikriausiai lėmė didesnį PTC klaidų lygį.

Šie skirtumai gali pagerinti klasifikavimo rezultatus, nes tiek klasteriai, tiek jų santykiai apibūdina PTC ir FTC mazgų kapsules. Todėl prieš stratifikuotą 10 kartų kryžminį patvirtinimą modelio optimizavimui buvo fiksuotas klasterių santykis, kuris buvo nustatytas visam mokymo duomenų rinkiniui naudojant *k*-vidurkius. Gretimų audinių ir stiklelio SHG vaizdai buvo pridėti prie "netikslinės" klasės, siekiant išvengti išankstinio apdorojimo veiksmų, kuriais siekiama pašalinti SHG vaizdus, kurie nėra svarbūs taikiniui ir todėl gali sukelti žymeklių triukšmą. "FTC", "PTC" ir "netikslinės" klasės buvo subalansuotos prieš klasifikatoriaus optimizavimą, nors išliko tam tikra

duomenų disproporcija. Šis sumažintas duomenų rinkinio dydis, lyginant su visų audinių ir kapsulės metodais, gali turėti įtakos klasifikatoriaus veikimui.

MLP ir C-SVC 70/30 paskirstymo optimizavimo rezultatai pateikiami *7-4 lentelėje*. Šis paskirstymas apima daugiau duomenų patvirtinimo rinkinyje ir išsaugo klasterių santykius, todėl yra reprezentatyvesnis. Klasifikatorių našumas yra panašus į tą, kuris gautas taikant "su kapsulėmis susijusį" žymeklių metodą. Funkcijų atrankos taikymas reikšmingai nepagerina klasifikavimo tikslumo; tai reiškia, kad atsižvelgiant į klasterių santykius mokymo metu, sumažėja funkcijų perteklius, o visos likusios funkcijos išlieka svarbios tikslui.

7-4 lentelė Optimizuoto modelio našumo skaitmeninis įvertinimas (remiantis maksimaliu tikslumu), gautas duomenų skaidymo mokymui / patvirtinimui 70/30 duomenų rinkiniams, atsižvelgiant į klasterių santykį PTC ir FTC imtyse MLP ir C-SVC modeliams. Perspausdinta iš [***Paper D***].

ML modelis	Tikslumas (patvirtinimas), %	Tikslumas (mokymas), %	Artumas (makro)	Atkūrimas (mikro)	F1 (svertinis)	Tikslumas (FTC testas), %	Tikslumas (PTC testas), %	Pastaba
MLP	81.76	83.16	0.816	0.817	0.811	42.98	63.73	++
MLP*	81.32	83.98	0.810	0.813	0.814	65.12	50.22	+++
C-SVC	84.73	89.30	0.843	0.847	0.847	63.70	52.23	+++
C-SVC*	84.80	89.50	0.844	0.848	0.847	68.65	51.66	+++

Nors daugiaklasė klasifikacija reikšmingai nepagerino visuotinių bandymų duomenų tikslumo, neprižiūrima ML segmentacija išryškino PTC ir FTC kapsulių bei gretimų audinių skirtumus, paaiškindama klasifikatoriaus veikimo skirtumus. Gretimame audinyje nėra aptinkamų PTC arba FTC progresavimo požymių ir jį galima pašalinti iš analizės taikant dvejetainį k-vidurkį (II metodas) arba laikyti atskira klase daugiaklasėje klasifikacijoje (III metodas). Panašūs heterogeniniai kolageno modeliai PTC ir FTC kapsulėse apsunkina klasifikavimą, o PTC kalcifikacijos, kurios tekstūros savybėmis skiriasi nuo PTC kapsulės, visų klasifikatorių yra klaidingai klasifikuojamos kaip FTC. Mažesnis FTC heterogeniškumas, lyginant su PTC kapsulėmis, leidžia C-SVC atskirti PTC ir FTC, o tuo tarpu kitiems klasifikatoriams sunku atlikti perteklinį pritaikymą arba sumažinti duomenų dydį (MLP).

Tai buvo apibendrinta ketvirtajame disertacijos teiginyje: *Prižiūrimas mašininio mokymosi modelis C-SVC leidžia 84,73 % tikslumu diferencijuoti papilines ir folikulines skydliaukės karcinomas, remiantis SHG vaizdavimu.*

7.8 Išvados

1. Plataus lauko SHG mikroskopija kartu su kiekybine vaizdų analize yra patikimas ir be žymeklių veikiantis metodas fibroziniam PAH remodeliavimui įvertinti. MCT gydytų žiurkių plaučių audinyje buvo nustatyti nuo laiko

priklausantys kolageno kiekio ir morfologijos pokyčiai. FOS analizė parodė progresuojantį kolageno kaupimąsi, o SOS analizė – perivaskulinio kolageno tinklo tankėjimą ir vėlesnį jo išplitimą į alveolių sritį. FFT analizė taip pat parodė dinaminę skaidulų orientacijos moduliaciją, kuriai būdingas pradinis išsidėstymas, po kurio vėlyvoje ligos stadijoje seka dezorganizacija. Šie rezultatai pabrėžia SHG pagrįstų metodų potencialą neardomajam fibrozės vertinimui sergant PAH ir susijusiomis plaučių ligomis.

2. Plataus lauko PSHG mikroskopijos taikymas vaizduojant ištisus skydliaukės mazgų pjūvius histologiniuose preparatuose leido išskirti kiekybinius parametrus, apibūdinančius kolageno orientaciją ir ultrastruktūrą mazgo kapsulėje. Naudojant cilindrinį kolageno modelį, kapsulės invazines sritis galima veiksmingai atskirti nuo neinvazinių sričių statistine analize ir neprižiūrimu ML. Šis metodas leidžia objektyviai ir atkuriamai įvertinti kolageno ultrastruktūros pokyčius kapsulės invazijos metu skydliaukės navikuose ir gali būti pagrindas kuriant automatines diagnostikos priemones skydliaukės patologijai diagnozuoti.

3. Plataus lauko SHG mikroskopiją, tekstūros analizę ir neprižiūrimą mašininį mokymąsi apjungiantis sukurtas metodas yra skirtas kiekybiniam kolageno kapsulės struktūros PTC įvertinimui. Dviejų pakopų *k*-vidurkių klasterizacija atskleidė ryškų kapsulės ir apibrėžtų sričių heterogeniškumą su skirtingais struktūriniais požymiais, atitinkančiais nepažeistas, invazijos apimtas ir potencialiai priešinvazines vietas. Visų pirma, šiuo metodu buvo nustatytos mažai pastebimos mikroinvazijos sritys, kurios nebuvo aptiktos atliekant pradinį histopatologinį tyrimą. Siūlomo neprižiūrimo ML metodo gebėjimas aptikti tokias sritis pabrėžia jo, kaip papildomos diagnostikos priemonės, potencialą, siekiant pagerinti tikslumą, sumažinti stebėtojo kintamumą ir paremti automatizuotų klasifikavimo sistemų kūrimą.

4. Prižiūrimi ML algoritmai, taikomi iš SHG gautiems intensyvumo ir tekstūros požymiams, leidžia efektyviai diferencijuoti PTC ir FTC diagnozę. Nors klasifikavimo užduotį apsunkina požymių perteklius ir dėl gretimų audinių įtraukimo, kalcifikacijų ir tarpnavikinių kapsulių panašumo atsirandantys žymėjimo netikslumai, klasifikatoriaus našumas gerokai pagerėjo taikant specialų duomenų apdorojimą ir neprižiūrimą segmentavimą, siekiant pašalinti neinformatyvias sritis. C-SVC klasifikatorius pasiekė didžiausią patvirtinimo tikslumą, įrodantį jo atsparumą triukšmui ir šališkumui.

7.9 Autoriaus indėlis

Doktorantas atliko didžiąją dalį eksperimentinio darbo, įskaitant plataus lauko SHG ir TPEF vaizdų duomenų rinkimą, formalią analizę ir mokslinių publikacijų rašymą. Doktorantas taip pat atliko visus skaičiavimus, apskaičiavimus ir vizualizaciją, analizę ir interpretaciją. Plataus lauko SHG nustatymo modifikaciją PTC mėginių PSHG vaizdavimui ir PSHG vaizdų duomenų rinkinio gavimui atliko Dr. Danielis Rutkauskas (Fizinių ir technologijos mokslų centras, Vilnius, Lietuva) ir Dr. Radu Hristu

(Mikroskopijos-mikroanalizės ir informacijos apdorojimo centras, Bukarešto nacionalinis mokslo ir technologijų politechnikos universitetas, Bukareštas, Rumunija). Plaučių audinio pjūvių paruošimą ir imunohistocheminį dažymą atliko Dr. Nadezda Amaegberi (Baltarusijos valstybinis universitetas, Minskas, Baltarusija), Dr. Tatyana Vladimirskaja ir Olga Yatsevich (Baltarusijos medicinos magistrantūros akademija, Minskas, Baltarusija). PTC ir FTC mazgų audinių pjūvius paruošė Dr. Lucian George Eftimie (Centrinė universitetinė skubios pagalbos karo ligoninė, Bukareštas, Rumunija; Nacionalinis kūno kultūros ir sporto universitetas, Bukareštas, Rumunija).

7.10 Apie autorių

Yaraslau Padrez gimė Borisove, Baltarusijoje. 2014 m. jis baigė valstybinę mokymo įstaigą "Borisovo licėjus" ir įstojo į Baltarusijos valstybinio universiteto Fizikos fakultetą (Minskas, Baltarusija). 2020 m. įgijo branduolinės fizikos ir technologijos specialisto laipsnį (5,5 metų studijų). 2019–2020 m. gavo Pasaulio mokslininkų federacijos stipendiją. 2021 m. jis pradėjo doktorantūros studijas Valstybiniame tyrimų institute Fizinių ir technologijos mokslų centre, Molekulinių darinių fizikos skyriuje.

NOTES

NOTES

Vilniaus universiteto leidykla
Saulėtekio al. 9, III rūmai, LT-10222 Vilnius
El. p. info@leidykla.vu.lt, www.leidykla.vu.lt
bookshop.vu.lt, journals.vu.lt
Tiražas 40 egz.

NUO / SINCE 1579

Fragmentas iš Vilniaus universiteto auklėtinio Alberto Diblinskio (1601–1665) vieno geriausių XVII a. astronomijos veikalų *Centuria astronomica* (Vilnius, 1639), kuriame pateikta astronomijos pasiekimų apžvalga, remiantis stebėjimais teleskopu, atliktais kartu su kitu VU mokslininku, matematiku ir astronomu Osvaldu Krygeriu (apie 1598–1655).

VU biblioteka, BAV 47.10.21

Fragment from *Centuria astronomica* (Vilnius, 1639), one of the most well-known works on astronomy from the 17th c., written by Vilnius University graduate Albertas Diblinskis (1601–1665). It presents an overview of achievements in the field of astronomy, based on observations using a telescope together with another VU scientist, mathematician, and astronomer Osvaldas Krygeris (c. 1598–1655).

VU Library, BAV 47.10.21

bookshop.vu.lt journals.vu.lt

https://doi.org/10.15388/vu.thesis.816