Structural Biomarkers for Breast Cancer Determined by X-Ray Diffraction

Jonathan Friedman[a], Benjamin Blinchevsky[a], Maria Slight[a], Aika Tanaka[a], Alexander Lazarev[a], Wei Zhang[a], Byron Aram[a], Melanis Ghadimi[b], Thomas Lomis[b], Lev Mourokh*[a,c], Pavel Lazarev[a]
[a]EosDx, Inc., 1455 Adams Drive, Menlo Park, CA, USA 94025; [b]San Fernando Valley Cancer Foundation, 15211 Vanowen St Ste 208, Van Nuys, CA, USA 91405; [c]Physics Department, Queens College of the City University of New York, 65-30 Kissena Blvd, Flushing, NY, USA 11367

*lev.murokh@qc.cuny.edu; phone 1 718 997-4893; fax 1 718 997-3349

ABSTRACT

False positives from breast cancer screenings lead to billions of dollars of waste and suffering every year. X-ray diffraction of breast tissue for cancer detection is a promising technique that can potentially be used to reduce the significant number of unnecessary biopsies by first scanning suspicious areas, rather than performing a biopsy straightaway. Breast tissue diffraction patterns contain information about the structure and density of constituent fiber molecular structures, such as fatty acids and collagen. These structural biomarkers are known to change due to the presence of tumors. We ran a pilot study with biopsies from 38 cancer patients that were scanned using a low-cost diffractometer. Our diagnostic algorithm achieved an overall performance of 96.3% sensitivity, 91.6% specificity, and 93.4% positive predictive value based on a random train-test split. We believe X-ray diffraction technology is mature enough to be integrated into the clinical setting in the near future.

Keywords: X-ray diffraction, early cancer diagnostics, structural biomarkers, principal component analysis, ROC curve.

1. INTRODUCTION

Cancer is the second leading cause of death worldwide, accounting for nearly 10 million deaths in 2020[1]. By 2040, the global burden is expected to grow to 27.5 million new cancer cases and 16.3 million cancer deaths simply due to the growth and aging of the population[2]. The "War on Cancer" was announced more than 50 years ago with the signing of the National Cancer Act of 1971 by United States President Richard Nixon. Since then, billions of dollars have been spent on cancer-related research, and trillions have been spent on treatment. Still, despite certain improvements in cancer survival and the development of successful treatment methods for certain types of cancer, this war is far from victory. This is especially true for breast cancer. According to the World Health Organization (WHO), breast cancer is the most common type in women in the developed and developing world. In 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally[3].

In this work, we present a systematic analysis of human breast tissue samples using X-ray diffraction (XRD) to determine the structural biomarkers associated with the changes in extracellular matrix (ECM). Many components of ECM[4], such as glycoproteins, lipids, collagen, and keratin, exhibit a periodicity, leading to pronounced XRD patterns. These compounds experience cancer-induced changes, which the XRD measurements can monitor. In several papers, researchers studied local structural responses of breast tissues[5-12] using both small- (SAXS) and wide-angle x-ray scattering (WAXS). Numerous characteristic peaks were identified, and their relations to the periodicities of various ECM components were discussed. It should be emphasized that the developed EosDx prototype (Fig. 1) can perform both SAXS and WAXS measurements on the same machine just by changing the distance between the sample and the detector.

Figure 1. EosDx X-Ray Diffractometer.

2. METHODOLOGY

2.1 Tissue preparation

Breast tissue samples (biopsies and mastectomy) were obtained from the San Fernando Valley Cancer Foundation following regulatory standards. Specimens were transported to the XRD laboratory on dry ice on surgery day. If needed, specimens were thawed in a water bath at room temperature.

The XRD clinical team processed specimens in a sterile environment within a laminar flow hood. Tissues were microscopically examined and manually probed, guided by pathology reports, in order to identify control-like or tumor-like regions. Appropriately sized sample holders were selected based on tissue dimensions, and each was labeled with a unique barcode for tracking purposes. Mylar windows were fitted to sample holders to secure specimens. Mylar material was specifically chosen since it does not produce a diffraction signal at the angles of interest in our study.

Samples were preserved in CryoStor CS10 medium at -80°F for long-term storage. Our approach was designed to ensure the integrity and traceability of tissue samples for XRD analysis.

2.2 Diffraction measurements

We used a simple, custom-built, low-cost, transmission-mode diffractometer comprising a Xenocs source, Genix 3D Cu with Fox 12-53 Cu Mirror ($\lambda = 0.1540562$ nm), an Advacam MiniPIX TPX3 detector (256x256 pixel sensor array, with 55 μm square pixel dimensions), and a sample stage. For affordability and simplicity, we used air as the scattering medium. For signal normalization purposes, a beam stop was not used. The collimated beam had an approximate spot size of 1 mm in diameter on samples.

The diffractometer was calibrated using a sample of powdered silver behenate. Breast tissue samples were scanned for 2 or 4 minutes at various sample-to-detector distances. In this work, we present the WAXS results obtained with a distance of approximately 12 mm. Scanning time was decreased for control-like tissues since enough signal was captured after 2 minutes. Measurement data was saved as an array of 256x256 integers.

2.3 Data analysis

The analysis aimed to produce a diagnostic algorithm capable of taking a raw diffraction input image and outputting a binary cancer prediction. We used data from 38 cancer patients with 292 diffraction scans in total. We randomly selected 30 patients for training and 8 patients for testing (the “blind” set), see Table 1 for dataset compositions**.**

Raw measurements were preprocessed using the following steps: denoising, azimuthal integration, conversion of units, interpolation, normalization, and standard scaling. Analysis consisted of two steps, principal component analysis (PCA) and logistic regression. All steps are summarized below.

Table 1. Dataset compositions.

Dataset	Cancer patient count	Tumor measurement count	Control measurement count	Total measurement count
Training	30	122	92	214
Blind	8	39	39	78
Total	**38**	**161**	**131**	**292**

Denoising consisted of detecting and ignoring pixels with singular values. Azimuthal integration converted the 2D measurements into 1D intensity profiles by computing the mean intensity as a function of the radius using the beam center location from corresponding calibration measurements. Conversion of units used the sample-to-detector distance from corresponding calibration measurements to transform the pixel length units into momentum transfer units (q). Conversion of units was especially needed since multiple sample-to-detector distances were used throughout the course of this study. Similarly, interpolation was also needed to compare specific q values due to the digital nature of the detector. Normalization via the Euclidean norm was used to help control for differences in flux due to exposure time or tissue density. In subsequent steps, standard scaling was used to prevent arbitrary q values from having too much weight.

PCA was performed on the q range of 10 to 23 nm^{-1} at a resolution of 256 points. We used three PCA components so that individual tissue diffraction measurements were represented by three coordinates in PCA space. PCA reduces the dimensionality and changes the basis so that the lowest dimensions have the highest variance. We fit a logistic regression classifier to the training data using a linear decision boundary (i.e., a plane) in 3D PCA space. The clinical team provided the diagnostic labels "control" or "tumor" for model training and testing. We used sensitivity, specificity, positive predictive value (PPV), and the F1-score as performance metrics. Sensitivity is the proportion of the tumor samples that were correctly identified. Similarly, specificity measures the proportion of control samples that were correctly identified. PPV measures the accuracy of cancer predictions. Finally, the F1-score is a weighted mean of sensitivity and specificity. We chose the threshold which yielded the performance closest to an ideal classifier on the receiver operating characteristics (ROC) curve. The area under the ROC curve (AUC) is also an important metric, as the closer it is to one, the better. Performance was assessed on a blind set by a separate team using the same preprocessing and analysis pipeline from the training phase, which included all preprocessing and analysis steps with a pre-trained logistic regression classifier.

3. RESULTS

Typical XRD patterns are presented in the top panels of Fig. 2, with the control-like sample on the left and the tumor-like sample on the right. The relative positions of the rings can be compared using two bright faulty pixels. Corresponding 1D dependencies of the intensities on q, obtained after azimuthal integration and conversion of units, are shown in the middle panels. The bottom panels contain the magnified region of the q values used for the PCA analysis. Measurements of control-like tissues typically produced a diffraction ring near 14.1 per nm, whereas measurements of tumor-like tissues typically produced a diffraction ring near 20 per nm. The healthy peak is associated with the intermolecular distance in the fatty acid molecules (adipose)[13]. The tumor peak corresponds to the signal from the oxygen-oxygen electron density correlation between adjacent molecules on the tetrahedral structure of liquid water[14]. It should be noted that the most prominent peak in the control-like sample is sharper, whereas the most prominent peak in the tumor-like sample is broader.

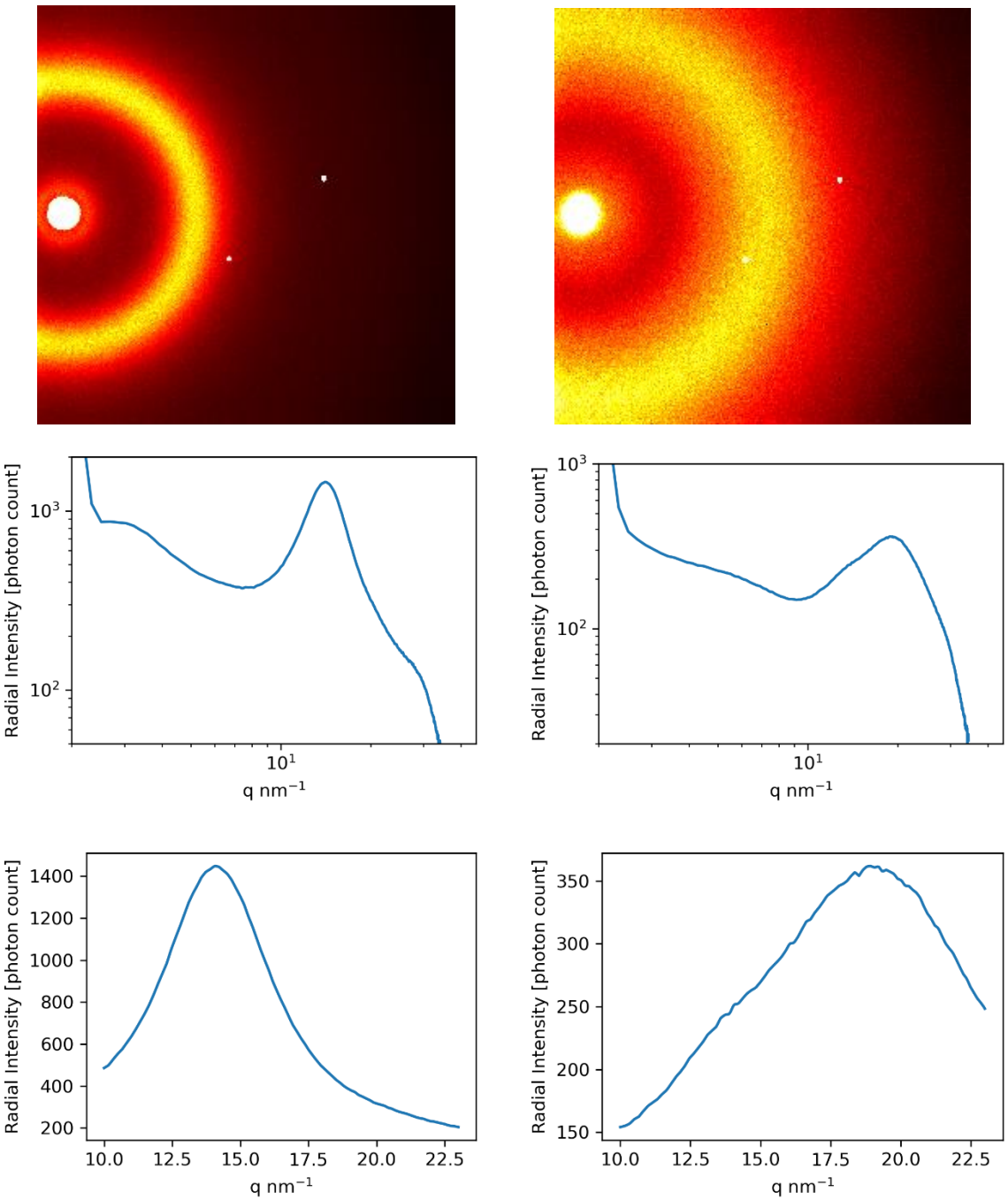

Figure 2. Typical control-like sample diffraction pattern (left) and tumor-like sample diffraction pattern (right). The top images are raw measurements with log-rescaled intensity; the middle images are after azimuthal integration; and the bottom images contain data from the *q*-range 10 to 23 nm^{-1}.

The data were run through the preprocessing and analysis pipeline, with the obtained PCA loading curves presented in Fig. 3 and the variance shown in Table 2.

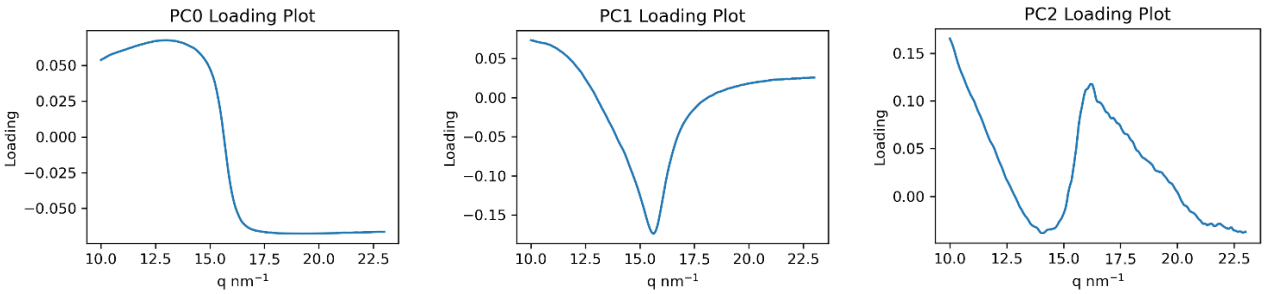

Figure 3. PCA loading plots for PC0, PC1, and PC2.

Table 2. Explained variance for the first three dimensions of PCA.

PC0	**PC1**	**PC2**	**Total**
84.2%	12.5%	2.6%	99.3%

After using these loading plots, each data set can be represented as a point in the 3D space of the principal components, as shown in Fig. 4.

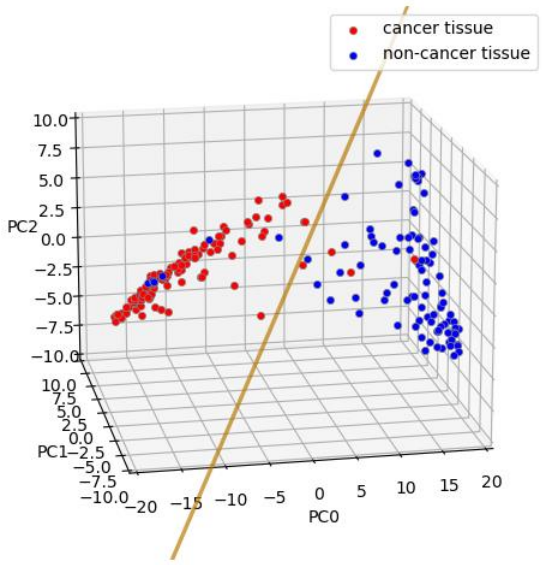

Figure 4. PCA-transformed data in 3 dimensions and the decision boundary.

It is evident from Fig. 4 that we achieve an excellent separation of control and tumor clusters. Table 3 contains the detailed performance metrics, including sensitivity, specificity, positive predictive value (PPV), and F1 score.

Table 3. Algorithm performance on the training set, blind set, and entire dataset.

Dataset	**Sensitivity**	**Specificity**	**PPV**	**F1**
Training	95.9%	93.5%	95.1%	0.955
Blind	97.4%	87.2%	88.4%	0.927
Overall	**96.3%**	**91.6%**	**93.4%**	**0.948**

Varying the decision boundary, the ROC curves are obtained for the training and blind datasets, as shown in Fig. 5. The AUC values are 0.96 for the training set and 0.97 for the blind set, indicating high model performance.

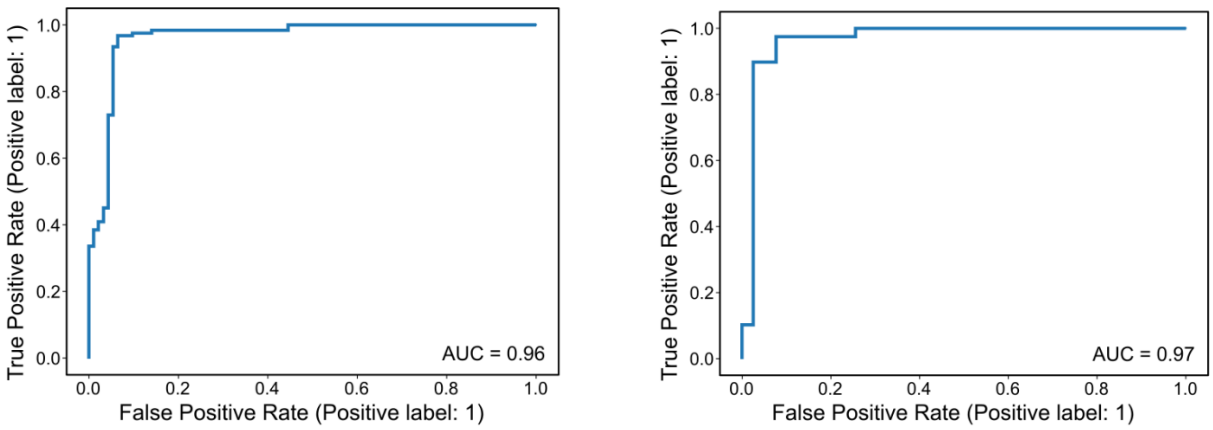

Figure 5. ROC curves on training (left) and blind (right) data, with the AUC values indicated.

4. CONCLUSIONS

In conclusion, we performed X-ray diffraction measurements of the tissue samples from 38 patients with breast cancer. The clinical team visually separated the tissues into the control and tumor ones. The XRD patterns were examined using principal component analysis and logistic regression. We achieved 95.9% sensitivity, 93.5% specificity, and 95.1% positive predictive value (PPV) for our training set. For the blind set, the results are 97.4%, 87.2%, and 88.4%, respectively. It should be emphasized that a separate team performed the blind set analysis.

In this work, we limited ourselves to the WAXS results with the *q*-range of 10 to 23 nm^{-1}, with the most prominent features associated with adipose. Next, we will examine the SAXS to study the contributions of collagen and glycoproteins.

Our results represent a promising step towards commercialization and widespread adoption of diffraction technology for early cancer detection. However, future studies are necessary to increase the data volume, match our results with "gold-standard" histopathology, and compare them with genuinely healthy samples.

REFERENCES

[1] Ferlay, J., et al., "Global Cancer Observatory: Cancer Today," Lyon: International Agency for Research on Cancer; 2020, https://gco.iarc.fr/today

[2] Ferlay, J., et al., "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," Int. J. Cancer 136, E359-386 (2015).

[3] https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[4] Frantz, C., Stewart, K. M., and Weaver, V. M., "The extracellular matrix at a glance," J. Cell Sci. 123, 4195-4200 (2010).

[5] Kidane, G., Speller, R. D.. Royle, G. J., and Hanby, A. M., "X-ray scatter signatures for normal and neoplastic breast tissues," Phys. Med. Biol. 44, 1791 (1999).

[6] Lewis, R. A., et al., "Breast cancer diagnosis using scattered X-rays," J. Synchrotron Radiat. 7, 348–352 (2000).

[7] Conceicao, A. L. C., et al., "Multivariate analysis of the scattering profiles of healthy and pathological human breast tissues," Nucl. Inst. Meth. Phys. Res. A 652, 870–873 (2011).

[8] Pani, S., et al., "Characterization of breast tissue using energy-dispersive X-ray diffraction computed tomography," Appl. Radiat. Isot. 68, 1980–1987 (2010).

[9] Sidhu, S., et al., “Classification of breast tissue using a laboratory system for small-angle x-ray scattering (SAXS),” Phys. Med. Biol. 56, 6779 (2011).
[10] Scott, R., et al., “Relationships between pathology and crystal structure in breast calcifications: an in situ X-ray diffraction study in histological sections,” NPJ Breast Cancer 2, 16029 (2016).
[11] Moss, R. M., et al., “Correlation of X-ray diffraction signatures of breast tissue and their histopathological classification,” Sci. Rep. 7, 12998 (2017).
[12] Conceicao, A. L. C., et al., “The influence of hydration on the architectural rearrangement of normal and neoplastic human breast tissues,” Heliyon 5, e01219 (2019).
[13] Tartari, A., Casnati, E., Bonifazzi, C., and Baraldi, C., “Molecular differential cross sections for x-ray coherent scattering in fat and polymethyl methacrylate,” Phys. Med. Biol. 42, 2551-2560 (1997).
[14] Kidane, G., Speller, R. D., Royle, G. J., and Hanby, A. M., “X-ray scatter signatures for normal and neoplastic breast tissues,” Phys. Med. Biol. 44, 1791-1802 (1999).