scientific reports

Check for updates

OPEN

Machine learning-enabled cancer diagnostics with widefield polarimetric second-harmonic generation microscopy

Kamdin Mirsanaye[1,2], Leonardo Uribe Castaño[1,2], Yasmeen Kamaliddin[1,2], Ahmad Golaraei[1,2,3], Renaldas Augulis[4], Lukas Kontenis[5,6], Susan J. Done[3,7,8], Edvardas Žurauskas[9], Vuk Stambolic[3,7], Brian C. Wilson[3,7] & Virginijus Barzda[1,2,5]✉

The extracellular matrix (ECM) collagen undergoes major remodeling during tumorigenesis. However, alterations to the ECM are not widely considered in cancer diagnostics, due to mostly uniform appearance of collagen fibers in white light images of hematoxylin and eosin-stained (H&E) tissue sections. Polarimetric second-harmonic generation (P-SHG) microscopy enables label-free visualization and ultrastructural investigation of non-centrosymmetric molecules, which, when combined with texture analysis, provides multiparameter characterization of tissue collagen. This paper demonstrates whole slide imaging of breast tissue microarrays using high-throughput widefield P-SHG microscopy. The resulting P-SHG parameters are used in classification to differentiate tumor from normal tissue, resulting in 94.2% for both accuracy and F1-score, and 6.3% false discovery rate. Subsequently, the trained classifier is employed to predict tumor tissue with 91.3% accuracy, 90.7% F1-score, and 13.8% false omission rate. As such, we show that widefield P-SHG microscopy reveals collagen ultrastructure over large tissue regions and can be utilized as a sensitive biomarker for cancer diagnostics and prognostics studies.

Cancer is amongst the leading causes of death, affecting approximately 1 in 5 people worldwide; a figure that is predicted to increase by ~47% by 2040[1]. The most common cancer diagnostic techniques rely on the gold standard histopathology of hematoxylin and eosin (H&E) stained tissue sections examined with white light microscopy[2]. H&E histopathology focuses predominantly on characteristics of the cell nuclei and the tissue cell arrangements. However, the ultrastructure and texture of the background collagen-rich extracellular matrix (ECM) can also serve as an additional biomarker.

Collagen is a major constituent of the ECM, which undergoes structural alterations during tumorigenesis[3]. Several stains, including Movat's pentachrome[4], Masson's trichrome[5], picrosirius red[7], as well as immunohistochemical labels[6] have been used to highlight collagen. However, the use of ECM collagen as a biomarker is not widely employed in cancer diagnostics due to subtle structural variations which may be too difficult to detect with white light microscopy.

Second-harmonic generation (SHG) microscopy provides an alternative imaging modality, which enables label-free visualization of the collagenous ECM with high specificity[8,9]. The SHG signal depends on the inherent 3D structure of noncentrosymmetric collagen fibrils, characterized by the nonlinear susceptibility tensor, as well as incident laser polarization[10]. As such, polarimetric SHG (P-SHG) can be utilized in laser scanning microscopy to measure the nonlinear susceptibility tensor elements for each imaged voxel of the tissue. It was further used to identify cancer-associated ECM alterations of multiple human tumor types in lung[11], thyroid[12], breast[13–15], pancreas[16], and ovary[17]. However, current P-SHG microscopy techniques rely on raster scanning of the imaged area, which are slow for whole slide imaging and high-throughput clinical use.

[1]Department of Physics, University of Toronto, Toronto, Canada. [2]Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, Canada. [3]Princess Margaret Cancer Centre, University Health Network, Toronto, Canada. [4]National Centre of Pathology, Vilnius, Lithuania. [5]Laser Research Centre, Faculty of Physics, Vilnius University, Vilnius, Lithuania. [6]Light Conversion, Vilnius, Lithuania. [7]Department of Medical Biophysics, University of Toronto, Toronto, Canada. [8]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada. [9]Department of Pathology, Forensic Medicine and Pharmacology, Faculty of Medicine, Vilnius University, Vilnius, Lithuania. ✉email: virgis.barzda@utoronto.ca

Recently, widefield SHG microscopy was shown to significantly reduce the time required for large-area imaging compared to laser scanning systems[18]. Based on this technology, we have developed a high-throughput quantitative imaging technique by integrating P-SHG with widefield microscopy. Widefield P-SHG microscopy enables rapid scan-less imaging of large sample regions (~ several millimeters) using 16 orthogonal polarization states. Moreover, the subsequent image processing avoids time consuming pixel-by-pixel model fitting, and provides a series of polarimetric parameters that highlight the ultrastructural properties of the collagenous ECM. Furthermore, morphological organization of collagen fibers in the ECM of cancer tissue can be investigated using texture analysis of polarimetric parameters[19–22]. Texture analysis of P-SHG polarimetric parameters was previously utilized in studies of lung cancer, revealing significant features that aid in characterization of tumor tissue[23].

In this work, polarimetric and texture parameters from widefield P-SHG imaging were used to train a logistic regression classifier, which was able to differentiate normal from tumor tissue with 94.2% accuracy and F1-score, and 6.3% false discovery rate, in human breast tissue microarray slides. The trained classifier was then further evaluated by predicting the presence of tumor on an independent data set, yielding 91.3% accuracy, 90.7% F1-score, and 13.8% false omission rate. Implementation of machine learning into the image postprocessing enhances differentiation of normal and tumor tissues, potentially enabling automated screening of, for example, tissue microarrays, as well as mapping areas of altered collagen to improve histopathologic diagnoses.

Results

Experimental design and overview. We have developed a widefield P-SHG microscopy technique for rapid high-throughput quantitative imaging, which we applied to breast tissue microarrays. The method generates a set of information-rich polarimetric and texture parameters, and utilizes these to classify normal and tumor tissues.

Figure 1 shows an overview of the experimental workflow. The tissue microarray slide was imaged under widefield P-SHG microscopy with 16 unique combinations of input and output polarization states (Fig. 1a). These were combined to generate images of SHG Stokes vector elements, in accordance with reduced double Stokes-Mueller polarimetry, as illustrated in Fig. 1b (see Supplementary Information for more details). Images of the following 5 distinct polarimetric parameters were extracted and used to characterize the ultrastructure of collagen in each image pixel: (1) average SHG intensity from all incident circular polarizations, (2) R-ratio, which is the ratio of 2 achiral second-order nonlinear optical susceptibility elements ($\chi^{(2)}_{zzz}/\chi^{(2)}_{zxx}$, where the z-axis is parallel to the collagen fiber axis), (3) degree of circular polarization (DCP), (4) SHG circular dichroism (SHG-CD), and (5) SHG linear dichroism (SHG-LD). Each polarimetric parameter carries unique ultrastructural information about the ECM collagen.

The SHG intensity obtained with circular incident polarization is highly sensitive to the molecular organization of collagen in the focal volume and is independent of the in-plane orientation of collagen fibers. Polarity of neighboring collagen fibrils, tilt angle of fibrils out of the image plane, crossing of fibrils, and distances between the fibrils influence the SHG intensity[10,24,25]. It has been shown that the SHG intensity is reduced in various solid tumors[26,27]. The R-ratio describes the structural organization of the collagen fibers in the focal volume and it is sensitive to the ultrastructure of the fibers, fiber tilt angles out of the image plane and crossing of the fibers. R-ratio has been successfully used to differentiate normal and malignant tissues in lung, breast, thyroid, and pancreas[11–13,16]. The value of DCP is closely related to the R-ratio in non-scattering tissue and reflects the disorder and depolarization in the sample[28,29]. SHG-CD depends on the R-Ratio, out-of-plane fiber tilt angle and polarity, and the phase retardance between the achiral and chiral susceptibility components[10,24,30,31]. SHG-CD has been utilized in investigations of ovarian cancer[25]. Here, we introduce SHG-LD as a new parameter in P-SHG microscopy to study the in-plane organization of collagen fibers in the ECM. Recent investigation show that the collagen fiber orientation computed from SHG-LD is slightly altered due to R-Ratio and complex chiral susceptibilities[24]. A similar definition of SHG-LD was previously used in surface SHG measurements of Langmuir–Blodgett films of chiral polymers functionalized with a nonlinear optical chromophore[32]. In this work, we use SHG intensity of circularly polarized light to assess the presence of aligned collagen fibers in the tissue. R-Ratio is used to determine subtle variations in molecular composition of collagen fibers, and DCP is employed as a measure of depolarization as a result of fiber chirality and out-of-plane orientation. Although complex, SHG-CD and SHG-LD of the ECM are treated as approximations for the out-of-plane and in-plane fiber orientation, assuming the presence of a single collagen chiral handedness throughout.

The images of each computed polarimetric parameter were further subjected to texture and statistical analyses. Texture analysis is a well-known method for characterizing tissue morphology by analyzing the variation in the neighboring pixel values[33]. It is commonly used as a classification tool through its ability to recognize patterns that may be indistinguishable to the human eye[34–36]. Fractal-based texture analysis has shown to be useful in microscopic image analysis[37]. In this work, we will focus on grey-level co-occurrence matrix (GLCM-based) texture analysis. Here, each polarimetric image is comprised of over 4 million pixels. Hence, to enable high-resolution mapping of texture parameters and statistical investigations, all polarimetric images were divided into 64 sub-images (Fig. 1c). The mean, mean absolute deviation (MAD), and 5 of the most useful textures of the polarimetric parameters, including contrast, correlation, entropy, angular second moment (ASM), and inverse difference moment (IDM) were computed over all sub-images to characterize the collagen in the tissue (Fig. 1d)[19,20].

The final stage of the analysis involved machine learning-assisted diagnostics, as shown in Fig. 1e. For this, the combination of mean, MAD and texture of 5 polarimetric parameters of normal and tumor tissue were used to train a logistic regression classifier. The trained classifier was used to perform predictions on an independent set of images and map the ultrastructural properties of collagen, leading to differentiation of normal and tumor tissue.

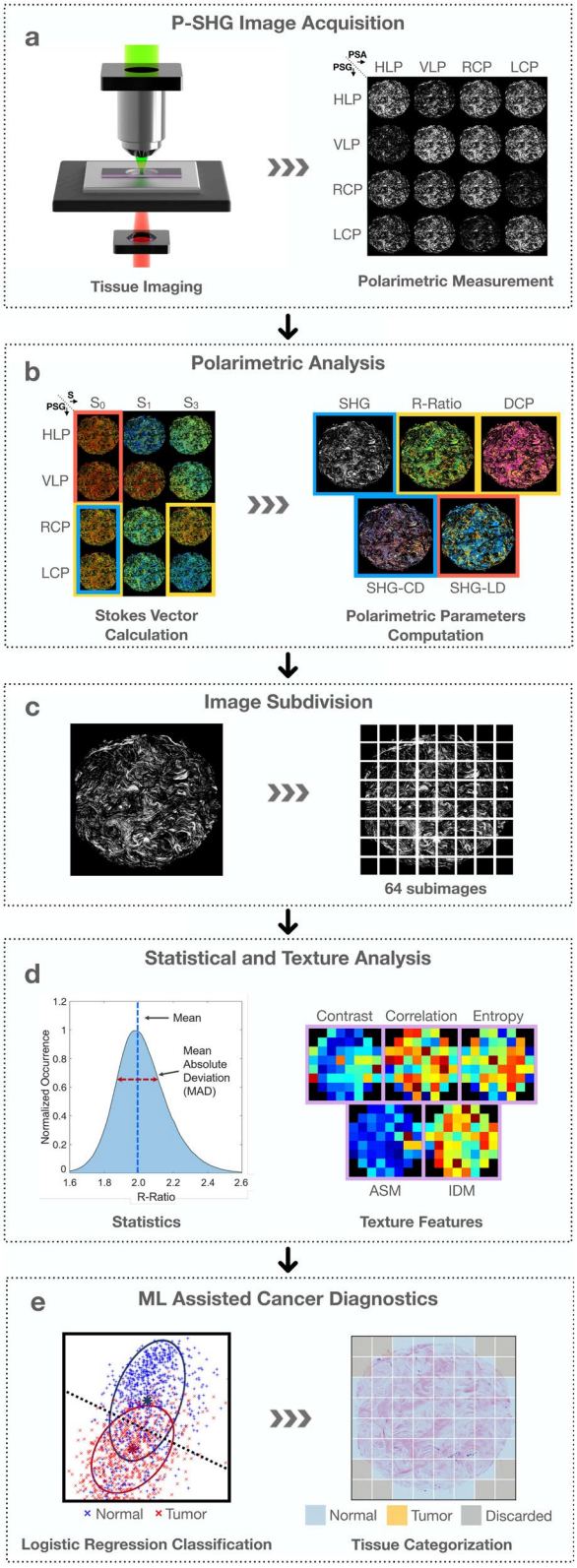

Figure 1. From imaging to classification. (**a**) Widefield polarimetric SHG imaging of the sample at 16 unique polarization state combinations, prepared by the polarization state generator (PSG) and the polarization state analyzer (PSA), which correspond to horizontal linearly polarized (HLP), vertically linearly polarized (VLP), right and left circularly polarized states (RCP and LCP), respectively. (**b**) Calculation of SHG Stokes vector elements to compute polarimetric parameter images. On the left, image shows breast tissue core pseudo-color images of s_0, s_1 and s_3 component at PSG states of HLP, VLP, RCP and LCP respectively. The right image shows polarimetric parameter maps of the core for SHG intensity, R-Ratio, DCP, SHG-CD and SHG-LD. (**c**) Subdivision of polarimetric images into 64 sub-images, to allow high-resolution texture analysis and statistical significance testing. (**d**) Calculation of mean and mean absolute deviation of polarimetric parameter, as well as contrast, correlation, entropy, angular second moment (ASM), and inverse difference moment (IDM) texture parameters of each sub-image. (**e**) Training of a logistic regression classifier using the polarimetric and texture parameters, and the subsequent prediction to differentiate normal and tumor tissue.

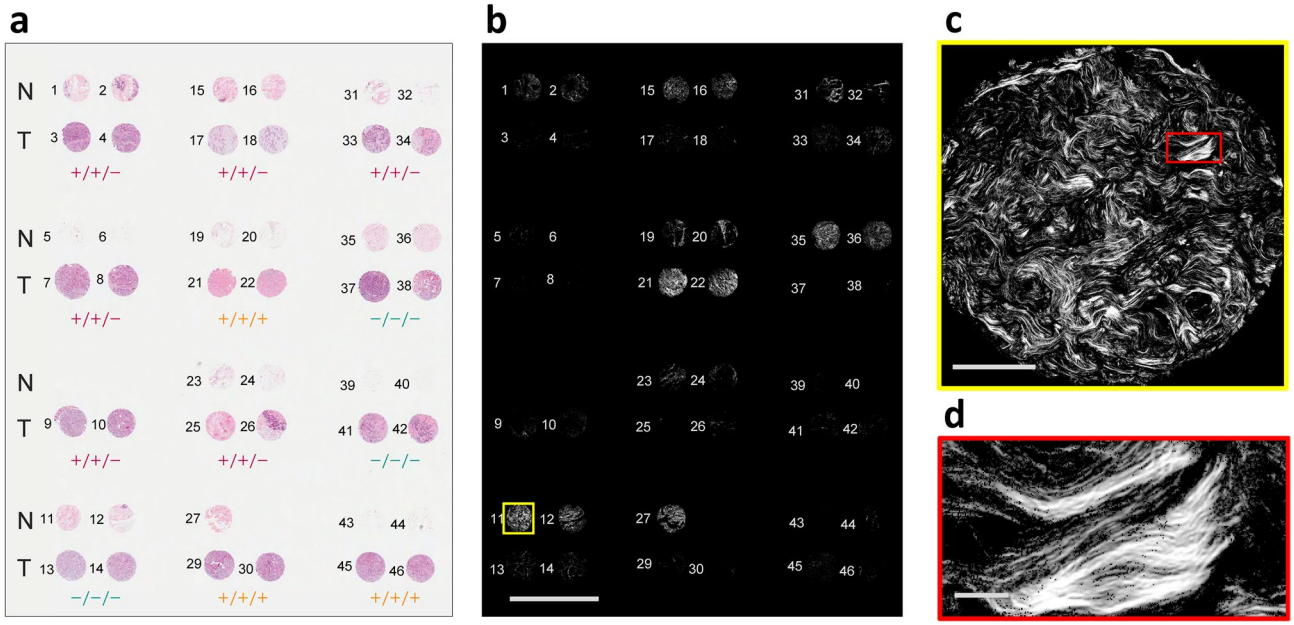

Figure 2. High-throughput widefield P-SHG imaging of tissue microarray. (**a**) A H&E-stained breast tissue microarray containing normal and three distinct tumor subtypes. Enumeration indicates the number assigned to each breast tissue core. The positive sign indicates overexpression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), denoted by ER/PR/HER2. (**b**) Image of circularly-polarized SHG intensity of the entire microarray slide shows distinctive low signal associated with tumor cores, except cores 21 and 22 that possess a large amount of collagenous stroma. Scale bar: 2.5 mm. (**c**) A single widefield SHG image consisting of 2048 × 2048 pixels and an area of approximately 670 μm × 670 μm. Scale bar: 200 μm. (**d**) A magnified region of the tissue, highlighted by red rectangle in **c** shows high resolution of the imaging technique. Scale bar: 25 μm.

Whole slide P-SHG imaging of tissue microarray. The collagen content of a breast tissue microarray was imaged using widefield P-SHG microscopy, resulting in images of whole breast tissue cores that were 0.6 mm in diameter, without the need for raster scanning. Figure 2a shows the annotated H&E-stained microarray of normal (N) and tumor (T) cores. The latter were characterized by estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) expression, and 3 of the most common subtypes were considered: ER+/PR+/HER2+ (triple positive or +/ + /+, also known as HER2 positive), ER+/PR+/HER2− (double positive or +/ + /−, also known as hormone-receptor positive), and ER−/PR−/HER2− (triple negative or −/ − /−)[38]. The breast tissue microarray slide was comprised of 45 tissue cores from 12 different patients. Overall, 15 normal cores and 20 tumor cores (5 triple negative cores, 11 double negative cores, and 4 triple positive cores) were used for further analysis.

The SHG intensity images of all cores are seen in Fig. 2b. Each core image is comprised of 2048 × 2048 pixels, corresponding to an area of approximately 670 μm × 670 μm, as shown in Fig. 2c, which corresponds to core 11, as indicated by yellow border in Fig. 2b. The resolving power of the microscope (approximate pixel size of 0.3 μm × 0.3 μm) is shown by an enlarged rectangular region in Fig. 2d, whose location is highlighted with the red rectangle in Fig. 2c. For better viewing of the SHG images, pixels that exhibited signal-to-noise ratio (SNR) <1 were removed. It is clearly seen that tumor tissue has markedly lower SHG intensity than normal, consistent with previous reports in multiple tumor types[11–13,16,39]. This is likely due to collagen degradation in the tumor microenvironment which results in randomly oriented fibrils, so resembling a centrosymmetric arrangement. It is also clear that there are considerably fewer pixels with SHG signal present in the tumor tissues, rendering pixel density as an important classification parameter.

Polarimetric SHG imaging and analysis. Representative polarimetric images of normal, triple negative, double positive, and triple positive groups are shown in Fig. 3. The H&E-stained core images (Fig. 3a) were segmented to reveal stained fibrillar components of the ECM using a published convolutional neural network (CNN) based technique[40] (Fig. 3b). H&E segmented images serve as a visual approximation of the total collagen content of the tissue. The true arrangement of the ordered collagen fibers can be visualized with P-SHG microscopy as presented in Fig. 3c. SHG identifies only the ordered collagen structures in the tissue; hence, highlighting a subset of the segmented collagen image.

As shown in Fig. 3d, the tumor cores have larger R-ratio values than normal, similar to earlier reports of human breast tissues imaged with a scanning P-SHG microscope[13]. In addition to the extracted R-ratio, DCP of the SHG signal is obtained. As shown in Fig. 3e, the measured DCP was on average lower in normal tissue compared to each of the tumor groups.

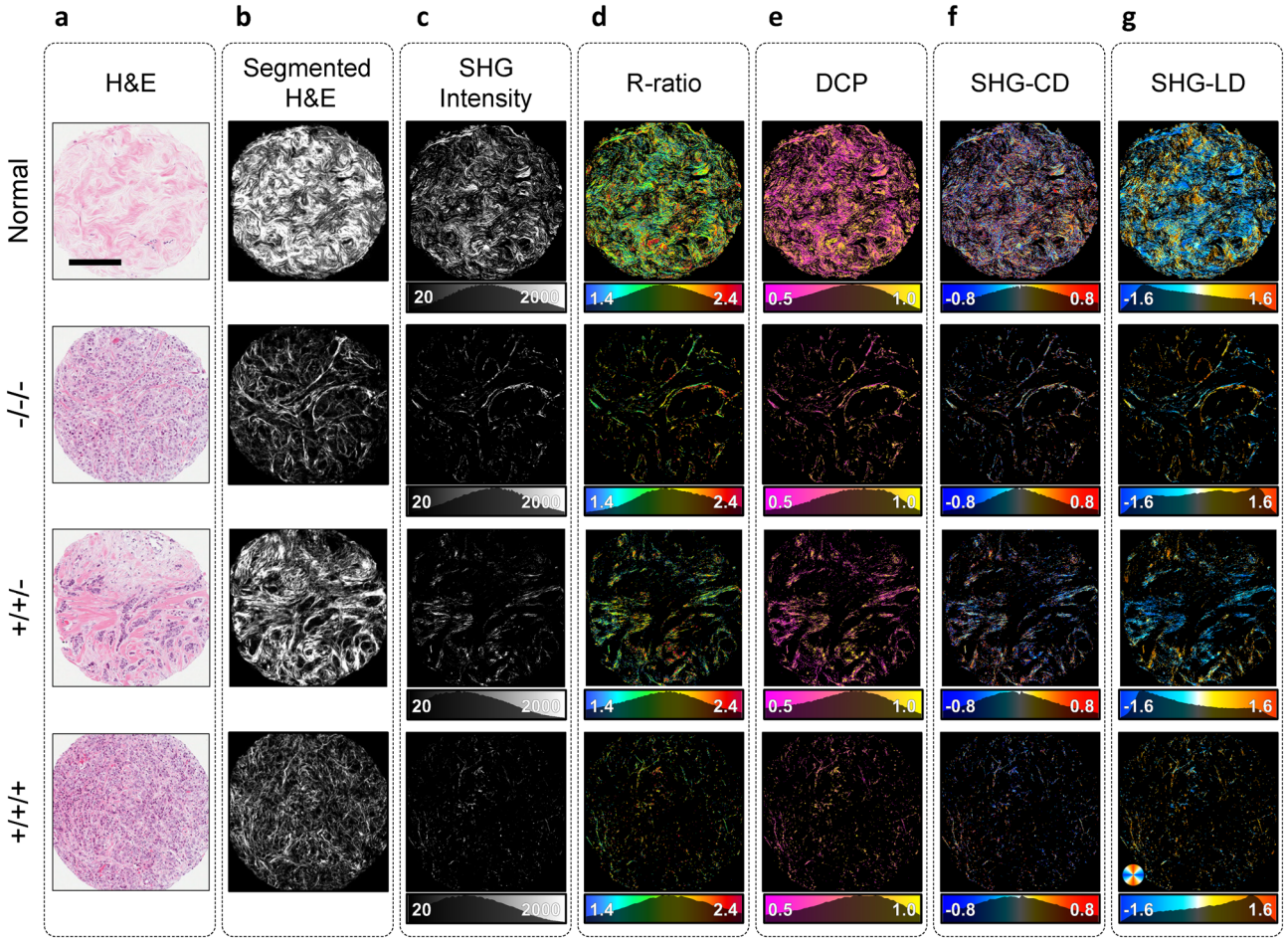

Figure 3. Widefield P-SHG polarimetric parameters. (**a**) Representative H&E white light images of normal and three tumor breast cores with different expressions of the receptors ER/PR/HER2. Scale bar: 200 μm. (**b**) Segmented images display the portion of H&E images stained by Eosin, which serves as a good estimate for total collagen content of the tissue. (**c**) SHG intensity of circularly-polarized incident light, highlighting ordered collagen fibers. Normal core produces much larger SHG signal. (**d**) R-ratio images of cores range from 1.4 to 2.4, indicating the presence of collagen in the tissue. (**e**) Degree of circular polarization images show larger means in tumor due to higher R-ratio and low depolarization. (**f**–**g**) SHG-CD and SHG-LD images depict the out-of-plane and in-plane fiber orientations, respectively. Due to unbiased sectioning of the tissue, collagen fibers are at random orientations, resulting in SHG-CD distributions centered around zero, and SHG-LD distributed randomly.

SHG-CD values (Fig. 3f) were calculated using polarimetric images with circular input polarizations, and are directly associated with chiral and achiral optical susceptibilities, as well as the out-of-image-plane tilt angle of collagen fibers[10,31]. Tissue samples were sectioned without any fiber orientation bias, resulting in random direction of collagen fibers. Therefore, SHG-CD mean values did not provide useful information for classification of tumor tissue. However, collagen in normal tissue appears wavier, leading to a broader distribution of the SHG-CD. Thus, MAD of SHG-CD distribution was instead used to differentiate between normal and tumor tissues.

The SHG-LD images were constructed using the linear input polarizations (Fig. 3g), and highlight the in-plane collagen fiber orientation. Similar to SHG-CD, the SHG-LD distribution appears wider in normal tissue compared to the tumor, reflecting larger variations of the in-plane collagen fiber orientation in normal tissue. The SHG-LD measurements exhibited diverse distribution forms, due to random sectioning of breast cores. Hence, in order to further analyze the degree of variability of in-plane collagen fiber orientation, the MAD of SHG-LD distributions were computed and compared between normal and tumor groups.

Texture analysis and statistical testing. Statistical multiple comparison tests were performed to evaluate the significance of parameter value differences between tumor and normal cores. Each core image was comprised of over 4 million pixels, however, pixels with SNR < 3 were discarded from analysis to ensure reliability of the statistical testing. Each large-area image of the cores was subdivided into 64 sub-images, each with 256×256 pixels (approximately 84 μm×84 μm in area). Sub-images located in the corner of each breast core P-SHG image, as well as those without viable signal were removed from the analysis. The number of pixels with SNR > 3 in each sub-image was used as an important metric, referred to as the SHG pixel density (PD) to indicate the

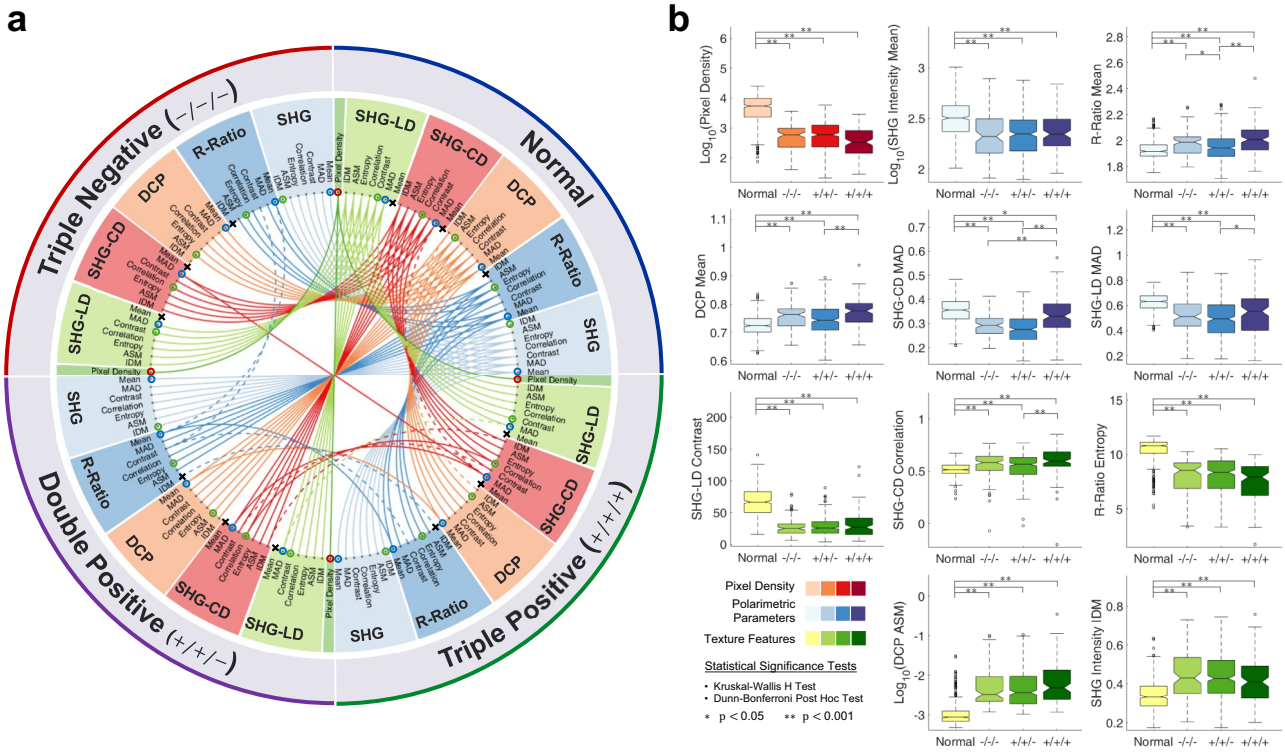

Figure 4. Parameter multiple comparisons testing. (**a**) Circular plot of all multiple comparisons test between all four groups. Kruskal–Wallis H test and Dunn–Bonferroni post hoc test were performed. The dashed links indicate significant differences ($0.001 < p < 0.05$), while solid links depict highly significant differences ($p < 0.001$). All non-significant differences are omitted. It is evident that normal tissue is highly significantly different from all tumor groups across most of the measured parameters. Parameters that are not significantly different between normal and any of the tumor groups are denoted by x's and removed from further investigation. (**b**) Boxplots of selected parameters show detailed differences between all four considered groups. Color of the plots correspond to small circles around parameters in a. It is evident that triple positive is the most different tumor group from normal. Colored areas highlight the interquartile range and the center line denotes the median of the data. The whiskers extend to 1.5× interquartile range, and the circles show the outliers.

abundance of ordered SHG-producing collagen in the tissue. The PD was found to be on average considerably higher in normal tissue than tumor; therefore, it was included in the set of diagnostic polarimetric and texture parameters.

The mean, MAD, and texture features of the polarimetric parameters were used in multiple comparisons testing and machine learning-assisted classification. Given the large number of resulting parameters (36), the significance test results are shown as a circular plot in Fig. 4a. The circular plot is divided into four quadrants, representing normal, triple negative, double positive, and triple positive groups. The mean, MAD, and textures of each polarimetric parameter are highlighted by distinct colors. Most differences between normal and tumor groups were statistically highly significant ($p < 0.001$), as indicated by solid colored links. Links between tumor groups are mostly dashed lines, which indicate significant differences ($0.001 < p < 0.05$). Non-significant differences are omitted from the diagram. It is evident that all three considered molecular subtypes of breast cancer are well-differentiated from normal with comparable levels of accuracy using widefield P-SHG.

For a more detailed presentation of the data, boxplots of subsets of the computed parameters are shown in Fig. 4b. Sub-image means of both R-ratio and DCP showed similar trends of being significantly higher in tumor than normal tissue, with triple positive breast cancer possessing the largest R-ratio and DCP. The R-ratio results are in agreement with a previous investigation on small breast tissue regions[13]. The MAD of both SHG-CD and SHG-LD highlighted larger variation of the collagen fibers orientation in normal tissue, compared to tumor tissue. Between tumor groups, triple positive had the largest variation in fiber orientation, approaching that of the normal group. Three of the parameters (SHG-CD mean, SHG-LD mean, and R-ratio IDM) were not significantly different between normal and any of the tumor groups, as indicated by **x**'s on their corresponding nodes in Fig 4a. As such, these parameters were omitted from classification.

Classification and prediction of normal and tumor breast tissue. To showcase the efficacy of collagen as a diagnostic biomarker and display the detection capabilities of widefield P-SHG microscopy, the polarimetric parameter statistics (mean and MAD) and texture features were used to train a binary logistic regression classifier. Prior to classifier training, two cores (one from normal group and one from tumor group) were removed to form a prediction dataset, which was used to further investigate the predictive power of the classifier.

a

	Performance Metrics	Trained Classifier	Prediction
Threshold-Independent	Brier Score (±0.002)	0.045	0.067
	AUROC (±0.001)	0.984	0.990
Threshold-Dependent (Threshold at 50%)	Accuracy (±0.004)	0.942	0.913
	F1-score (±0.004)	0.942	0.907
	True Negative Rate (±0.005)	0.936	0.980
	True Positive Rate (±0.005)	0.947	0.846
	False Negative Rate (±0.005)	0.053	0.154
	False Positive Rate (±0.005)	0.064	0.020
	Negative Predictive Value (±0.005)	0.947	0.862
	Positive Predictive Value (±0.004)	0.937	0.978
	False Omission Rate (±0.005)	0.053	0.138
	False Discovery Rate (±0.004)	0.063	0.022

b

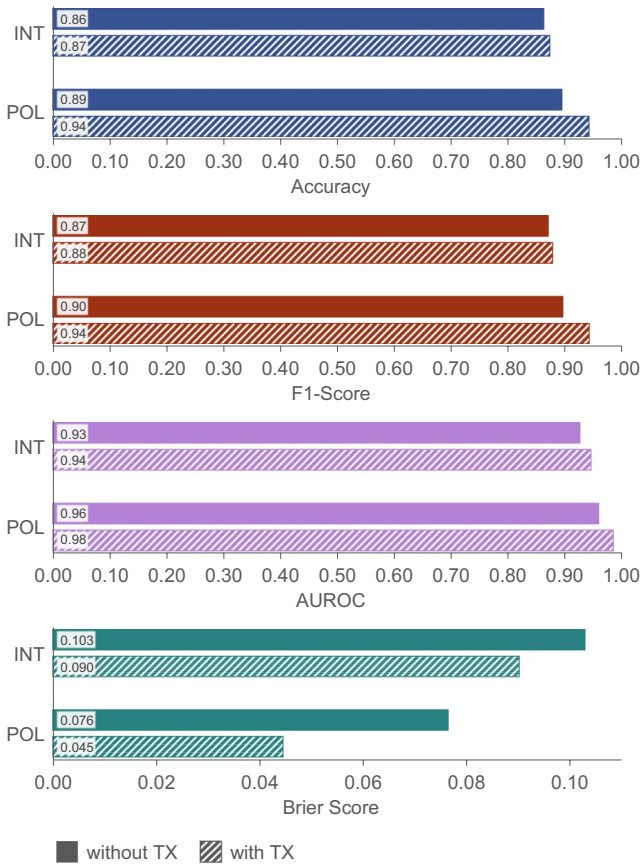

Figure 5. Tissue classification performance. (**a**) Tissue classification and prediction with a binary logistic regression classifier, trained using all polarimetric and texture parameters. Reported values are mean ± standard deviation. (**b**) Testing the impact of various parameters using 4 subsets of the measured parameters, including SHG intensity (INT) and all 5 polarimetric parameters (POL), with and without texture parameters (TX). All groups include pixel density. Differences in accuracy, F1-Score, AUROC, and Brier score between groups are statistically highly significant ($p \ll 0.001$), as determined by the Kruskal–Wallis H test and the Dunn–Bonferroni post hoc test. Using all polarimetric and texture parameters improves the classification performance across all metrics. All reported accuracy and F1-score values have < 0.4% standard error, and AUROC and Brier score have standard errors of < 0.2% and < 4.5%, respectively.

All tumor cores were combined into a single tumor group and by treating each sub-image as an individual data point, 544 normal and 543 tumor data points (samples), across 33 different polarimetric and texture parameters (predictors) were used for classification. The training was repeated 1000 times with random stratified partitioning of the dataset for fivefold cross-validation, and the standard deviation of performance metrics were used to evaluate the classifier stability[41].

Using the complete dataset, the classifier differentiated tumor from normal tissue with 94.2% accuracy at 50% posterior probability threshold, as indicated in Fig 5a. The threshold-dependent classification performance metrics (true negative rate, true positive rate, negative predictive value, and positive predictive value) were within 93–95%. The complimentary metrics (false positive rate, false negative rate, false omission rate, and false discovery rate) were within 5–7%. In addition, the F1-score of 94.2% demonstrates the excellent robustness and accuracy of the trained classifier in identifying tumor tissue regions.

Two threshold-independent metrics, namely the area under the curve of the receiver operating characteristic (AUROC) and Brier score, were computed to further examine the classifier performance. An AUROC value of 1 is ideal, while 0.5 indicates a completely random classification. The Brier score ranges from 0 (perfectly accurate) to 1 (completely inaccurate). As shown in Fig. 5a, the trained classifier generated an AUROC of 0.984 and Brier score of 0.045, indicating well-differentiated normal and tumor groups and low misclassification rate.

To determine the importance of polarimetric and texture parameters in classifier performance, 4 subsets of the data were considered (Fig. 5b): (1) SHG intensity and PD, which are the most commonly used parameters, (2) SHG intensity and PD, together with corresponding texture analysis, (3) the 5 polarimetric parameters (including SHG intensity) and PD, and (4) the 5 polarimetric parameters and PD with corresponding texture parameters (complete dataset). Classification using only the SHG intensity resulted in the lowest accuracy,

F1-score and AUROC values, and the largest Brier Score. Adding texture analysis provided a modest improvement of (∼ 1%) across all metrics. A more significant improvement was evident from using all 5 polarimetric parameters (including the SHG intensity), with accuracy and F1-scores reaching > 90%. Finally, combining all polarimetric and texture parameters, produced the best overall performance with accuracy and F1-score > 94%, AUROC > 98%, and Brier score of 0.045.

The trained classifier was then used on the prediction set of normal and tumor cores that had not been part of the training set. As depicted in Fig. 5a, the prediction highlighted sub-images with normal and tumor properties, resulting in 91.3% accuracy, 0.990 AUROC, and 0.067 Brier score. Normal tissue was better predicted than tumor tissue, with a true negative rate of 98.0% compared to a true positive rate of 84.6%. However, the positive predictive value was 97.8%, resulting in an F1-score of 90.7%.

Discussion. For the first time, we have presented the application of widefield P-SHG microscopy of the collagenous extracellular matrix as a potential histopathological biomarker for cancer diagnostics. The instrument provides rapid high-resolution large-area imaging of tissue samples such as the microarrays shown here, while providing comprehensive ultrastructural information through 33 quantifiable parameters. The lack of moving optics, such as scanning mirrors, in widefield imaging significantly contributes to the robustness of the technology. This enables simple construction and customization of the microscope, that is well suited for user-friendly high-throughput applications.

To assess the effects of high-powered laser exposure on biological tissue during widefield P-SHG microscopy, we have measured both Multi-Photon Excitation Fluorescence (MPF) and SHG, and the details are reported elsewhere[24]. In short, 16 polarization states were measured twice for improved SNR and averaged, sustaining ∼ 8 minutes of laser exposure in the process. This resulted in less than 5% MPF bleaching and an increase in measured SHG intensity to a much lesser extent (less than 1%). The minimal SHG intensity increase is due to bleaching of the eosin dye, which likely reduces the reabsorption of the SHG signal. Despite the high laser powers, the relatively low NA excitation lens illuminates a very large section of the tissue (∼ 1 mm in diameter), thus, effectively reducing the energy density at the region of interest. Without visual signs of tissue damage and with negligible signal change, we conclude that widefield P-SHG microscopy can be used for imaging under optimized experimental conditions.

Overall trends in many of the measured parameters showed a clear distinction between tumor and normal tissues. In particular, triple positive tissue had the greatest differences from normal tissue across most parameters. However, this distinction was small compared to the overall difference between normal and all tumor groups, as further illustrated by multiple comparisons tests on the circular plot of Fig. 4a.

Classification enabled an efficient use of all measured parameters, allowing for accurate predictions of normal and tumor tissue with performance comparable to other methods involving machine learning-assisted cancer diagnostics[42–44]. The pixel density and average SHG intensity were amongst the most informative measured parameters, with substantial differences in the ECM between normal and tumor tissues. It is important to note that simply the absence of SHG signal and, thus, lower pixel density and average SHG intensity are not always indicative of the presence of tumor; sparsely distributed collagen fibers and adipose tissue often do not result in sufficient SHG signal, therefore, such regions must be discarded to avoid inflation of false positive rates. Here, empty regions in the corners of the breast core images, and cores which were mainly comprised of adipose tissue were manually discarded. However, in future developments, H&E images may be used simultaneously with widefield P-SHG to identify and discard such areas automatically. Moreover, most breast cores showed a significant reduction in ordered collagen, as indicated by lower SHG intensity compared to normal tissue. However, 2 cores showed a high degree of stroma, and due to rarity of such tissue types in the microarray, they were discarded from classification training data.

The calculated classification metrics showed robust and accurate tissue differentiation with 94.2% accuracy, 6.3% false discovery rate, AUROC of 0.98, and Brier score of 0.045. Capabilities of the trained classifier were further demonstrated in predicting an independent test dataset with an accuracy of 91.3%, false omission rate of 13.8%, AUROC of 0.990, and a Brier score of 0.067. It is important to note that the number and size of the sub-images used, affect the predictive power of the classifier. For example, subdividing the images into fewer but larger sub-images would improve the accuracy and specificity, at the cost of decreased classification stability, resulting in reduced reliability. We found that subdivision level of 64 sub-images per image delivers highly accurate classification with sufficient robustness. We present a detailed investigation of the optimum number of sub-images for classification of the P-SHG data in the supplementary information document (see Supplementary Fig. 2).

A natural extension to this work is introduction of chiral nonlinear susceptibility components. Consequently, additional polarimetric parameters such as chiral second order nonlinear optical susceptibility ratio ($C = \chi^{(2)}_{xyz}/\chi^{(2)}_{zxx}$) may be introduced to further improve molecular identification, based on varied polarity of the collagen fibers throughout the tissue sample. The C-ratio has been previously investigated in scanning P-SHG systems, revealing information on collagen fiber polarity and organization[10,48].

In this study, we demonstrated widefield P-SHG microscopy as a potential tool in cancer diagnostics. However, it is important to also identify the applications of the microscope in studying other pathologies related to remodeling of the collagen structure in the ECM and in other collagenous tissues. As an example, widefield P-SHG microscopy may be used to identify large-scale ultrastructural changes in pathologies relating to abnormalities in fibrillar collagen types I, II, and III such as, arterial aneurysms, chondrodysplasias, osteogenesis imperfecta, osteoporosis, osteoarthrosis, intervertebral disc disease, and Ehlers–Danlos syndrome[45]. Scanning P-SHG has also been used to study muscle ultrastructure in rat and drosophila larvae[46,47]. Future applications of the widefield P-SHG may include high-throughput investigations of human muscle pathologies such as muscular dystrophy and multiple sclerosis. In addition, continuous and uniform illumination of the entire imaged area provided

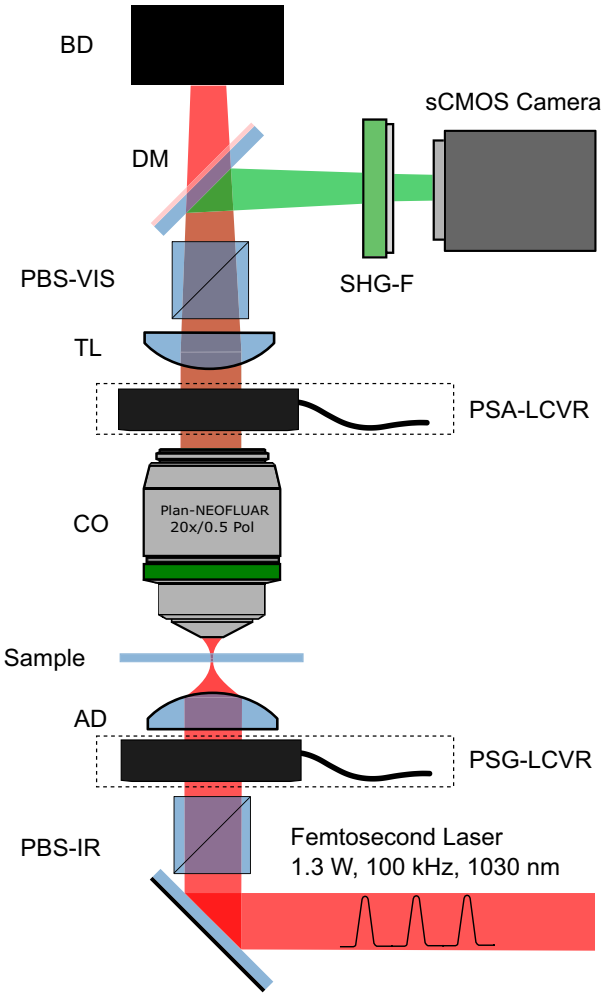

Figure 6. Widefield P-SHG microscope setup. Schematic representation of the experimental setup. The laser beam denoted in red is focused before the sample after passing polarizing beam splitter (PBS-IR), liquid crystal variable retarder of polarization state generator (PSG-LCVR) and achromatic doublet (AD). The produced SHG signal, depicted in green is collected by a collection objective (CO), passed through the liquid crystal variable retarder of polarization state analyzer (PSA-LCVR), tube lens (TL), polarization beam splitter (PBS-VIS), dichroic mirror (DM), filter (SHG-F), and projected onto the camera. The laser beam is terminated at a beam dump (BD).

by the widefield P-SHG microscope, enables dynamic imaging of processes involving fast kinetics, such as live ultrastructural imaging of contracting muscle fibers[18] and deformation of collagen fibers during application of external forces[51].

Methods

Widefield P-SHG microscopy. A custom microscope was designed and constructed for widefield polarimetric SHG imaging as shown in Fig. 6. A high-power amplified laser (PH1-15W, Light Conversion) was used for large area illumination. The laser power was set to 1.3 W, at 100 kHz repetition rate, and 13 μJ pulse energy. The laser beam was collimated to 4 mm diameter and coupled to the microscope. The beam passed through an infrared polarizing cube (PBS102, Thorlabs) to produce a linearly polarized state, followed by an infrared liquid crystal variable retarder (LCC1223-B, Thorlabs), referred to as the polarization state generator (PSG-LCVR), which was positioned with its fast axis at 45° to the incident linear polarization. An achromatic doublet (AC254-030-AB, Thorlabs) with a focal length of 30 mm was used to focus the beam on the sample. Widefield imaging was achieved by placing the sample above the focal plane and adjusting the illumination area (~1 mm in diameter) through axial translation of the excitation achromatic doublet. A Plan-Neofluar 20×/0.50 air collection objective (420350-9900-000, Zeiss) was used to collect the SHG signal radiated in the forward direction. The polarization state analyzer (PSA-LCVR), comprised of a LCVR in the visible wavelength range (LCC1223-A, Thorlabs), was used to probe the outgoing polarization of the SHG signal (also positioned with its fast axis at 45° to the incident linear polarization). The SHG signal was passed through a tube lens (452960-0000-000, Zeiss) and a visible range polarizing beam splitter (PBS251, Thorlabs). A dichroic mirror (FF662-FDi02-t3-25x36, Semrock) was used to separate the visible SHG signal from infrared fundamental light. The infrared light trans-

mitted through the dichroic mirror was blocked by a beam dump. The reflected SHG signal was filtered by two BG40 colored glass filters (FGB37-A, Thorlabs), a 515/10 nm interference filter (65-153, Edmund Optics), and projected onto a sCMOS camera (ORCA-Flash 4.0 V2, Hamamatsu).

To carry out polarimetric measurements, four orthogonal incoming polarization states were generated including left circular polarized (LCP), horizontally linearly polarized (HLP), right circularly-polarized (RCP), and vertically linearly polarized (VLP), corresponding to quarter-wave ($\lambda/4$), half-wave ($\lambda/2$), three-quarter-wave ($3\lambda/4$), and full-wave (λ) retardances of the PSG-LCVR, respectively. For each incoming polarization state, four SHG signal polarization states (set by the PSA-LCVR) corresponding to the same set of retardance values (quarter-wave, half-wave, three-quarter-wave and full-wave) were measured, resulting in 16 combinations of polarization states. Each polarization state was imaged with a 10 s exposure time, and 5 s of delay time was used for switching polarization states between exposure times, so that polarimetric measurement of each 670 μm$\times$670 μm imaged area was completed in 4 min. Therefore, P-SHG imaging of the entire microarray was achieved in approximately 3 h. It is important to mention that whole-microarray imaging time may be significantly reduced by decreasing the polarization switch delay time to <1 s, in which case, whole-array P-SHG measurement would be complete in <2.5 h.

Polarimetric parameter calculations. The polarization state of the SHG signal can be characterized by a Stokes vector[49,50]

$$s = (s_0, s_1, s_2, s_3)^T, \tag{1}$$

where s_0 is the total intensity, s_1, s_2 are the linearly polarized Stokes vector components, and s_3 is the circularly-polarized Stokes vector component. The measured SHG of the 16 polarization state combinations were used to compute SHG Stokes vector elements. Using these elements, polarimetric parameters of interest were calculated to reveal ultrastructural properties of interest, including: the average SHG intensity produced with circularly-polarized incident light; a ratio of two achiral second order nonlinear optical susceptibility tensor elements, $\chi^{(2)}_{zzz}/\chi^{(2)}_{zxx}$ (R-ratio), where z points along the fiber axis; the degree of circular polarization (DCP); SHG circular dichroism (SHG-CD); and SHG linear dichroism (SHG-LD). R-ratio and DCP have been shown previously to provide valuable information on the ultrastructural and molecular organization of collagen fibers, while SHG-CD and SHG-LD probe the three-dimensional fiber orientation[11–13,16,48,51,52]. For detailed calculation of all the Stokes vector elements, refer to supplementary information document.

The orientation-independent SHG, I^{CP}, of each core was computed from the average SHG signal from right and left (RCP and LCP) circular polarizations of the incident light

$$I^{CP} = \frac{1}{2}(s_0^{RCP} + s_0^{LCP}), \tag{2}$$

where the superscripts refer to the PSG polarization states, and the subscripts identify Stokes vector components of SHG signal. The R-ratio was calculated in terms of 1st and 4th elements of the measured SHG Stokes vector[55]:

$$R = \frac{\chi^{(2)}_{zzz}}{\chi^{(2)}_{zxx}} = 1 + 2\left(\frac{s_0^{RCP} + s_0^{LCP}}{s_3^{RCP} - s_3^{LCP}}\right) + 2\sqrt{\left(\frac{s_0^{RCP} + s_0^{LCP}}{s_3^{RCP} - s_3^{LCP}}\right)^2 - 1}. \tag{3}$$

DCP was defined as the average of the magnitude of circular Stokes element, over the total intensity, such that[56]

$$DCP = \frac{1}{2}\left(\frac{|s_3^{RCP}|}{s_0^{RCP}} + \frac{|s_3^{LCP}|}{s_0^{LCP}}\right). \tag{4}$$

In addition, we computed SHG-CD, which was conventionally expressed as[53–55]

$$SHG_{CD} = 2\left(\frac{s_0^{RCP} - s_0^{LCP}}{s_0^{RCP} + s_0^{LCP}}\right), \tag{5}$$

and SHG-LD as:

$$SHG_{LD} = 2\left(\frac{s_0^{VLP} - s_0^{HLP}}{s_0^{VLP} + s_0^{HLP}}\right), \tag{6}$$

where both SHG-CD and SHG-LD range from − 2 to 2.

Texture analysis. Textural parameters were extracted from a grey-level co-occurrence matrix (GLCM), which is a second-order statistical representation of the grey-level distribution in a region of interest[33]. The GLCM was built by counting the occurrence of a pixel of grey level i followed by a pixel of grey level j at a distance d along a direction specified by angle θ. In the case of nearest neighbors ($d = 1$), the four angles of analysis were $\theta = \{0°, 45°, 90°, 135°\}$. GLCMs of all four angles were constructed and averaged, forming a direction-independent GLCM. The probability density function, $P_{d,\theta}(i,j)$, of finding certain grey-level pairs was obtained through normalization of the resulting GLCM. From this, we calculated the five textural parameters of interest, as described by Haralick et.al[33].

A custom-built texture analysis program was written in MATLAB, taking advantage of the available functions. The size of the GLCM depends on the number of grey levels (N_g) specified by the user. Since the program runtime was highly dependent on N_g, a testing routine was carried out to determine the optimal N_g, represented by discretization level of continuous polarimetric parameters (refer to supplementary information and Supplementary Fig. 1 for detailed calculation of N_g).

The presence of significant background in widefield P-SHG imaging results in highly skewed texture parameter distributions, leading to reduced differentiation between the groups. As such, the background pixels were converted to "not a number" (NaN) and were not included in the analysis. Consequently, the texture parameters only reflect the structure of the signal-producing entities, such as collagen fibers.

The measured texture parameters were contrast, correlation, entropy, angular second moment (ASM) and inverse difference moment (IDM). The contrast, ν, is defined by:

$$\nu = \sum_{i,j=0}^{N_g-1} (i-j)^2 P_{d,\theta}(i,j), \tag{7}$$

which is a measure of pixel value differences between grey-level pairs. Contrast is highly sensitive to variation in neighboring pixel values and it can be used to probe collagen fiber density in the tissue.

Correlation, ρ, quantifies a linear dependence between grey-level pairs, is expressed as:

$$\rho = -\sum_{i,j=0}^{N_g-1} P_{d,\theta}(i,j)\left[\frac{(i-\mu)(j-\mu)}{\sigma^2}\right], \tag{8}$$

where μ is the mean and σ is the standard deviation of the grey levels. The correlation ranges from − 1 to 1 for perfectly negatively or perfectly positively correlated images, respectively. In the context of widefield polarimetric SHG microscopy, correlation may be used to showcase well-defined structures and patterns in the acquired images.

Entropy, S, is defined by:

$$S = -\sum_{i,j=0}^{N_g-1} P_{d,\theta}(i,j)\log_2\left(P_{d,\theta}(i,j)\right), \tag{9}$$

which measures the level of disorder or lack of spatial organization of the grey levels. Entropy increases as disorder increases; however, it is also indicative of the size of the pixel clusters, which often represent bundles of collagen fibers.

Conversely, angular second moment (ASM) is a measure of orderliness and ranges from 0 to 1, where 1 is characteristic of a completely uniform image. ASM also varies with the size of pixel clusters of comparable values and is a measure of the average size of similar collagen fiber bundles in an image. ASM is expressed as:

$$ASM = \sum_{i,j=0}^{N_g-1} \left(P_{d,\theta}(i,j)\right)^2. \tag{10}$$

Finally, the inverse difference moment (IDM) describes the homogeneity of an image. Similar to ASM, the range of IDM is from 0 to 1, where 1 is indicates of a completely uniform image. IDM is provided by:

$$IDM = \sum_{i,j=0}^{N_g-1} \frac{P_{d,\theta}(i,j)}{1+(i-j)^2}. \tag{11}$$

Statistical analysis and machine learning. Each polarimetric image was divided into 64 sub-images for high-resolution texture analysis, and to compute the statistics of polarimetric parameters. The performance and stability of the logistic regression classifier in differentiating tumor from normal tissue were considered to establish the optimal discretization level of 64 sub-images for each polarimetric image (see supplementary information and Supplementary Fig. 2 for more detail). The results indicated a range of optimal discretization levels from 4 to 64 sub-images per image. The latter was chosen, since it results in a large AUROC and Brier scores, while performing with the highest stability.

Sub-images without SHG signal were discarded and the mean and mean absolute deviation (MAD) of all remaining sub-images were computed using standard MATLAB toolboxes. To reduce the effects of extreme outliers, data points below the 1st, and above the 99th percentiles of parameters in each group were discarded. These statistics were then used in further multiple comparisons tests to establish statistical significance between various groups. There were 544 normal, 103 triple negative, 325 double positive, and 115 triple positive data points across 36 polarimetric and texture parameters. Most parameters did not form normal distributions as indicated by Q–Q plots, and the Shapiro–Wilk test. Hence, MATLAB's Kruskal–Wallis H test with 1086 degrees of freedom, along with the Dunn–Bonferroni post hoc test, were used to determine the significance of the difference between normal and all three tumor groups, for all polarimetric and texture parameters, separately. Differences with p-values < 0.05 and < 0.001 were considered to be significant and highly significant, respectively.

Breast tissue microarray. All human tissues used in this study were obtained with written informed patient consent. The study was approved by the Research Ethics Board of University Health Network, Toronto, Canada. The study was performed in accordance with relevant guidelines and regulations and in compliance with the tenets of the Declaration of Helsinki. The breast tissue microarray slide contained 45 tissue cores from 12 different patients, collected using a 0.6 mm diameter biopsy needle and stabilized in phosphate-buffered formalin. The sample was embedded in paraffin and stored at room temperature. The cores were cut to 5 μm thickness, stained with H&E and imaged with a bright-field microscope scanner (Aperio Whole Slide Scanner, Leica Biosystems) for reference. The microarray contained normal tissue, as well as tissue from 3 different tumor subtypes, including ER−/PR−/HER2− (triple negative or −/ − /−), ER+/PR+/HER2− (double positive or +/ + /−, also known as hormone-receptor positive), and ER+/PR+/HER2+ (triple positive or +/ + /+, also known as HER2 positive), where ER, PR, and HER2 indicate estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2, respectively. Cores that were mainly comprised of adipose tissue did not produce significant SHG signal and were removed from the analysis. In addition, two tumor cores containing significant collagenous stroma with minor tumor foci were excluded, due to their rarity in the microarray. In total, 15 normal cores, as well as 5 triple negative cores, 11 double negative cores, and 4 triple positive cores (20 tumor cores in total) were considered. The tissue assessment and tumor identification were performed by expert pathologists (S.J.D. and E.Z.).

Data availability

All data generated and analysed during this study are included in this published article and its supplementary information files. Raw image files are available from V.B. upon request.

Code availability

The custom analysis software, including the complete data and instructions are available on GitHub: https://github.com/kamdinmirsanaye/Widefield_P-SHG_of_BreastTissue.

Received: 2 November 2021; Accepted: 3 May 2022
Published online: 18 June 2022

References

1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **10**, 0–41 (2021).
2. Rosai, J. Why microscopy will remain a cornerstone of surgical pathology. *Lab. Investig.* **87**, 403–408 (2007).
3. Frantz, C., Stewart, K. M. & Weaver, V. M. The extracellular matrix at a glance. *J. Cell Sci.* **123**, 4195–4200 (2010).
4. Whatcott, C. *et al.* Desmoplasia in primary tumors and metastatic lesions of pancreatic cancer. *Clin. Cancer Res.* **21**(15), 3561–3568 (2015).
5. Rochette, A. *et al.* Asporin is a stromally expressed marker associated with prostate cancer progression. *Br. J. Cancer* **116**(6), 775–784 (2017).
6. Zhou, Z. H. *et al.* Reorganized collagen in the tumor microenvironment of gastric cancer and its association with prognosis. *J. Cancer* **8**(8), 1466–1476 (2017).
7. Lattouf, R. *et al.* Picrosirius Red staining: A useful tool to appraise collagen networks in normal and pathological tissues. *J. Histochem. Cytochem.* **62**(10), 751–758 (2014).
8. Freund, I., Deutsch, M. & Sprecher, A. Connective tissue polarity. Optical second-harmonic microscopy, crossed-beam summation, and small-angle scattering in rat-tail tendon. *Biophys. J.* **50**(4), 693–712 (1986).
9. Campagnola, P. *et al.* Three-dimensional high-resolution second-harmonic generation imaging of endogenous structural proteins in biological tissues. *Biophys. J* . **82**(1), 493–508 (2002).
10. Golaraei, A. *et al.* Complex susceptibilities and chiroptical effects of collagen measured with polarimetric second-harmonic generation microscopy. *Sci. Rep.* **9**, 12488 (2019).
11. Golaraei, A. *et al.* Characterization of collagen in non-small cell lung carcinoma with second harmonic polarization microscopy. *Biomed. Opt. Express* **5**(10), 3562–3567 (2014).
12. Tokarz, D. *et al.* Ultrastructural features of collagen in thyroid carcinoma tissue observed by polarization second harmonic generation microscopy. *Biomed. Opt. Express* **6**, 3475–3481 (2015).
13. Golaraei, A. *et al.* Changes of collagen ultrastructure in breast cancer tissue determined by second-harmonic generation double Stokes–Mueller polarimetric microscopy. *Biomed. Opt. Express* **7**, 4054–4068 (2016).
14. Mercatelli, R., Triulzi, T., Pavone, F., Orlandi, R. & Cicchi, R. Collagen ultrastructural symmetry and its malignant alterations in human breast cancer revealed by polarization- resolved second-harmonic generation microscopy. *J. Biophoton.* **13**(8), e202000159 (2020).
15. Tsafas, V. *et al.* Polarization-dependent second-harmonic generation for collagen-based differentiation of breast cancer samples. *J. Biophoton.* **13**(10), e202000180 (2020).
16. Tokarz, D. *et al.* Characterization of pancreatic cancer tissue using multiphoton excitation fluorescence and polarization-sensitive harmonic generation microscopy. *Front. Oncol.* **9**, 272 (2019).
17. Campbell, K. R., Chaudhary, R., Handel, J. M., Patankar, M. S. & Campagnola, P. J. Polarization-resolved second harmonic generation imaging of human ovarian cancer. *J. Biomed. Opt.* **23**(6), 1–8 (2018).
18. Zhao, H. *et al.* Live imaging of contracting muscles with widefield second harmonic generation microscopy using a high power laser. *Biomed. Opt. Express* **10**, 5130–5135 (2019).
19. Mostaço-Guidolin, L. *et al.* Evaluation of texture parameters for the quantitative description of multimodal nonlinear optical images from atherosclerotic rabbit arteries. *Phys. Med. Biol.* **56**(16), 5319–5334 (2011).
20. Mostaço-Guidolin, L. *et al.* Collagen morphology and texture analysis: From statistics to classification. *Sci. Rep.* **3**(1), 2190–2190 (2013).
21. Cicchi, R. *et al.* Scoring of collagen organization in healthy and diseased human dermis by multiphoton microscopy. *J. Biophoton.* **3**(1–2), 34–43 (2010).
22. Cicchi, R. *et al.* From molecular structure to tissue architecture: Collagen organization probed by SHG microscopy. *J. Biophoton.* **6**(2), 129–142 (2013).
23. Golaraei, A. *et al.* Polarimetric second-harmonic generation microscopy of the hierarchical structure of collagen in stage I–III non-small cell lung carcinoma. *Biomed. Opt. Express* **11**, 1851 (2020).

24. Castaño, Uribe *et al.* Wide-field multi contrast nonlinear microscopy for histopathology. *Preprint*. https://doi.org/10.1101/2022.04.16.488489 (2022).
25. Campbell, K. R. & Campagnola, P. J. Wavelength-dependent second harmonic generation circular dichroism for differentiation of col I and col III isoforms in stromal models of ovarian cancer based on intrinsic chirality differences. *J. Phys. Chem. B* **121**(8), 1749–1757 (2017).
26. Lin, S. J. *et al.* Discrimination of basal cell carcinoma from normal dermal stroma by quantitative multiphoton imaging. *Opt. Lett.* **31**(18), 2756–2758 (2006).
27. Nadiarnykh, O., LaComb, R. B., Brewer, M. A. & Campagnola, P. J. Alterations of the extracellular matrix in ovarian cancer studied by Second Harmonic Generation imaging microscopy. *BMC Cancer* **10**, 94 (2010).
28. Mazumder, N. *et al.* Stokes vector based polarization resolved second harmonic microscopy of starch granules. *Biomed. Opt. Express* **4**(4), 538–547 (2013).
29. Mazumder, N. & Kao, F.-J. Stokes polarimetry-based second harmonic generation microscopy for collagen and skeletal muscle fiber characterization. *Lasers Med. Sci.* **36**(6), 1161–1167 (2020).
30. Lee, H. *et al.* Chiral imaging of collagen by second-harmonic generation circular dichroism. *Biomed. Opt. Express* **4**(6), 909–916 (2013).
31. Schmeltz, M. *et al.* Circular dichroism second-harmonic generation microscopy probes the polarity distribution of collagen fibrils. *Optica* **7**(11), 1469–1476 (2020).
32. Verbiest, T., Kauranen, M. & Maki, J. J. Linearly polarized probes of surface chirality. *J. Chem. Phys.* **103**, 8296 (1995).
33. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **SMC–3**, 610–621 (1973).
34. Kupidura, P. The comparison of different methods of texture analysis for their efficacy for land use classification in satellite imagery. *Remote Sens.* **11**, 1233 (2019).
35. Kabir, S., He, D. C., Sanusi, M. A. & Wan Hussina, W. M. A. Texture analysis of IKONOS satellite imagery for urban land use and land cover classification. *Imaging Sci. J.* **58**, 163–170 (2010).
36. Bharati, M. H., Liu, J. J. & MacGregor, J. F. Image texture analysis: Methods and comparisons. *Chemom. Intell. Lab. Syst.* **72**, 57–71 (2004).
37. Cross, S. S. The application of fractal geometric analysis to microscopic images. *Micron* **25**(1), 101–113 (1994).
38. DeSantis, C. E. *et al.* Breast cancer statistics, 2019. *CA A Cancer J. Clin.* **69**, 438–451 (2019).
39. Hompland, T., Erikson, A., Lindgren, M., Lindmo, T. & de Lange Davies, C. Second-harmonic generation in collagen as a potential cancer diagnostic parameter. *J. Biomed. Opt.* **13**(5), 054050 (2008).
40. Keikhosravi, A. *et al.* Non-disruptive collagen characterization in clinical histopathology using cross-modality image synthesis. *Commun. Biol.* **3**(1), 414–414 (2020).
41. Kim, J. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput. Stat. Data Anal.* **53**(11), 3735–3745 (2009).
42. Zheng, B., Yoon, S. & Lam, S. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst. Appl.* **41**(4), 1476–1482 (2004).
43. Statnikov, A. *et al.* A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* **21**(5), 631–643 (2005).
44. Chen, C. *et al.* Deep learning in label-free cell classification. *Sci. Rep.* **6**(1), 21471–21471 (2016).
45. Myllyharju, J. & Kivirikko, K. Collagens and collagen-related diseases. *Ann. Med. (Helsinki)* **33**(1), 7–21 (2001).
46. Samim, M., Prent, N., Dicenzo, D., Stewart, B. & Barzda, V. Second harmonic generation polarization properties of myofilaments. *J. Biomed. Opt.* **19**(5), 056005–056005 (2014).
47. Dubreuil, M. *et al.* Polarization-resolved second harmonic microscopy of skeletal muscle in sepsis. *Biomed. Opt. Express* **9**(12), 6350–6358 (2018).
48. Mirsanaye, K. *et al.* Polar organization of collagen in human cardiac tissue revealed with polarimetric second-harmonic generation microscopy. *Biomed. Opt. Express* **10**(10), 5025–5030 (2019).
49. Shi, Y., McClain, W. M. & Harris, R. A. Generalized Stokes–Mueller formalism for two-photon absorption, frequency doubling, and hyper-Raman scattering. *Phys. Rev. A* **49**, 1999–2015 (1994).
50. Samim, M., Krouglov, S. & Barzda, V. Nonlinear Stokes–Mueller polarimetry. *Phys. Rev. A* **93**, 013847 (2016).
51. Gusachenko, I. *et al.* Polarization-resolved second-harmonic generation in tendon upon mechanical stretching. *Biophys. J* . **102**(9), 2220–2229 (2012).
52. Tuer, A. E. *et al.* Nonlinear optical properties of type I collagen fibers studied by polarization dependent second harmonic generation microscopy. *J. Phys. Chem. B* **115**(44), 12759–12769 (2011).
53. Petralli-Mallow, T., Wong, T., Byers, J., Yee, H. & Hicks, J. Circular dichroism spectroscopy at interfaces: A surface second harmonic generation study. *J. Phys. Chem.* **97**(7), 1383–1388 (1952).
54. Verbiest, T. *et al.* Nonlinear optical activity and biomolecular chirality. *J. Am. Chem. Soc.* **116**(20), 9203–9205 (1994).
55. Golaraei, A., Kontenis, L., Karunendiran, A., Stewart, B. A. & Barzda, V. Dual- and single-shot susceptibility ratio measurements with circular polarizations in second-harmonic generation microscopy. *J. Biophoton.* **13**(4), e201960167 (2020).
56. Chipman, R. A. *et al. Handbook of Optics Ch. 15* (McGraw-Hill, 1995).

Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2017-06923, DGDND-2017-00099, CHRPJ 462842-14), the Canadian Institutes of Health Research (CIHR) (CPG-134752), and European Regional Development Fund with the Research Council of Lithuania (01.2.2.-LMT-K-718-02-0016). We thank Light Conversion for providing the laser for our experiments.

Author contributions

V.B. developed the concept of widefield P-SHG microscopy. L.K. optimized laser source for the setup. V.B., B.C.W., and K.M. designed the experiment. K.M., L.U.C., and V.B. constructed the widefield P-SHG microscope. V.S. prepared the H&E-stained breast tissue microarray slide. S.J.D. and E.Z. selected normal and tumor breast cores for P-SHG imaging. K.M., L.U.C., and A.G. calibrated the microscope and imaged the breast tissue microarray. K.M. developed the widefield P-SHG software, analyzed data, performed statistical significance testing, and carried out the classification. Y.K. developed the texture analysis software and analyzed data. R.A. performed image segmentation of H&E-stained tissue images. K.M., L.U.C., Y.K., A.G., and V.B. interpreted the results and wrote the article. All authors contributed to the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-13623-1.

Correspondence and requests for materials should be addressed to V.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.