IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 March 2025, accepted 22 June 2025, date of publication 26 June 2025, date of current version 7 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3583435

RESEARCH ARTICLE

Supervised Machine Learning Thyroid Carcinoma Diagnosis Using Wide-Field SHG Microscopy

YARASLAU PADREZ[1], (Member, IEEE), RADU HRISTU[2], IGOR TIMOSHCHENKO[1], LUCIAN G. EFTIMIE[3,4], DANIELIS RUTKAUSKAS[1], AND LENA GOLUBEWA[1,5], (Member, IEEE)

[1]Department of Molecular Compound Physics, Center for Physical Sciences and Technology, 10257 Vilnius, Lithuania
[2]Center for Microscopy-Microanalysis and Information Processing, National University of Science and Technology POLITEHNICA Bucharest, 060042 Bucharest, Romania
[3]Pathology Department, Central Military Emergency University Hospital "Dr. Carol Davila," 010242 Bucharest, Romania
[4]Faculty of Physical Therapy, National University of Physical Education and Sports, 600127 Bucharest, Romania
[5]Institute of Chemical Physics, Vilnius University, 01513 Vilnius, Lithuania

Corresponding author: Yaraslau Padrez (yaraslau.padrez@ftmc.lt)

This work was supported in part by the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) for the National Research Council (CNCS) project number PN-IV-P1-PCE-2023-1137, within the National Research Development and Innovation Plan 2022-2027 (PNCDI IV), the Research Council of Lithuania (LMTLT) project number No S-PD-24-158, and by RealCare, European Health and Digital Executive Agency Project number 101137466.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Carol Davila University Central Emergency Military Hospital, Bucharest, Romania, under Protocol No. 380/09.06.2020, and performed in line with the Declaration of Helsinki.

ABSTRACT Papillary (PTC) and follicular (FTC) thyroid carcinomas require different treatment strategies, but their accurate differentiation remains a challenge in conventional histopathology. Misclassification can lead to overtreatment of low-risk PTCs or inadequate treatment of FTCs, increasing the risk of recurrence and metastasis. Since the structure of the collagen capsule surrounding thyroid nodules provides diagnostically valuable information, label-free imaging with second harmonic generation (SHG) microscopy combined with machine learning (ML)-based analysis offers a promising approach for automated classification. In this study, we extracted intensity and texture features from SHG images of thyroid nodules scanned as a whole and optimized several ML classifiers, including logistic regression, support vector classification (C-SVC), multilayer perceptron, random forest, XGBoost, and LightGBM, using hyperparameters tuning with stratified 10-fold cross-validation. One of the major challenges in classification was label noise resulting from 1) mislabeling of adjacent tissue, 2) PTC calcifications mimicking FTC features, and 3) capsule heterogeneity. To address this issue, we applied unsupervised segmentation to exclude mislabeled regions and consider capsular heterogeneity as a diagnostic feature. Recursive feature elimination and mutual information selection further refined the feature set and improved classification accuracy. Among all models, C-SVC achieved the highest accuracy (84.73%) with robust generalization to unknown data, significantly outperforming standard ML approaches (60–70%). These results demonstrate the feasibility of SHG microscopy-based ML classification as a reliable adjunct to existing histopathologic methods, which could improve diagnostic accuracy and patient outcomes.

INDEX TERMS Association rule learning, automated machine learning, boosting, biomedical image processing, cancer, image analysis, wide-field second harmonic generation microscopy.

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

I. INTRODUCTION

Papillary and follicular thyroid carcinomas (PTC and FTC) are both well-differentiated carcinomas and account for

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

approximately 88% of all thyroid tumors [1]. The prognosis for FTC is often worse than for PTC [2], [3] and requires a total thyroidectomy [4]. However, low-risk PTCs are often over-diagnosed and thus overtreated with total thyroidectomy and/or radioactive iodine [5], whereas a less aggressive approach is feasible and may prevent postoperative complications and hypoparathyroidism [6]. To prevent overtreatment, a precise diagnosing of FTC and PTC is needed.

The collagen capsule surrounding a thyroid nodule is an important histopathological feature, as it provides insight into tumor invasiveness, malignancy potential, and prognosis. In PTC, the capsule may be incomplete, infiltrative, or absent, whereas in FTC, the capsule is typically more well-defined. Differences in the capsular structure may correlate with the biological behavior of these tumors, including their metastatic potential and response to treatment.

A comparison of the capsules surrounding PTC and FTC nodules which are often evaluated through histopathological examination could yield valuable insights into their pathogenesis, clinical outcomes, prognostic indicators and ultimately may improve diagnostic accuracy.

Second harmonic generation (SHG) microscopy enables label-free visualization of fibrillar collagens which comprise the capsule around the nodule and can be used to supplement traditional histopathology. Intensity and texture analysis of SHG microscopy images of tissue sections based on the first, second and high order statistics (F-, S-, HOS) [7], [8] enables extraction of a set of numerical features which quantitatively describe collagen network properties and allow interpretation of specific changes in tumor microenvironment accompanying tumor progression.

Texture analysis of collagen structures imaged by SHG microscopy has a strong diagnostic potential for classification of benign and malignant thyroid nodules. Statistically significant differences in SHG texture features have been shown among various thyroid benign and malignant pathologies [9], [10], [11], [12], [13], [14], [15]. However, comparison of randomly selected small ROIs (approx. 100 μm - 200 μm for the above studies) may be flawed and not provide a correct and complete picture of the characteristics of collagen structures in these neoplasms. The heterogeneity in collagen structures along the whole capsules of PTCs [16] and around follicular adenomas (FA) and FTC [15] was demonstrated.

Wide-field variant of SHG microscopy [16], [17], [18], [19], [20], [21], [22], [23] allows the rapid accumulation of separate images of approximately 100 μm $\times$ 100 μm. These images can be arranged into a sub-centimeter-sized mosaic of the tissue sample [16], providing an alternative to whole slide scanning often used for imaging tissue sections stained with hematoxylin and eosin (H&E) [24]. This approach also overcomes the ROI size limitation, which is critical for analyzing thyroid nodule capsules. However, the manual analysis of large datasets of SHG images is limited and requires application of machine learning (ML) algorithms.

Unsupervised ML techniques, such as principal component analysis (PCA) and k-means clustering, applied to texture features of SHG images of whole PTC nodules, enabled identification of patterns associated with altered helical pitch angles of collagen fibers, particularly in areas of PTC capsular invasion [16]. These methods also highlighted the structural heterogeneity of the collagen capsule around PTC, identifying characteristic patterns in collagen textures and networks related to invasions, tight collagen barriers and altered areas [16].

Supervised ML classifiers (e.g., logistic regression (LR), linear support vector machines (SVM), random forest (RF), gradient boosting, and multilayer perceptron (MLP)) are often applied to distinguish between different types of cancer [25], [26], [27], [28], with classification primarily based on MRI or CT texture image analysis [29]. Although studies have reported efforts to classify thyroid nodules using features extracted from SHG and multiphoton fluorescence images [30], research on ML classification of thyroid neoplasms based on SHG imaging remains limited.

Data quality plays a crucial role in classifier performance. The accuracy of measurements (apparatus noise), the presence of irrelevant or redundant features (feature noise), and mislabeling (label noise) can all hinder model training and performance. Improving data quality often leads to better classifier performance, lower computational costs, and enhanced model generalization [31], [32]. While apparatus-related noise is difficult to eliminate, data quality can be improved by selecting only highly discriminative features and addressing label noise.

In real-world situations, feature diversity is required since there is often no prior knowledge about which features are relevant to the target. However, increasing feature diversity can introduce irrelevant features, which are not associated with the target but still affect the ML classifier, as well as redundant features that do not provide new information about the target [31]. High dimensionality of the feature space can lead to overfitting and hampered model performance [33]. Removing feature noise from data is an important step in the design of classification models, often achieved by reducing the dimensionality of the feature space. Dimensionality reduction is often done by feature extraction (e.g. PCA; linear discriminant analysis (LDA); etc.) which creates new features from the original ones, or through feature selection which reduces the number of initial features by removing less important ones. The latter approach is generally preferred, as it keeps the original features interpretable, while the new features lack biophysical meaning [31].

Label noise appears if a subset in the sample is labelled differently than the sample. The opposite situation may also occur: the sample is labelled as a whole, but it contains subsets that do not correspond to the assigned label, e.g., an imaged sample labelled as a tumor, may contain areas of normal tissue. Label noise may originate from automated

labelling, but even manual labelling by experts can lead to error rates of tens of percent in data labels [34].

Given these challenges, diagnosing PTC and FTC from SHG images is a complex task, since data is likely prone to both feature and data noise. However, existing research [9], [10], [11], [12], [13], [14], [15], [16] highlights the high diagnostic potential of SHG image analysis for distinguishing thyroid cancer types. Well-designed and properly trained classification models could supplement the diagnosis and contribute to the disclosure of mechanisms of FTC and PTC progression.

In this study, we used three monolithic (LR; C-support vector classification, C-SVC and MLP) and three ensemble (RF; XGBoost and LightGBM) classifiers to: (i) diagnose PTC and FTC based on texture features extracted from SHG images of whole thyroid tissue section scans; (ii) evaluate the impact of label noise and feature noise (irrelevance/redundance) on the predictive performance of these models; (iii) assess the feasibility of SHG-based discrimination between PTC and FTC in real-life applications.

We show that the label noise originating from the similarity of collagen structures in PTC and FTC samples and the heterogeneity of the capsules significantly affect the performance of ML classifiers. Feature selection improves the accuracy of all classifiers if label noise is removed. C-SVC and MLP classifiers outperform other classifiers, with C-SVC showing better performance even when feature and label noise are not excluded. The reduction of label and feature noise increases the accuracy of the C-SVC classifier up to almost 85% on the validation set. The developed classification model shows good generalization and provides satisfactory classification of the unknown samples, making it a reliable complementary technique for conventional methods of differential diagnosis of thyroid carcinoma.

II. EXPERIMENTAL SECTION

A. THYROID TISSUE SAMPLE PREPARATION

The tissue sections of the PTC and FTC nodules were prepared according to the standard histological procedure [35]. The formalin-fixed, paraffin-embedded 4 μm – 7 μm thick tissue sections were placed on glass slides and stained with hematoxylin and eosin (H&E). Bright field imaging of the slides was performed using the Aperio LV1 IVD Whole Slide Scanner (Leica Biosystems) equipped with a 20× objective. The classification of the samples as PTC or FTC was performed by an expert histopathologist (LGE).

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

B. WIDE-FIELD SHG IMAGING SETUP

SHG images were obtained with a custom-built wide-field non-linear microscopy setup based on a modular microscope from ASI (Applied Scientific Instruments) [18]. Measurement conditions and SHG microscopy setup are described in detail in [16]. For the analysis, 45360 separate SHG images of sizes 150 μm × 150 μm were collected. In total, 10 samples of PTC (5) and FTC (5) nodule sections were imaged for this work. The scanned areas of PTC and FTC tissue sections ranged in size from 2.1 mm × 2.1 mm to 8.4 mm × 8.4 mm. Bright field and SHG images are shown in Fig. S1 and Fig. S2, Supplementary Material 1.

The SHG images of the PTC capsules selected for this study were previously analyzed by unsupervised ML algorithms [16], [17], which facilitated exploratory data analysis prior to optimization of classifiers and interpretation of the results of ML model performance.

C. SHG IMAGES PRE-PROCESSING

The edges of the 150 μm × 150 μm SHG images were cropped to 117 μm × 117 μm (800 pixels × 800 pixels) to minimize illumination inhomogeneity caused by the Gaussian laser beam intensity profile. Further preprocessing included applying a threshold of 110 counts [36] and a two-stage 99th percentile reduction applied to the grey levels of the entire 16-bit SHG image dataset, as described in [16].

D. INTENSITY AND TEXTURE FEATURE EXTRACTION

For a complete quantitative evaluation of the textural changes in collagen structures, statistical moments of different orders were applied:

1) First-order statistics (FOS): based on image intensity, includes the following parameters: mean (μ_1), standard deviation (σ), skewness (g_1), and kurtosis (g_2).

2) Second-order statistics (SOS): derived from the grey level co-occurrence matrix (GLCM) [37], calculated for five-pixel distances (1, 3, 6, 9, and 12 px). Parameters include energy (E), inertia (I), correlation (C), local homogeneity (L) and entropy (H) [7].

3) High-order statistics (HOS): computed from the grey level run-length matrix (GLRLM) [8], including short run emphasis (*SRE*), long run emphasis (*LRE*), grey-level non-uniformity (*GLN*), run length non-uniformity (*RLN*), and run percentage (*RP*) [7].

A total of 34 texture and intensity features (4 FOS, 25 SOS, and 5 HOS) were extracted from each single SHG image, forming a feature vector that characterizes each image. A complete description of these parameters can be found in [16].

E. MACHINE LEARNING ANALYSIS

The schematic workflow outlining dataset preparation, strategies of handling feature noise and label noise, model optimization, training, testing and generalization are shown in Fig. 1.

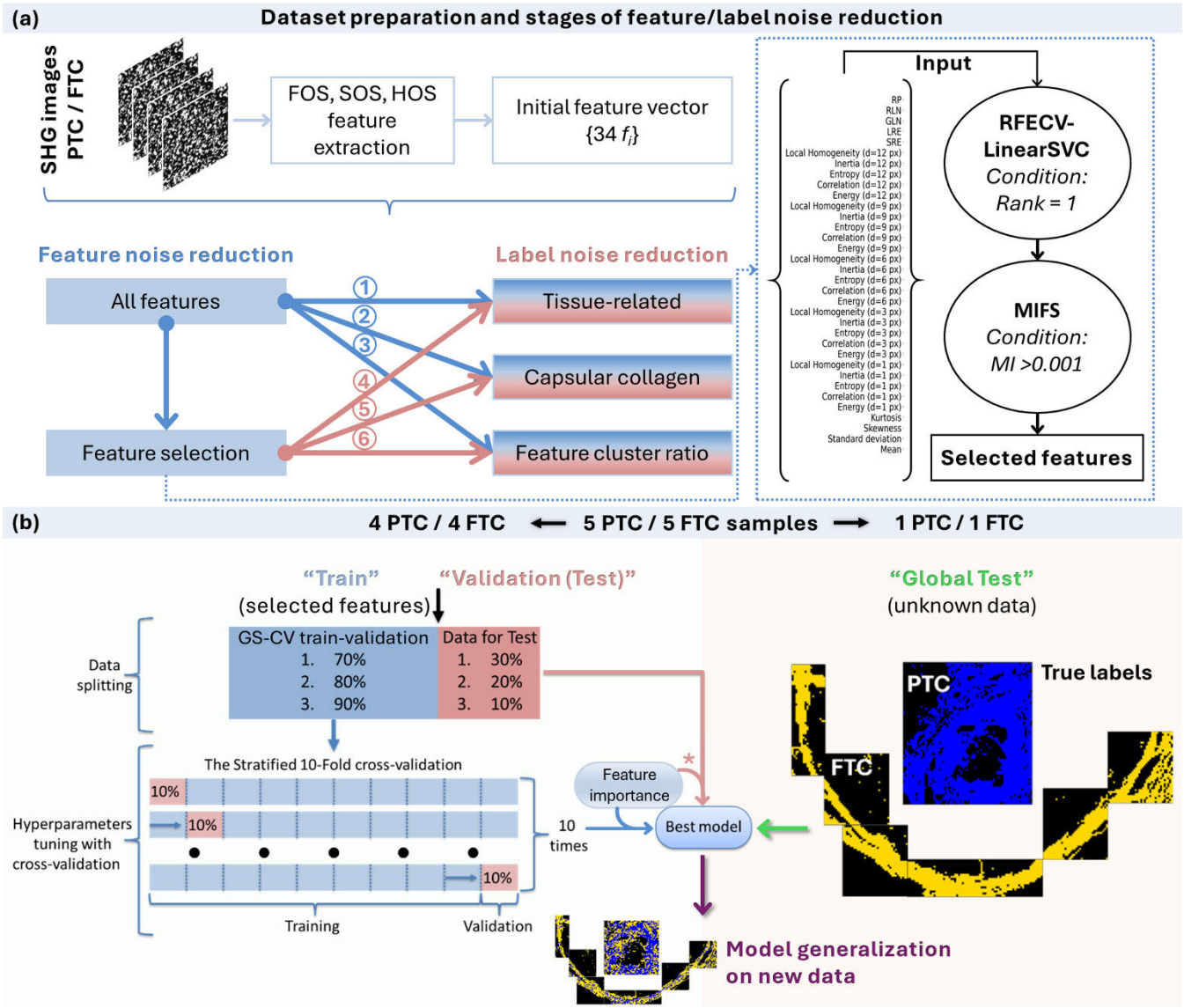


FIGURE 1. The schematic workflow of (a) dataset preparation, feature extraction, feature selection, and label noise reduction, and (b) optimization and generalization of the ML classification models. An asterisk indicates that, for permutation importance analysis (feature importance analysis for C-SVC 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.

1) MANAGING LABEL NOISE

Since the most obviously mislabeled data relate to the SHG images of glass slide, glass-related SHG images were excluded as previously described in [16] 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) [16] 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 the 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. Since heterogeneity of collagen capsules of PTC was previously demonstrated [16], an assumption that FTC capsule may also be heterogeneous 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 III-C.

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 on the training set.

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 (Fig. 1a). 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. [38]. This method ranks features in descending order by recursively considering feature subsets of decreasing sizes [38]. 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 [39].

Next, Mutual Information Feature Selector (MIFS) [40] 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 [26]. 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 [41]. A selection threshold of MI>0.001 (in nat units) was applied in this study.

Feature selection was applied independently on the training sets for each label noise reduction approach (subsection II-E1) prior to optimizing the classifiers by stratified k-fold cross-validation (subsection II-E5). 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 Fig. 1a).

3) PREPARATION OF THE TRAINING, VALIDATION AND TEST DATASETS

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 [36].

Since different strategies for managing label noise (see subsection II-E1) 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 S1, Supplementary Material 1.

4) MACHINE LEARNING CLASSIFIERS

The following ML classifiers were used in the study: ensemble classifiers Random Forest (RF) [42], Extreme Gradient Boosting (XGBoost) [43], Light Gradient-Boosting Machine (LightGBM) [44], and monolithic classifiers Logistic Regression (LR), Ñ-Support Vector Classifier (C-SVC) [45] and Multilayer Perceptron (MLP) [46], which relates to Deep Learning algorithms. All models have been developed in the Python platform libraries scikit-learn 1.6.1 [47]. 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.

5) OPTIMIZATION OF ML CLASSIFIERS WITH HYPERPARAMETER TUNING AND CROSS-VALIDATION

The hyperparameters are top-level parameters of the ML classifier that control model development process and must be optimized before training the best/final model [48]. The hyperparameters of each ML model, tuned and used in this study, are detailed in Table S2, Supplementary Material 1. Hyperparameters not listed in Table S2 were set to their default values.

Hyperparameter tuning was performed using either grid search or halving grid search. The 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 (SH) algorithm [49], was used for models with larger hyperparameter combinations (RF, XGBoost and MLP) [50].

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 [51] 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 [52]. Stratified

TABLE 1. Features selected by RFECV-LinearSVC and MIFS within each label correction approach.

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.

10-fold cross-validation can handle multi-class problems, so it was also applied in the ''PTC/FTC/Normal tissue'' classification [47], ensuring carcinoma-specific cluster ratios were preserved in both training and validation sets.

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 Fig. 1b.

The hyperparameter values for all optimized classifiers are summarized in Table S3, Supplementary Material 1.

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 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 [53].

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 [54]. The performance of the classifier was represented by the area under the ROC curve (AUC) values [53].

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) [47].

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, as demonstrated for PTC samples in [16], there is no single universal approach for estimating the most important features. For instance, multicollinearity may affect the results of permutation feature importance analysis (PIA) if applied to LR or tree-based classifiers. Thus, feature contribution to decision-making was treated independently using model-specific approaches, as detailed in Table S4, Supplementary Material 1.

8) 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.

III. RESULTS AND DISCUSSION

All FTC and PTC samples were imaged using the SHG microscopy setup and the combined images of all samples are shown in Fig. S1, S2 (Supplementary Material 1). To ensure diverse sample description, 34 intensity and texture features (4 FOS, 25 SOS and 5 HOS) extracted from each 117 μm $\times$ 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 1.

A. TISSUE-RELATED SHG IMAGES

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 principal components (PCs) covering more than 92% of data variance, enabled segmentation of the tissue section image into tissue- and non-tissue related SHG images [16]. A typical SHG image segmentation is shown in Fig. 2. 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 (Fig. 2).

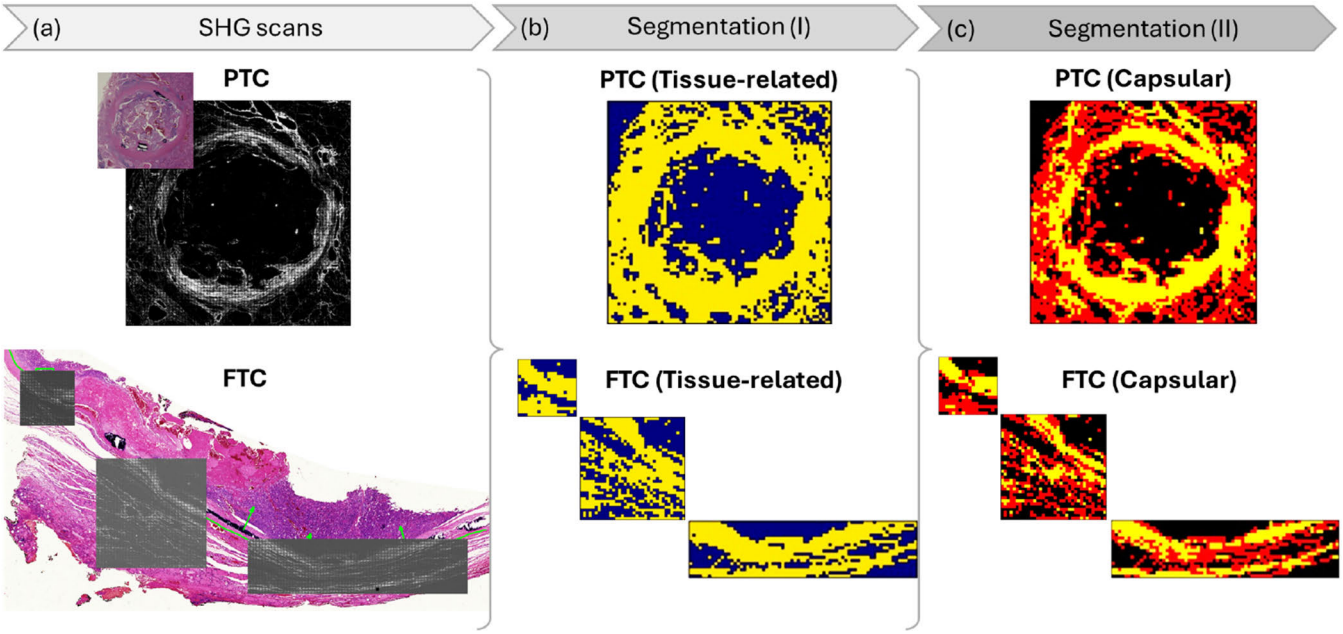


FIGURE 2. Typical SHG and brightfield images of PTC and FTC nodule sections (a) and the result of the separation of tissue-related (b) or capsule-related (c) SHG images based on k-means clustering performed on the principal components PC1-PC5. The images separated for the analysis are colored yellow in (b) and (c). Bright field, SHG image (a) and tissue-related image segmentation (b) of a PTC nodule are reproduced with permission 2024 Y. Padrez et al. Published by Elsevier Ltd. Licensed under CC BY 4.0 [16].

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 (Supplementary Material 1) and Fig. S1a-1 – Fig. S1a-18 (Supplementary Material 1). The highest accuracy values were achieved with a 90/10 (training/validation) data split and classifier optimization using the complete feature vectors (Table 2 "I. Tissue-related", Fig. 3).

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 (Fig. 3a-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.

On the unknown test set, MLP demonstrates correct classification rates slightly above 50% (Fig. 3e), while C-SVC exhibited low discriminative power for PTC but performed well in distinguishing FTC (Fig. 3d). Visual inspection of the classified PTC images revealed that the normal tissue surrounding the circular PTC capsule was frequently misclassified as FTC (yellow), 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 (Fig. 4).

Calcifications are more frequently observed in PTCs than in FTCs and are generally accepted as a reliable indicator of PTC [55]. 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, hemorrhage and subsequent fibrosis within the tumor. Collagen I serves as a scaffold for mineralization – deposition of mineral salts [56], such as calcium carbonate phosphates [57], calcium hydroxyapatite [55], or other calcium compounds within the fibrous extracellular matrix. Mineral deposits do not generate SHG [58], [59], indicating that SHG signal detected in PTC samples originate from fibrillar collagens which favored the formation of macrocalcifications.

The areas of the PTC capsule which are associated with calcifications (Fig. 4) were classified as FTC (Fig. 3) 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 mislabeling: 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.

TABLE 2. 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.

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.

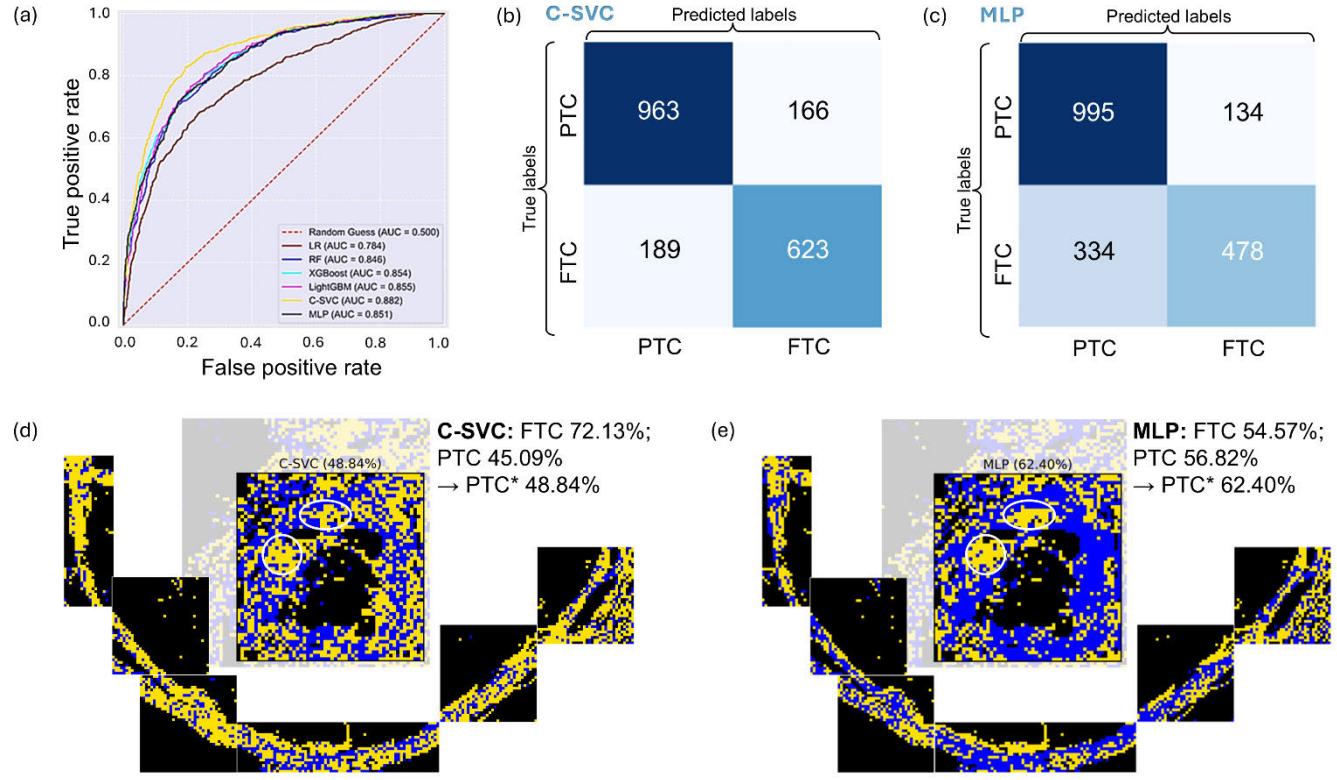


FIGURE 3. Performance of the best developed ML models optimized based on accuracy: (a) ROC curves of all ML models, (b) confusion matrix for C-SVC; (c) confusion matrix for MLP; (d) C-SCV classification performed on new data set (test set); (e) MLP classification performed on new data set (test set). Blue colored tile images mark image 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.

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 (Fig. 5). 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 [42]. 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.

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 1). However, the performance of the classifiers decreases in all

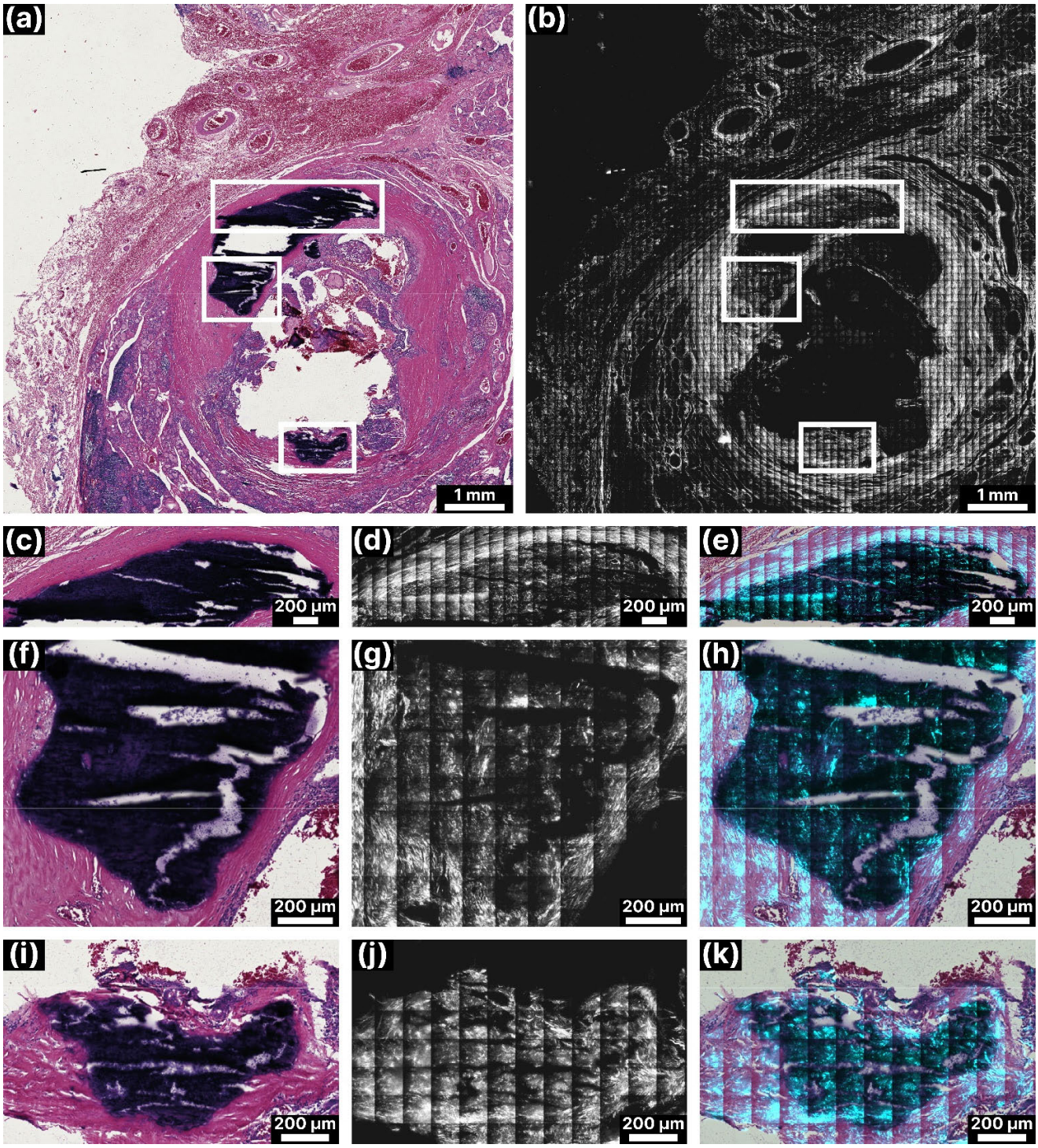


FIGURE 4. Bright-field (a) and SHG microscopy (b) of a whole PTC nodule with calcifications: (c)-(e), (f)-(h) and (i)-(k) – bright-field, SHG and merged images of calcifications marked with white rectangles in (a) and (b). Bright field and SHG image of a PTC nodule are reproduced with permission 2024 Y. Padrez et al. Published by Elsevier Ltd. Licensed under CC BY 4.0 [16].

cases, indicating that there is another, more significant source of noise beyond feature noise, e.g. label noise (Table S5 in Supplementary material 1, and Fig. S1b-1 – S1b-18 in Supplementary material 2b). The proportion of mislabeled data is likely considerable, and the classification models mainly fail in handling this mislabeled 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 classification accuracy. The perinodular tissue may contain signatures of tumor

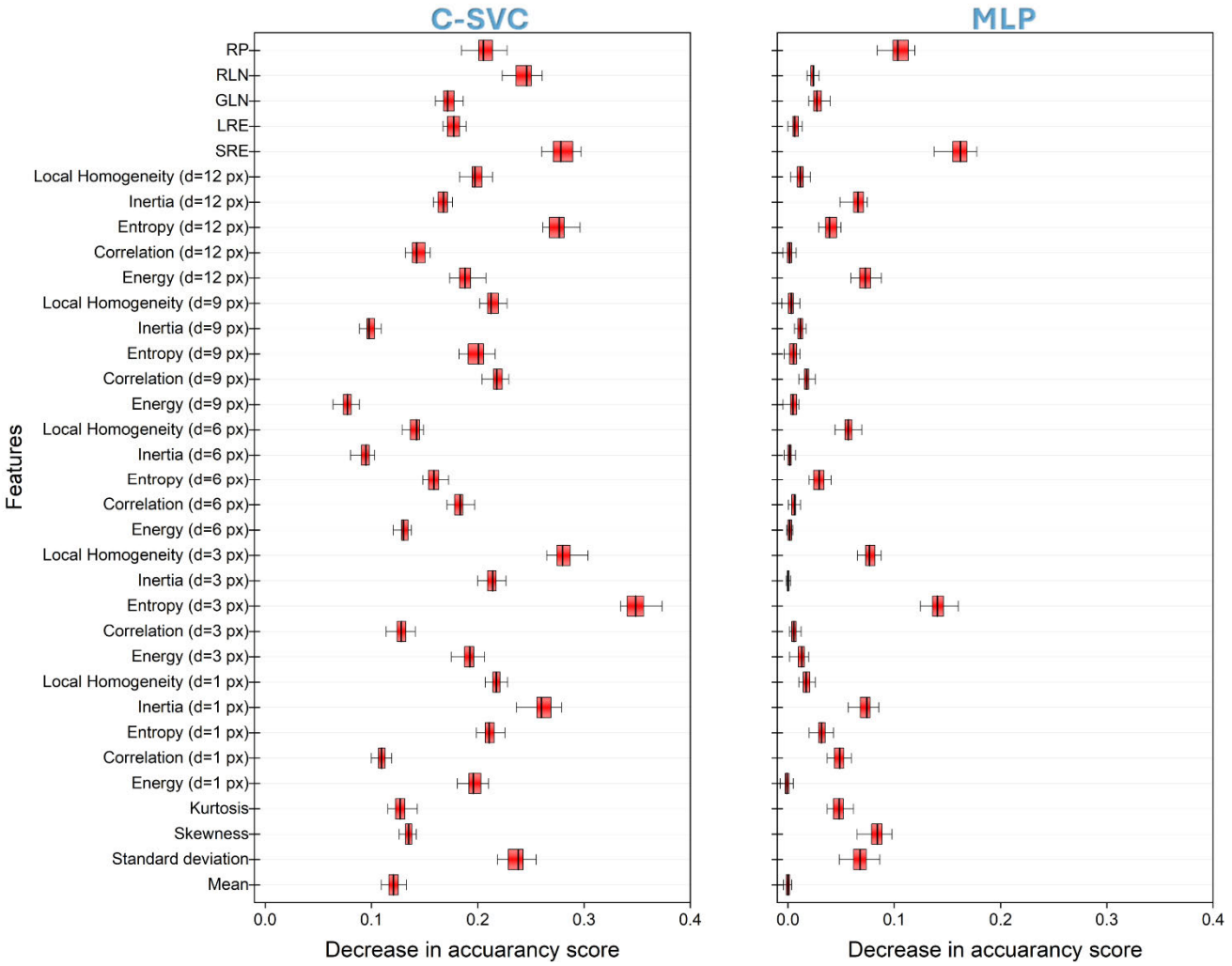


FIGURE 5. Permutation importance analysis of C-SVC and MLP models.

progression, such as collagen network remodeling induced by increased MMP9 secretion from macrophages recruited to the tumor growth sites [60], providing additional information for classification models. However, since these changes may be similar in both carcinoma types [61], labelling surrounding tissues as PTC or FTC could lead to mislabeling and misclassification.

B. CAPSULE-RELATED SHG IMAGES

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 [16]: PCA was performed to the SHG dataset, followed by binary k-means clustering on the acquired PCs. A typical result is shown in Fig. 2c. 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 carried out 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 S1, Supplementary Material 1). 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 [62]. 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 summarized in Table S6 (Supplementary Material 1) and Fig. S2a-1 – S2a-18 (Supplementary Material 2a). 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 2 ''II. Capsule-related''. Overall, models perform better than those trained on all tissue-related data (Fig. 6a). However, ensemble models

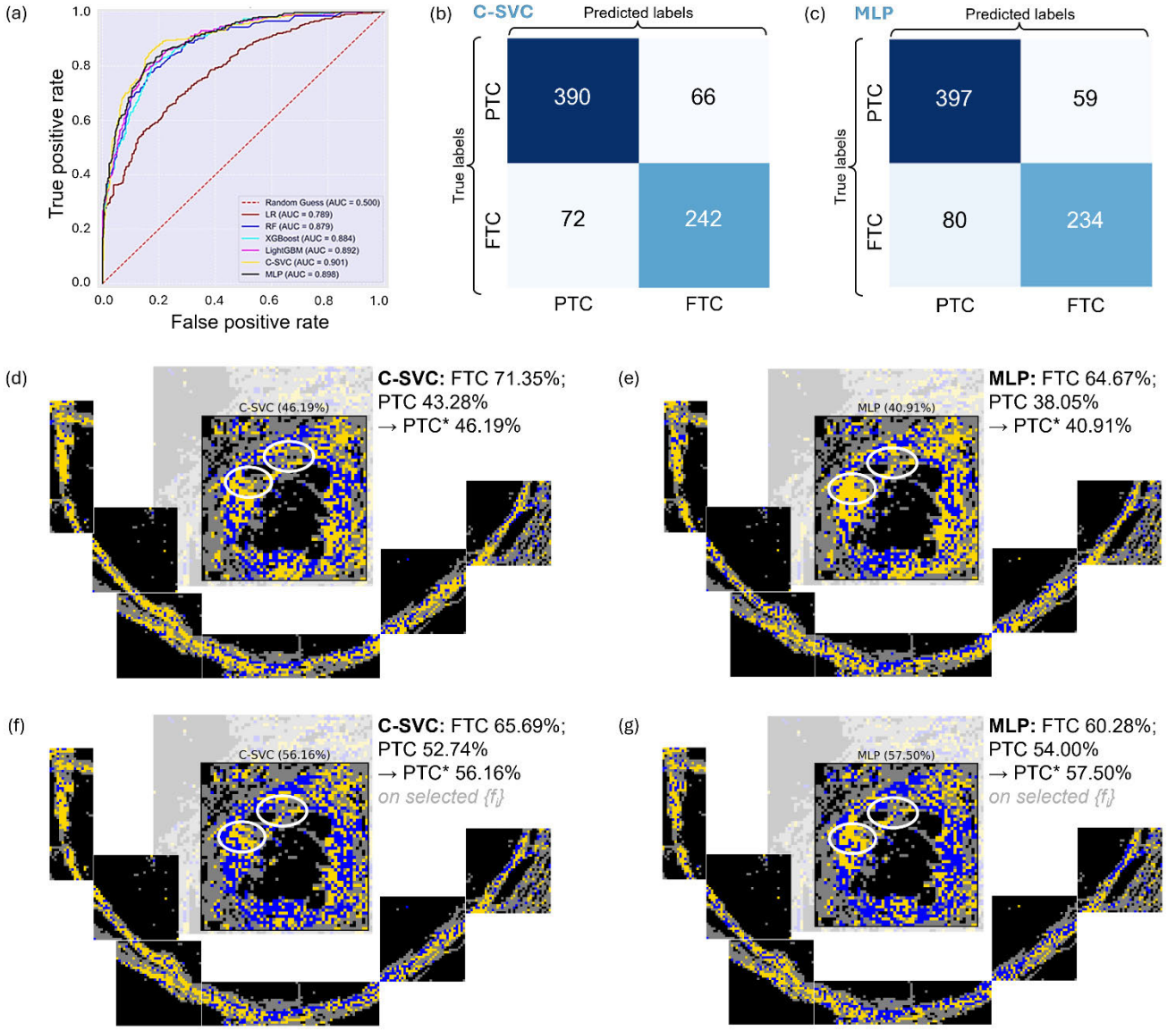


FIGURE 6. Performance of the ML models developed on capsule-related datasets for a split of (90/10): (a) ROC curves of all ML models, (b) confusion matrix for C-SVC; (c) confusion matrix for MLP; (d) C-SVC classification performed on the new data set (test set); (e) MLP classification performed on the 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.

(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 (Fig. S2a-13, Supplementary material 2a) 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 2 "II. Capsule-related", Fig. 6). In addition, the recall and F-1 score metrics increase for the MLP model, indicating enhanced performance when trained on capsule-related datasets.

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 (Fig. 6d, e). The C-SVC model's classification results for PTC were similar to those from the all-tissue-related datasets (Fig. 6d), 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 (Fig. 6d).

PIA reveals that C-SVC is better at identifying feature relevance and addressing feature multicollinearity (Fig. 7).

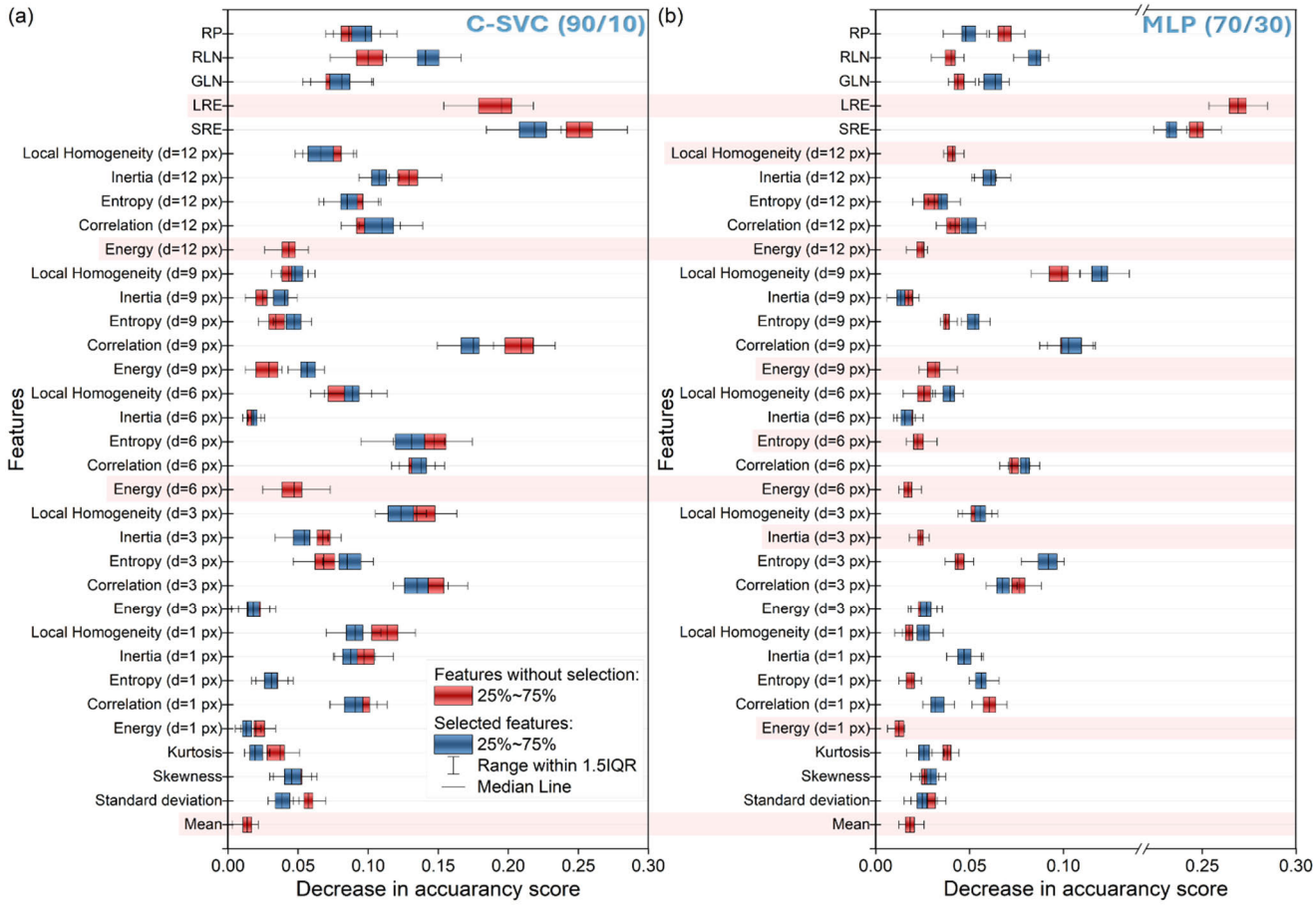


FIGURE 7. Permutation importance analysis of the performance of (a) C-SVC (90/10) and (b) MLP (70/30) classifiers, optimized based on the complete (red) and reduced (blue) feature sets.

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 (Fig. 7).

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 2 "II. Capsule-related", Table S6 (Supplementary Material 1), 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 (Fig. 6f). 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 (Fig. 6g). The classification results for all classifiers are summarized in Fig. S2b-1 – S2b-18 in Supplementary Material 2b.

PIA of C-SVC and MLP, trained on the reduced feature sets (Fig. 7), 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 explain 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 remaining 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

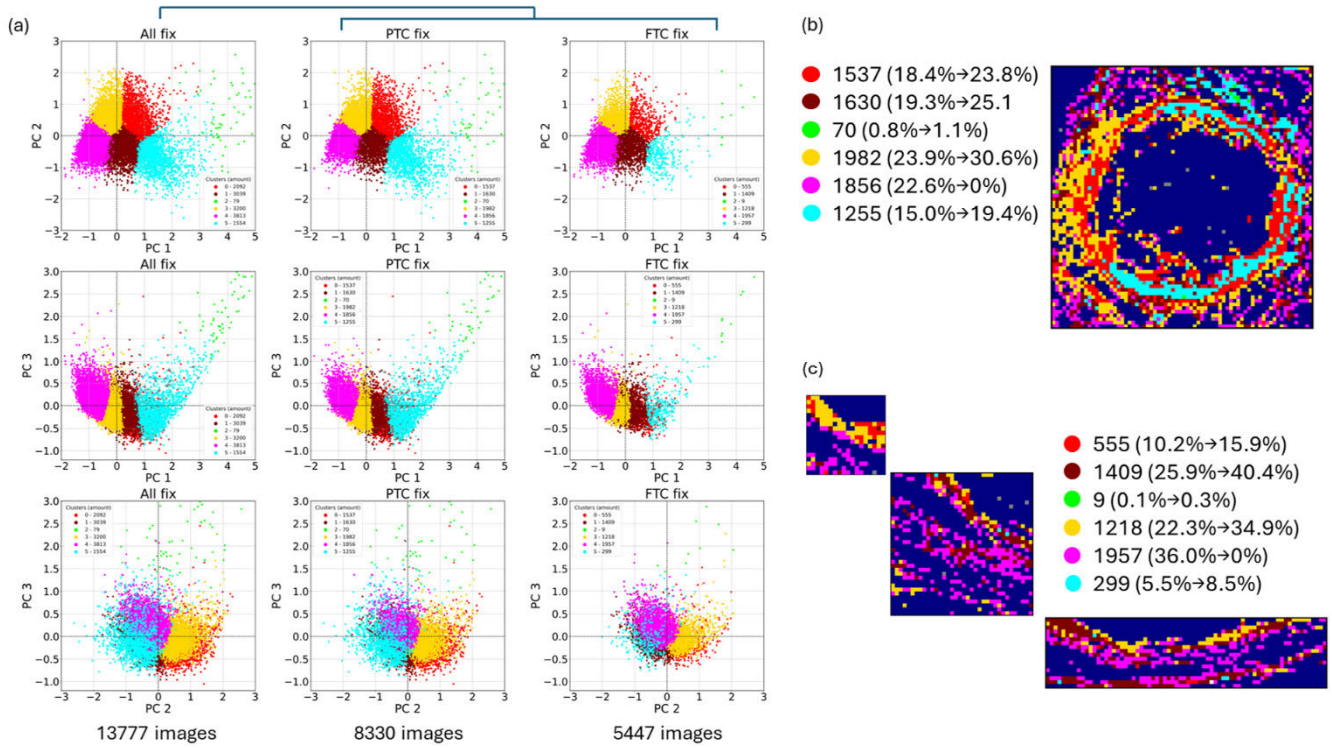


FIGURE 8. PCA analysis of the feature datasets of PTC and FTC samples and their clustering using k-means: (a) score plots of PC1 vs. PC2, PC3 vs. PC1 and PC3 vs. PC2 for all data and separately for PTC and FTC; (b) cluster map of random PTC and FTC samples. The numbers indicate the percentage of each 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.

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 (Fig. 6d-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 tumor capsules are heterogenous, with variations in collagen composition and structure influenced by tumor 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.

C. 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 mislabeling in the tissue-related approach.

Segmentation based on intensity and texture features was performed as described in [16] to prove the similarity of perinodular tissue. PCA and multi-class k-means clustering results (Fig. 8) show that while capsular collagen is heterogeneous in both PTC and FTC, adjacent tissue is separated in one class (colored magenta, Fig. 8b-sc).

Both carcinoma capsules consist of the same clusters, and this complicates classification of PTC and FTC capsules even when adjacent tissue is excluded from the analysis (e.g., Fig. 6d-g). Despite shared cluster composition, the cluster ratios differ between carcinoma types. PCA score plots representing all data (Fig. 8a) and examples of segmented SHG scans (Fig. 8b-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

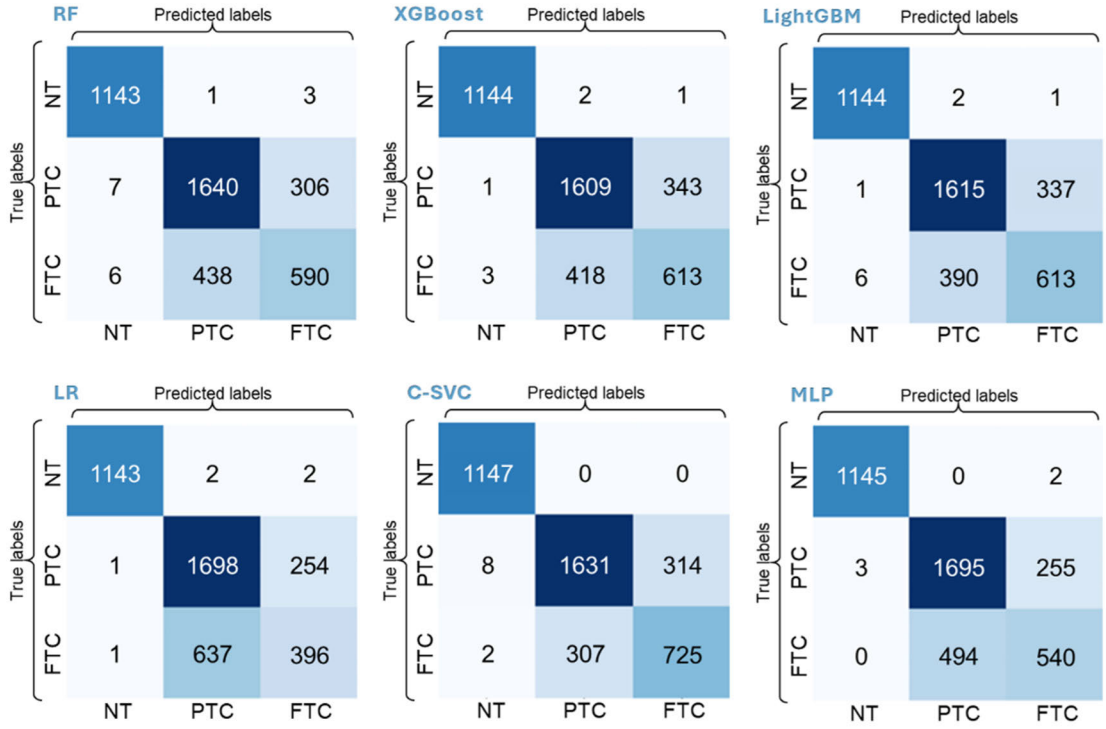


FIGURE 9. Confusion matrices calculated for all classifiers developed with a 70/30 split (optimized by maximum accuracy) for datasets considering the ratio of clusters in PTC and FTC samples.

differences in growth rates of the nodules. PTC tends to grow more slowly, while FTC exhibits increased aggressiveness and a higher tendency for metastasis [63]. FTC often presents with larger nodules at diagnosis [64], contributing to its faster growth. Furthermore, FTC has a tendency for hematogenous spread, contrasting with the lymphatic spread more commonly associated with PTC [63], 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 (Table S1, Supplementary Material 1) could affect the classifier performance.

Multi-class classification, which includes all SHG images and tends to correct the mislabeling 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 (Supplementary material 1) and Fig. S3a-1 – S3a-18 (Supplementary material 2a). The 70/30 split results for MLP and C-SVC shown in Table 3 and Fig. 9-10 are more representative, since this split includes more data in the validation set and preserves cluster ratios.

Confusion matrices (Fig. 9) 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 (Supplementary material 1)). 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 (Fig. 9), while C-SVC outperforms other classifiers achieving 84.73% accuracy on the validation set (Fig. 9, Table 3).

The C-SVC classifier performs better on the unknown test set compared to the all tissue and capsule-related approaches without feature selection (Fig. 10a), but worse than the capsule-related approach with feature selection (Fig. 6f). 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, Supplementary Material 1; Fig. S3b-1 – S3b-18, Supplementary Material 2b). However, C-SVC generalization performance was significantly improved and correct predictions increased up to 68.65 % and 55.26 % for FTC and PTC, respectively (Fig. 10b).

TABLE 3. 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.

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	+++

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.

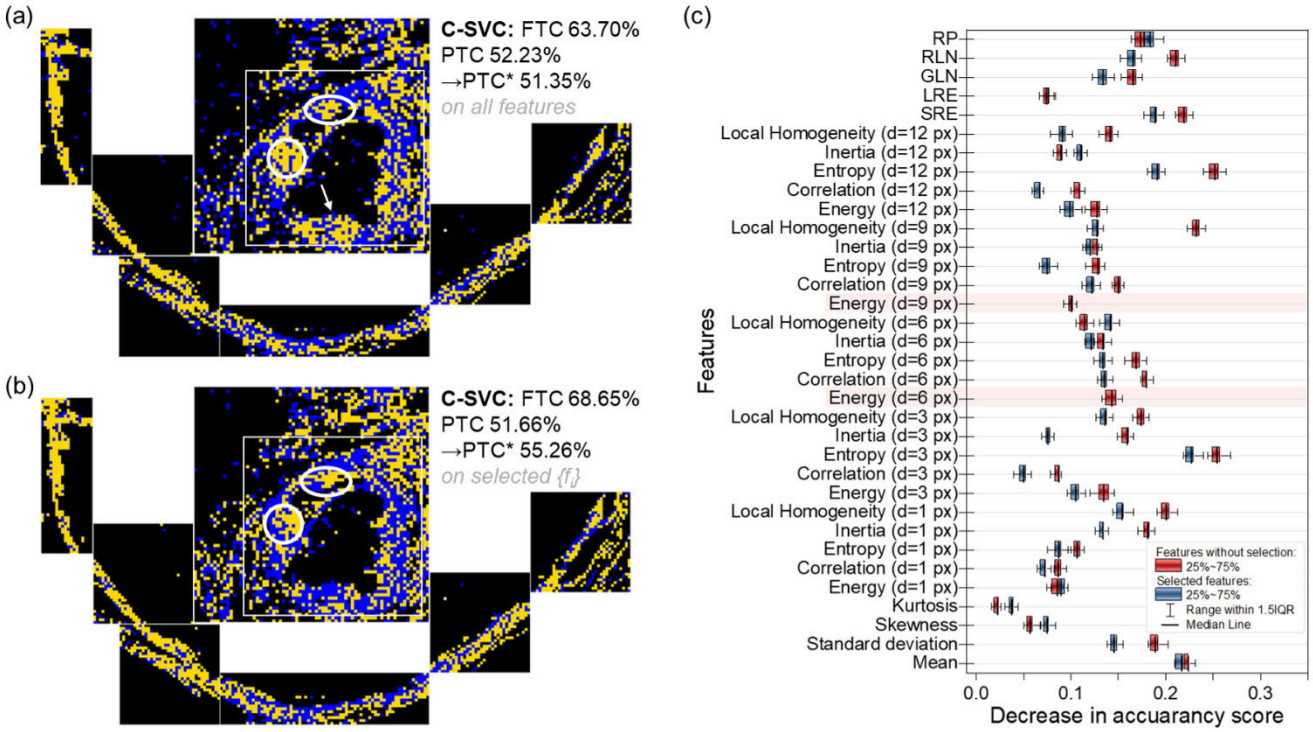


FIGURE 10. Performance of C-SVC classifier optimized for datasets considering the ratio of clusters in PTC and FTC samples: (a) classification of a global test set based on the complete set of features; (b) classification of a global test set based on the reduced set of features; (c) permutation importance analysis. Blue colored 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.

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 (Fig. 10c) relies on the full feature set, with high μ_1 and σ (FOS) contributions distinguishing non-target class from PTC and FTC capsules. Low PI scores for μ_1 and σ in the capsule-related approach and higher PI scores in other approaches support this conclusion. PI analyses 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.

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). PI analysis 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.

IV. CONCLUSION

Automated diagnosis of papillary and follicular thyroid carcinomas is challenging, but supervised machine learning classification can improve the diagnosis. Wide-field SHG imaging allows intensity and texture feature extraction for supervised classification to differentiate between PTC and FTC, though feature redundancy and mislabeling affect datasets. Mislabeling sources include: (i) adjacent tissue irrelevant to PTC or FTC, (ii) calcifications in PTC, and (iii) heterogeneity and similarity between PTC and FTC capsules. Two unsupervised ML segmentation approaches were proposed to automatically remove adjacent tissue from the analysis. Additionally, feature selection and consideration of the different heterogeneity in PTC and FTC capsules improved C-SVC accuracy on the validation data to 84.73%. The classification of the unknown data was also satisfactory, despite the limited training data. However, other classifiers (LR, RF, XGBoost, LightGBM, MLP) were more affected by label and feature noise. The strong performance of the C-SVC suggests that ML-based classification is a valuable tool for differential diagnosis of PTC and FTC. However, despite the apparent success of label noise reduction methods, errors related to the peculiarities of collagen networks in calcifications remain and reduce the accuracy of the models. Further improvements can be achieved by incorporating additional features from CARS, fluorescence and bright-field images, or clinical data.

V. DISCLOSURES

The authors declare no conflict of interest.

VI. SUPPLEMENTARY INFORMATION

The Supplementary Materials contain the following files:

- **Supplementary Material 1** – Images of PTC and FTC samples; Distribution of SHG images between training and validation datasets and figures; Description of the hyperparameters and algorithms of grid search of the classifiers; Hyperparameter configurations of the optimized classifiers after tuning, Algorithms description for feature importance analysis; Performance of the optimised classifiers for tissue-related SHG images (label correction approaches I-III).
- **Supplementary Material 2a** – Results of classification performance of ML classifiers optimized on the complete set of features.
- **Supplementary Material 2b** – Results of classification performance of ML classifiers optimized on the reduced sets of features.

REFERENCES

[1] L. Boucai, M. Zafereo, and M. E. Cabanillas, "Thyroid cancer: A review," *JAMA*, vol. 331, no. 5, p. 425, Feb. 2024, doi: 10.1001/jama.2023.26348.

[2] Z. Liu, W. Zeng, L. Huang, Z. Wang, M. Wang, L. Zhou, D. Chen, H. Feng, W. Zhou, and L. Guo, "Prognosis of FTC compared to PTC and FVPTC: Findings based on SEER database using propensity score matching analysis," *Am. J. Cancer Res.*, vol. 8, no. 8, pp. 1440–1448, Jan. 2018.

[3] E. M. Aboelnaga and R. A. Ahmed, "Difference between papillary and follicular thyroid carcinoma outcomes: An experience from Egyptian institution," *Cancer Biol. Med.*, vol. 12, no. 1, pp. 53–59, Mar. 2015, doi: 10.7497/j.issn.2095-3941.2015.0005.

[4] O. Gimm and H. Dralle, "Differentiated thyroid carcinoma," in *Surgical Treatment: Evidence-Based and Problem-Oriented*. Munich, Germany: W. Zuckschwerdt Verlag GmbH, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK6979/

[5] P. H. Dedhia, M. C. Saucke, K. L. Long, G. M. Doherty, and S. C. Pitt, "Physician perspectives of overdiagnosis and overtreatment of low-risk papillary thyroid cancer in the U.S.," *JAMA Netw. Open*, vol. 5, no. 4, Apr. 2022, Art. no. e228722, doi: 10.1001/jamanetworkopen.2022.8722.

[6] P. A. Singer, "Treatment guidelines for patients with thyroid nodules and well-differentiated thyroid cancer," *Arch. Internal Med.*, vol. 156, no. 19, p. 2165, Oct. 1996, doi: 10.1001/archinte.1996.00440180017002.

[7] R. W. Conners and C. A. Harlow, "A theoretical comparison of texture algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 3, pp. 204–222, May 1980, doi: 10.1109/TPAMI.1980.4767008.

[8] M. M. Galloway, "Texture analysis using gray level run lengths," *Comput. Graph. Image Process.*, vol. 4, no. 2, pp. 172–179, Jun. 1975, doi: 10.1016/s0146-664x(75)80008-6.

[9] R. Hristu, B. Paun, L. Eftimie, S. G. Stanciu, D. E. Tranca, and G. A. Stanciu, "Changes in the collagen structure of thyroid nodule capsules determined by polarization-resolved second harmonic generation microscopy," in *Proc. 20th Int. Conf. Transparent Opt. Netw. (ICTON)*, Jul. 2018, pp. 1–4, doi: 10.1109/ICTON.2018.8473854.

[10] R. Hristu, L. G. Eftimie, B. Paun, S. G. Stanciu, and G. A. Stanciu, "Pixel-level angular quantification of capsular collagen in second harmonic generation microscopy images of encapsulated thyroid nodules," *J. Biophotonics*, vol. 13, no. 12, Dec. 2020, Art. no. 202000262, doi: 10.1002/jbio.202000262.

[11] R. Hristu, L. G. Eftimie, S. G. Stanciu, D. E. Tranca, B. Paun, M. Sajin, and G. A. Stanciu, "Quantitative second harmonic generation microscopy for the structural characterization of capsular collagen in thyroid neoplasms," *Biomed. Opt. Exp.*, vol. 9, no. 8, p. 3923, Aug. 2018, doi: 10.1364/boe.9.003923.

[12] D. Tokarz, R. Cisek, A. Golaraei, S. Krouglov, R. Navab, C. Niu, S. Sakashita, K. Yasufuku, M. Tsao, L. Sylvia, V. Barzda, and B. C. Wilson, "Tumor tissue characterization using polarization-sensitive second harmonic generation microscopy," *Proc. SPIE*, vol. 9531, pp. 25–30, Jun. 2015, doi: 10.1117/12.2180969.

[13] D. Tokarz, R. Cisek, A. Joseph, S. L. Asa, B. C. Wilson, and V. Barzda, "Characterization of pathological thyroid tissue using polarization-sensitive second harmonic generation microscopy," *Lab. Invest.*, vol. 100, no. 10, pp. 1280–1287, Oct. 2020, doi: 10.1038/s41374-020-0475-7.

[14] D. Tokarz, R. Cisek, A. Golaraei, S. L. Asa, V. Barzda, and B. C. Wilson, "Ultrastructural features of collagen in thyroid carcinoma tissue observed by polarization second harmonic generation microscopy," *Biomed. Opt. Exp.*, vol. 6, no. 9, p. 3475, Sep. 2015, doi: 10.1364/boe.6.003475.

[15] R. Hristu, S. G. Stanciu, D. E. Tranca, L. G. Eftimie, A. Enache, and G. A. Stanciu, "Collagen organization in second harmonic generation images for the assessment of thyroid nodule capsular invasion," in *Proc. 23rd Int. Conf. Transparent Opt. Netw. (ICTON)*, Bucharest, Romania, Jul. 2023, pp. 1–4, doi: 10.1109/ICTON59386.2023.10207435.

[16] Y. Padrez, L. Golubewa, I. Timoshchenko, A. Enache, L. G. Eftimie, R. Hristu, and D. Rutkauskas, "Machine learning-based diagnostics of capsular invasion in thyroid nodules with wide-field second harmonic generation microscopy," *Computerized Med. Imag. Graph.*, vol. 117, Oct. 2024, Art. no. 102440, doi: 10.1016/j.compmedimag.2024.102440.

[17] L. G. Eftimie, Y. Padrez, L. Golubewa, D. Rutkauskas, and R. Hristu, "Widefield polarization-resolved second harmonic generation imaging of entire thyroid nodule sections for the detection of capsular invasion," *Biomed. Opt. Exp.*, vol. 15, no. 8, p. 4705, Aug. 2024, doi: 10.1364/BOE.523052.

[18] A. Dementjev, R. Rudys, R. Karpicz, and D. Rutkauskas, "Optimization of wide-field second-harmonic generation microscopy for fast imaging of large sample areas in biological tissues," *Lithuanian J. Phys.*, vol. 60, no. 3, Aug. 2020, Art. no. 3, doi: 10.3952/physics.v60i3.4301.

[19] Y. Padrez, L. Golubewa, T. Kulahava, T. Vladimirskaja, G. Semenkova, I. Adzerikho, O. Yatsevich, N. Amaegberi, R. Karpicz, Y. Svirko, P. Kuzhir, and D. Rutkauskas, "Quantitative and qualitative analysis of pulmonary arterial hypertension fibrosis using wide-field second harmonic generation microscopy," *Sci. Rep.*, vol. 12, no. 1, May 2022, Art. no. 1, doi: 10.1038/s41598-022-11473-5.

[20] H. Zhao, R. Cisek, A. Karunendiran, D. Tokarz, B. A. Stewart, and V. Barzda, "Live imaging of contracting muscles with wide-field second harmonic generation microscopy using a high power laser," *Biomed. Opt. Exp.*, vol. 10, no. 10, p. 5130, Oct. 2019, doi: 10.1364/boe.10.005130.

[21] K. Mirsanaye, L. U. Castaño, Y. Kamaliddin, A. Golaraei, R. Augulis, L. Kontenis, S. J. Done, E. Žurauskas, V. Stambolic, B. C. Wilson, and V. Barzda, "Machine learning-enabled cancer diagnostics with widefield polarimetric second-harmonic generation microscopy," *Sci. Rep.*, vol. 12, no. 1, p. 10290, Jun. 2022, doi: 10.1038/s41598-022-13623-1.

[22] K. Mirsanaye, L. U. Castaño, Y. Kamaliddin, A. Golaraei, L. Kontenis, E. Zurauskas, R. Navab, K. Yasufuku, M.-S. Tsao, B. C. Wilson, and V. Barzda, "Unsupervised determination of lung tumor margin with widefield polarimetric second-harmonic generation microscopy," *Sci. Rep.*, vol. 12, no. 1, p. 20713, Dec. 2022, doi: 10.1038/s41598-022-24973-1.

[23] L. U. Castaño, K. Mirsanaye, L. Kontenis, S. Krouglov, E. Žurauskas, R. Navab, K. Yasufuku, M.-S. Tsao, M. K. Akens, B. C. Wilson, and V. Barzda, "Wide-field Stokes polarimetric microscopy for second harmonic generation imaging," *J. Biophotonics*, vol. 16, no. 5, May 2023, Art. no. 202200284, doi: 10.1002/jbio.202200284.

[24] D. S. James and P. J. Campagnola, "Recent advancements in optical harmonic generation microscopy: Applications and perspectives," *BME Frontiers*, vol. 2021, Jan. 2021, Art. no. 3973857, doi: 10.34133/2021/3973857.

[25] B. Zhang, H. Shi, and H. Wang, "Machine learning and AI in cancer prognosis, prediction, and treatment selection: A critical approach," *J. Multidisciplinary Healthcare*, vol. 16, pp. 1779–1791, Jun. 2023, doi: 10.2147/jmdh.s410301.

[26] A. Nagpal and V. Singh, "A feature selection algorithm based on qualitative mutual information for cancer microarray data," *Proc. Comput. Sci.*, vol. 132, pp. 244–252, Jun. 2018, doi: 10.1016/j.procs.2018.05.195.

[27] O. O. Oladimeji, H. Ayaz, I. McLoughlin, and S. Unnikrishnan, "Mutual information-based radiomic feature selection with SHAP explainability for breast cancer diagnosis," *Results Eng.*, vol. 24, Dec. 2024, Art. no. 103071, doi: 10.1016/j.rineng.2024.103071.

[28] E. E. M. González, J. A. R. Ortiz, and B. A. G. Beltrán, "Machine learning models for cancer type classification with unstructured data," *Computación y Sistemas*, vol. 24, no. 2, pp. 403–411, Jun. 2020, doi: 10.13053/cys-24-2-3367.

[29] M. Sollini, L. Cozzi, A. Chiti, and M. Kirienko, "Texture analysis and machine learning to characterize suspected thyroid nodules and differentiated thyroid cancer: Where do we stand?" *Eur. J. Radiol.*, vol. 99, pp. 1–8, Feb. 2018, doi: 10.1016/j.ejrad.2017.12.004.

[30] H. R. Kethireddy, A. Tejaswee, L. G. Eftimie, R. Hristu, G. A. Stanciu, and A. Paul, "Differential diagnosis of thyroid tumors through information fusion from multiphoton microscopy images using fusion autoencoder,"attern Recognition (Lecture Notes in Computer Science), vol. 15313, A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C.-L. Liu, S. Bhattacharya, and U. Pal, Eds., Cham, Switzerland: Springer, 2025, pp. 80–93, doi: 10.1007/978-3-031-78201-5_6.

[31] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," in *Data Classification*. London, U.K.: Chapman & Hall, 2014.

[32] M. Wu, Z. Shi, H. Zhang, R. Wang, J. Chu, S. Q. Liu, H. Zhang, H. Bi, W. Huang, R. Zhou, and C. Wang, "Predicting the flavor potential of green coffee beans with machine learning-assisted visible/near-infrared hyperspectral imaging (Vis-NIR HSI): Batch effect removal and few-shot learning framework," *Food Control*, vol. 175, Sep. 2025, Art. no. 111310, doi: 10.1016/j.foodcont.2025.111310.

[33] R. Liu and D. F. Gillies, "Overfitting in linear feature extraction for classification of high-dimensional image data," *Pattern Recognit.*, vol. 53, pp. 73–86, May 2016, doi: 10.1016/j.patcog.2015.11.015.

[34] R. Cao, A. M. Bajgiran, S. A. Mirak, S. Shakeri, X. Zhong, D. Enzmann, S. Raman, and K. Sung, "Joint prostate cancer detection and Gleason score prediction in mp-MRI via FocalNet," *IEEE Trans. Med. Imag.*, vol. 38, no. 11, pp. 2496–2506, Nov. 2019, doi: 10.1109/TMI.2019.2901928.

[35] R. Hristu, S. G. Stanciu, A. Dumitru, L. G. Eftimie, B. Paun, D. E. Tranca, P. Gheorghita, M. Costache, and G. A. Stanciu, "PSHG-TISS: A collection of polarization-resolved second harmonic generation microscopy images of fixed tissues," *Sci. Data*, vol. 9, no. 1, Jul. 2022, Art. no. 1, doi: 10.1038/s41597-022-01477-1.

[36] M. Kociołek, M. Strzelecki, and R. Obuchowicz, "Does image normalization and intensity resolution impact texture classification?" *Computerized Med. Imag. Graph.*, vol. 81, Apr. 2020, Art. no. 101716, doi: 10.1016/j.compmedimag.2020.101716.

[37] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[38] C. Fernandez-Lozano, J. A. Seoane, M. Gestal, T. R. Gaunt, J. Dorado, and C. Campbell, "Texture classification using feature selection and kernel-based techniques," *Soft Comput.*, vol. 19, no. 9, pp. 2469–2480, Sep. 2015, doi: 10.1007/s00500-014-1573-5.

[39] E. W. Steyerberg, "Applications of prediction models," in *Clinical Prediction Models, in Statistics for Biology and Health*. New York, NY, USA: Springer, 2009, pp. 11–31, doi: 10.1007/978-0-387-77244-8_2.

[40] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994, doi: 10.1109/72.298224.

[41] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138, doi: 10.1103/physreve.69.066138.

[42] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[43] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[44] G. Ke, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 1–10. [Online]. Available: https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/

[45] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011, doi: 10.1145/1961189.1961199.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[47] F. Pedregosa, "Scikit-learn: Machine learning in Python," *Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

[48] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, Nov. 2020, doi: 10.1016/j.neucom.2020.07.061.

[49] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," 2016, *arXiv:1603.06560*.

[50] D. S. Soper, "Hyperparameter optimization using successive halving with greedy cross validation," *Algorithms*, vol. 16, no. 1, p. 17, Dec. 2022, doi: 10.3390/a16010017.

[51] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications* (Springer Texts in Statistics), vol. 103, 2nd ed., New York, NY, USA: Springer, 2021, doi: 10.1007/978-1-4614-7138-7.

[52] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: A review with examples from ecology," *Ecological Monographs*, vol. 93, no. 1, Feb. 2023, Art. no. e1557, doi: 10.1002/ecm.1557.

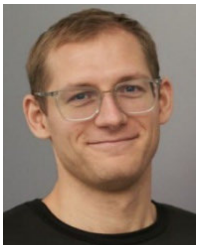
[53] A. C. Müller and S. Guido, *Introduction to Machine Learning With Python: A Guide for Data Scientists*. Sebastopol, CA, USA: O'Reilly Media, 2016.

[54] D. Berrar and P. Flach, "Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them)," *Briefings Bioinf.*, vol. 13, no. 1, pp. 83–97, Jan. 2012, doi: 10.1093/bib/bbr008.

[55] L. B. Ferreira, E. Gimba, J. Vinagre, M. Sobrinho-Simões, and P. Soares, "Molecular aspects of thyroid calcification," *Int. J. Mol. Sci.*, vol. 21, no. 20, p. 7718, Oct. 2020, doi: 10.3390/ijms21207718.

[56] L. Jing, L. Li, Z. Sun, Z. Bao, C. Shao, J. Yan, Q. Pang, Y. Geng, L. Zhang, X. Wang, and Z. Wang, "Role of matrix vesicles in Bone–Vascular cross-talk," *J. Cardiovascular Pharmacol.*, vol. 74, no. 5, pp. 372–378, Nov. 2019, doi: 10.1097/fjc.0000000000000720.

[57] S. De Santis, G. Sotgiu, A. Crescenzi, C. Taffon, A. C. Felici, and M. Orsini, "On the chemical composition of psammoma bodies microcalcifications in thyroid cancer tissues," *J. Pharmaceutical Biomed. Anal.*, vol. 190, Oct. 2020, Art. no. 113534, doi: 10.1016/j.jpba.2020.113534.
[58] A. D. Hofemeier, H. Hachmeister, C. Pilger, M. Schürmann, J. F. W. Greiner, L. Nolte, H. Sudhoff, C. Kaltschmidt, T. Huser, and B. Kaltschmidt, "Label-free nonlinear optical microscopy detects early markers for osteogenic differentiation of human stem cells," *Sci. Rep.*, vol. 6, no. 1, p. 26716, May 2016, doi: 10.1038/srep26716.
[59] K. S. Shin, M. Laohajaratsang, S. Men, B. Figueroa, S. M. Dintzis, and D. Fu, "Quantitative chemical imaging of breast calcifications in association with neoplastic processes," *Theranostics*, vol. 10, no. 13, pp. 5865–5878, 2020, doi: 10.7150/thno.43325.
[60] Z. Li, J. Wei, B. Chen, Y. Wang, S. Yang, K. Wu, and X. Meng, "The role of MMP-9 and MMP-9 inhibition in different types of thyroid carcinoma," *Molecules*, vol. 28, no. 9, p. 3705, Apr. 2023, doi: 10.3390/molecules28093705.
[61] T. R. Cox, "The matrix in cancer," *Nature Rev. Cancer*, vol. 21, no. 4, pp. 217–238, Apr. 2021, doi: 10.1038/s41568-020-00329-7.
[62] A. Bailly, C. Blanc, É. Francis, T. Guillotin, F. Jamal, B. Wakim, and P. Roy, "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models," *Comput. Methods Programs Biomed.*, vol. 213, Jan. 2022, Art. no. 106504, doi: 10.1016/j.cmpb.2021.106504.
[63] T. E. Luvhengo, I. Bombil, A. Mokhtari, M. S. Moeng, D. Demetriou, C. Sanders, and Z. Dlamini, "Multi-omics and management of follicular carcinoma of the thyroid," *Biomedicines*, vol. 11, no. 4, p. 1217, Apr. 2023, doi: 10.3390/biomedicines11041217.
[64] H. G. Lee, Y. J. Choi, and E. C. Chung, "Computed tomography features of follicular thyroid carcinoma: Comparison with papillary thyroid carcinoma," *J. Korean Soc. Radiol.*, vol. 64, no. 1, p. 17, 2011, doi: 10.3348/jksr.2011.64.1.17.



YARASLAU PADREZ (Member, IEEE) received the B.S. and M.S. degrees in nuclear physics and technology from Belarusian State University, Minsk, Belarus, in 2019. He is currently pursuing the joint Ph.D. degree in physics with the State Research Institute Center for Physical Sciences and Technology (FTMC), Vilnius, Lithuania, and Vilnius University. From 2021 to 2023, he was an Engineer, and since 2023, a Research Assistant with the Department of Molecular Compounds Physics, FTMC. His research interests include the development and application of machine learning algorithms for multiphoton and hyperspectral image analysis of biological samples and semiconductor lasers.

RADU HRISTU received the Ph.D. degree in electronics and telecommunications from the National University of Science and Technology POLITEHNICA Bucharest, Romania, in 2012. He is currently an Associate Professor with the Physics Department, National University of Science and Technology POLITEHNICA Bucharest. His research is conducted at the Center for Microscopy-Microanalysis and Information Processing, focusing on developing imaging applications and image analysis techniques for studying semiconductor materials, devices, and biological samples.

IGOR TIMOSHCHENKO received the B.S. and M.S. degrees in theoretical physics from Belarusian State University, Minsk, Belarus, in 2004. His research is conducted at the Centre for Physical Sciences and Technology (FTMC), Vilnius, Lithuania. His research interests include photonics, mathematical modeling, and machine learning.

LUCIAN G. EFTIMIE received the dual Ph.D. degree in electronics and telecommunications from the National University of Science and Technology POLITEHNICA Bucharest, Romania, and the University of Medicine and Pharmacy "Carol Davila," Bucharest, in 2021. He is currently the Head of the Pathology Department, Central Military Emergency University Hospital "Dr. Carol Davila," Bucharest, and is currently an Associate Professor with the National University of Physical Education and Sports, Bucharest, Romania. His research interests include developing novel imaging and image analysis methods to complement traditional histology approaches.

DANIELIS RUTKAUSKAS received the B.S. degree in physics from Vilnius University, Lithuania, in 1999, and the Ph.D. degree in single-molecule micro-spectroscopy from Free University, Amsterdam, The Netherlands, in 2006. Since 2011, he has been a Senior Researcher with the Institute of Physics, Center for Physical Sciences and Technology (FTMC), Vilnius, Lithuania. His research interests include the development and application of various advanced optical microscopy techniques for imaging and analysis of biological systems.

LENA GOLUBEWA (Member, IEEE) received the B.S. and M.S. degrees in biophysics from Belarusian State University (BSU), Minsk, Belarus, in 2011, and the joint Ph.D. degree in physics from the State Research Institute Center for Physical Sciences and Technology (FTMC), Vilnius, Lithuania, and Vilnius University, in 2023. She was a Research Assistant, from 2010 to 2016; and a Researcher with the Laboratory of Cell Engineering and Nanobiotechnology, from 2016 to 2020. From 2014 to 2020, she was a Senior Lecturer with the Department of Higher Mathematics and Mathematical Physics, BSU. She is currently a Researcher with the Department of Molecular Compounds Physics, FTMC. Her research interests include the development of nano-agents for cancer diagnosis and therapy, hyperspectral imaging of biological samples, and the application of machine learning techniques for image analysis.