Computational and Structural Biotechnology Journal 21 (2023) 3459–3465

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Evaluation of breast tumor morphologies from African American and Caucasian patients

A. Stone [a], C. Kalahiki [b], L. Li [b], N. Hubig [b], F. Iuricich [b], H. Dunn [c,*]

[a] Dept. of Animal & Veterinary Sciences, Clemson University, Clemson, SC, USA
[b] School of Computing, Clemson University, Clemson, SC, USA
[c] Dept. of Bioengineering, Clemson University, Clemson, SC, USA

ARTICLE INFO

Article history:
Received 19 March 2023
Received in revised form 25 June 2023
Accepted 28 June 2023
Available online 1 July 2023

Keywords:
Breast cancer
Racial disparities
Artificial intelligence
Extracellular matrix

ABSTRACT

The primary aim of this research was to investigate potential differences of breast tumor morphologies across African American and Caucasian racial groups by utilizing machine learning (ML) and artificial intelligence (AI) methods. While breast cancer disparities can partially be attributed to social determinants of health, tumor biology also contributes to survival outcomes. The rate of breast tumor growth is largely dependent on the extracellular matrix (ECM). Current research suggests the cellular components of the ECM may vary among racial and ethnic populations, and this may contribute to the incidence of cancer in African Americans. We utilized a supervised AI method to evaluate morphological differences between African American and Caucasian breast cancer tumors. Images used for analysis were downloaded from the Cancer Genome Atlas (TCGA) biorepository stored in the NIH Genomic Data Commons (GDC) data portal. We designed an ML classifier using the AlexNet model provided in PyTorch's torchvision package. The model was pre-trained and adapted via transfer learning resulting in a classification accuracy of 92.1% when using our breast cancer tumor image database split into 80% of training set and 20% of testing set. We interpreted the results of the AlexNet and ResNet50 models using LIME and saliency mapping as the explainers. Based on the images from our bi-racial testing set, this study confirmed significant variations of tumor and ECM regions in the different racial groups evaluated. Based on these findings, further analysis and characterization may provide new insight into disparities associated with the incidence of breast cancer.

1. Introduction

Approximately one in eight women in the United States will develop breast cancer in their lifetime, and this disease is the second leading cause of death among women [1]. Reports of incidence and mortality vary amongst different racial groups [2] where women of African American descent are disproportionately affected by breast cancer compared to their Caucasian counterparts. Although Caucasian women have a higher incidence rate overall, African American women have a 42% higher mortality rate than the comparable rate for Caucasian women [3]. African American women have higher incidence rates than Caucasian women before the age of 40 and are most frequently diagnosed with the most aggressive subtype, triple negative breast cancer. Extrinsic factors, including environment, lifestyle and socioeconomic status when combined with intrinsic factors such as germline genetics cannot be clearly delineated as the specific cause of cancer incidence [4]. Approximately 30% of breast cancer cases are associated with modifiable risk factors including the lack of physical activity, excess body weight, and alcohol consumption indicating some breast cancers may be preventable [5]. However, there is emerging evidence of ethnic differences in breast cancer indicating research should have a multifaceted approach targeting extrinsic and intrinsic causes of racial disparities in breast cancer [4].

The extracellular matrix (ECM), also known as the stroma or tumor microenvironment, is comprised of several different cell types such as immune cells, epithelial and endothelial cells, mesenchymal stem cells, cancer associated fibroblasts, inflammatory cells and pericytes [6]. Tumor growth rate is largely dependent on interactions between the cancer cells in the tumor and the ECM.

Abbreviations: AI, Artificial intelligence; ECM, Extracellular matrix; LIME, Local Interpretable Model-agnostic Explanations; ML, Machine learning; TCGA, The Cancer Genome Atlas; WSI, Whole slide images

* Corresponding author.
E-mail address: walkerd@clemson.edu (H. Dunn).

https://doi.org/10.1016/j.csbj.2023.06.019

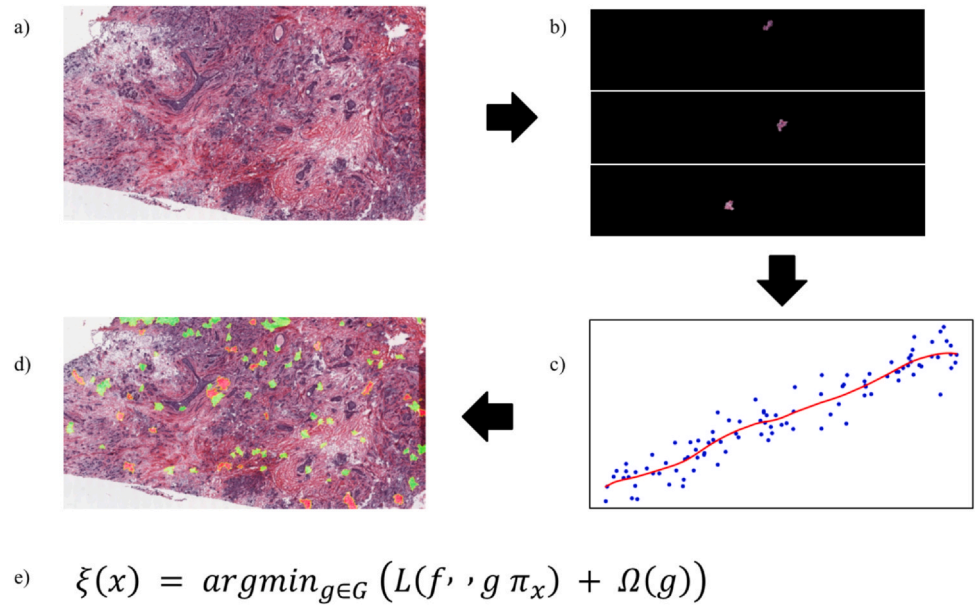

Fig. 1. LIME explainer model process from a) input H&E image, b) black box model for probabilities, c) locally weighted regression, d) explainer image with regions of interest highlighted in green, and e) equation for the LIME image explainer.

Communications between the ECM and cancer can drive tumor proliferation and cell migration leading to invasion and metastasis [7–9]. Additionally, the composition of the ECM changes significantly during breast cancer promoting metastatic spread [10], and the ECM can modify the effects of cancer treatments reducing response to chemotherapy [11]. Therefore, the ECM components are significant biological players in breast cancer development due to biochemical and mechanical cues promoting progression and metastasis [12].

Although socio-economic factors contribute to the racial disparities in breast cancer, emerging evidence suggests the ECM is partially responsible due to the different cellular components that vary among racial and ethnic populations [11, 13, 14]. The ECM in African American breast cancer populations is distinctly different from Caucasian populations due to higher vessel density, elevated cytokines and increased macrophage recruitment, resulting in poorer clinical outcomes [15]. African American patients also have higher levels of pro-inflammatory cytokines that support cell proliferation and invasion [16]. Elevated tumor associated macrophages in African American patients promote cancer cell growth and increase angiogenesis. A previous study reported microvessel density of African American tumors was higher than Caucasian patients, resulting in increased angiogenesis and macrophage recruitment [17].

Artificial intelligence (AI) and machine learning (ML) techniques are becoming more prevalent for breast cancer detection, prevention, and drug therapies. Artificial neural networks were first assessed in histopathological image analysis in 2012 to extract textural and morphological features. Convolutional neural networks are most effective for early detection resulting in more successful treatments [18]. Effective methodologies have been developed to detect other diseases, such as diabetes and leukemia [19]. Classification algorithms from hybrid natural inspired computing produced precise and accurate results. Deep learning techniques have also been applied to breast cancer with different image modalities, such as ultrasound and mammography [20]. While the human eye cannot distinguish such differences in Whole Slide Images (WSI), we involved state-of-the-art techniques for an explainable AI model to investigate such differences further.

Considering the varying cellular components in the ECM across ethnic groups, further investigation may provide insight into the origins of breast cancer disparities. Here, we propose the use of AI models to evaluate morphological patterns of breast cancer tissue. It is plausible the ECM may be one of the contributing factors to the increased incidence of breast cancer occurring in African Americans.

2. Methods

2.1. Dataset acquisition

Human cancer tissue images used for this analysis were based upon data generated by The Cancer Genome Atlas (TGCA) Research Network: https://www.cancer.gov/tcga. Whole slide breast cancer images were downloaded from the TCGA repository [21], and biospecimen files were selected for primary site, program, gender, race, and ethnicity. The date of patient recruitment and date of tissue collection prior to submission to the TCGA is unknown; however, whole slide images used for this study were downloaded in January 2021. These image files were not controlled data, and identifiable information of individual participants was not available for access. Cases selected were females who identified as white/Caucasian or black/African American with the breast as the selected primary site of cancer. To maintain consistency and avoid too many variables across samples, an additional case filter of AJCC pathologic stage IIA was selected to analyze tumors that were localized rather than including regional or distant metastases. The final data set consisted of 100 white/Caucasian and 100 black/African American whole slide images (WSI). Images were imported into HistomicsUI [22] and cropped to minimize the amount of white space in the image (Fig. 1). To reduce biases, all images were cropped to 1100 × 2011 pixels.

2.2. Deep learning analysis pipeline

The first objective of our analysis pipeline was to investigate whether a neural network could identify breast tumors originating from Caucasian and African American subjects. We utilized a supervised approach by training a neural network to identify the racial group of the subjects in our dataset, solely based on cropped images displaying the tumor and ECM of each subject. We tested several different models for the analysis pipeline using the Torchvision package from PyTorch [23] and training using the cropped image

files in our dataset. The models tested for this study included AlexNet, ResNet, Inception/GoogLeNet, and SqueezeNet.

- *AlexNet* - The AlexNet model is a convolutional neural network architecture developed by Krizhevsky et al. [24] to classify images. This model was designed to work with images containing three color channels and eleven convolutional layers over sixty-two million trainable parameters.
- *ResNet* - The ResNet model and its variants [25] achieved results superior to those of AlexNet on ImageNet data. For this current study, the focus included ResNet18 and ResNet50 variations for this work which established a baseline to explore how the increase in size between the variants of the ResNet model would impact results.
- *Inception and GoogLeNet* - The first inception model was proposed by in 2015 [26] and was presented as the new state-of-the-art in image classification, as it achieved improved results at the ImageNet ILSVRC14 challenge. The specific iteration used in the ILSVRDC14 challenge was known as GoogLeNet, which achieved superior accuracy while maintaining much smaller memory requirements.
- *SqueezeNet* - The SqueezeNet model [27] was developed to increase performance similar to AlexNet with fewer parameters. This model design was to create a more computationally efficient method to run on smaller devices.
- *ResNet50 with Attention Layers* - An attention layer was added to the ResNet50 model and metrics were collected under the same training parameters as the model without attention. The attention layer served to help the model memorize where important features occur over time.

This study utilized a pre-trained model that was derived from a large image dataset [28] due to the limited sample size. The pre-trained model utilized in these experiments was initially trained on a subset of the ImageNet data set, which does not contain any images of tissue samples or other items pertaining to this study. Models were trained using transfer learning [29] which is capable of adapting the model parameters to the new dataset, decreasing the time required to train the model to the task, and increasing accuracy.

2.3. Model interpretation using LIME and saliency maps

Local interpretable model-agnostic explanation (LIME) is an interpretation method used to explain individual prediction of black box machine learning models [30] as shown in Fig. 1. LIME generated a disconcerted dataset around the instance to explain and treat corresponding outputs of the black-box model as a label. LIME then trained an explainable model with the generated dataset. LIME was implemented into an image domain to find super-pixels or segments of significance that affect the prediction of the deep image classifier. Pixels of significance were highlighted green if they positively contributed to the image classification or red if they negatively impacted the image classification. Fig. 1a illustrates the original input image, and Fig. 1b shows three representative black box image segments from the original image which highlights selected segments with other parts of the images masked out in black. All segmented images are scored indicating the probability of the segment belonging to either African American or Caucasian racial groups. Fig. 1c is an illustration of locally weighted regressions of the segments from Fig. 1b where segments with higher weights are regarded as more important. As a result, Fig. 1d highlights positively important segments in green and negatively important segments in red. The equation in Fig. 1e is explained by $\xi(x)$ as the explanation of data instance x. The image to explain is x, as shown in Fig. 1a. The locally weighted regression model is g in Fig. 1c that was used to explain the deep learning model as f. The family of all possible explanations is g. The proximity measure of data instance x, or the locality of x is π_x. For this project, this is depicted by the image segments with other parts masked out as shown in Fig. 1b. $\Omega(g)$ is the complexity of our locally weighted regression model g. The fidelity function L measures the distance between the predictions from regression model g and deep learning model f given the locality of x, π_x. By minimizing $L + \Omega$, LIME ensures the fidelity and simplicity of locally weighted regression model g.

Saliency maps [31] are local explanation models designed at the pixel level to measure the spatial support of a particular class in each image. These maps are derived using the gradients of the model output over the model input, which identifies areas relevant for classification. Saliency maps were created for each histology image for an additional explanation of areas of significance for image classification. These maps alone were not useful in visualizing the model predictions; therefore, saliency output was split into thresholds in an attempt to identify pixels of significance at various levels of impact. A colormap of tiers of impact was created with modified opacity and overlaid onto the original image, which allowed for areas of greater significance to be highlighted and identified from the model prediction output. The original images were hematoxylin and eosin (H&E) stained, indicating tissues in shades of pink and purple hues. The saliency map tiers overlayed blue and orange/red pixels, with orange/red being the tier of highest significance, based on the identified pixels level of impact for image classification.

2.4. Analysis of model interpretations

Semantic segmentations were applied to all images for interpretation of the labeled regions identified by LIME or saliency maps. Each pixel of interest was labeled with a corresponding class to categorize the regions of interest. The data set of breast cancer tissue images was analyzed using semantic segmentations through application of the U-Net model [32]. This model was trained using the Breast Cancer Semantic Segmentation data [33] containing over 20,000 segmentation annotations of tissue regions from breast cancer images. These annotations contained regions corresponding to tissue regions of interest, including tumor, ECM, and lymphatics.

The hypothesis test statistic (t-statistic) was applied to AI models with the null hypothesis stating the African American and Caucasian cancer images were similar in the tumor and ECM regions. LIME regions were analyzed based on the areas that contributed the most to machine learning algorithms. The top five, ten, 20, 30, and 100 contributing regions were compared across both racial groups in the tumor and ECM. Saliency map results were analyzed as threshold levels referring to the intervals that provided the highest values and the most significant in defining racial groups. Pixel values above the threshold were included in computations, where the threshold value changed based on the top five, 20, or 30 segments containing significant pixels.

3. Results

3.1. Neural network model selection

The pre-trained models used in this study were provided in the torchvision package. Training operations used 100 epochs and batch sizes of 16, and following each training epoch, the accuracy of the training set was calculated. If the accuracy after an epoch was improved, the model was saved at the checkpoint. At the end of the model training, the version of the model with the best training accuracy was selected and loaded as the final model.

Results obtained from cropped images in our reference dataset with multiple neural networks following direct training are shown in Table 1. The models were pre-trained on ImageNet and adapted to the reference dataset using transfer learning. The F1 score is a

Table 1
Accuracy results obtained with different neural networks when classifying tissue images based on the racial group of the subject.

Model	Accuracy	F1 score	NMI score
AlexNet	92%	90%	0.66
ResNet18	84%	80%	0.38
ResNet50	92%	91%	0.59
GoogLeNet	90%	89%	0.53
SqueezeNet	86%	84%	0.41

harmonic mean of precision and recall. The equation below defines calculation for the F1 score where P is the precision and R is the recall.

$$\frac{2\cdot(P\cdot R)}{(P+R)}$$

The NMI score, or normalized mutual information score, is defined below where Y = class labels, C = cluster labels, H = entropy, and I(Y;C) refers to mutual information between Y and C.

$$NMI(Y, C) = \frac{2\cdot I(Y;C)}{[H(Y)+H(C)]}$$

Based on transfer learning, AlexNet was the most robust model with 92% accuracy, F1 score of 90% and the highest NMI score of 0.66. ResNet50 was the second model selected with 92% accuracy, 91% F1 score and NMI of 0.59. This information indicated that neural networks could identify differences in tumor tissues not visible by the human eye, and the features contained in the cropped images are distinguishable between Caucasian and African American subjects.

3.2. Model interpretation with LIME and semantic segmentation

LIME generated a sequence of super-pixel regions that were ranked based on their contribution to the model's output. Positive regions were those that correctly classified the racial group of the subject and were highlighted by green areas. If regions were incorrectly classified, the model highlighted the input with bright red areas. All images in our dataset contributed to positive identification, but not all images had regions of incorrect classifications. As indicated in Fig. 2, positive LIME segments were detected in the ECM, tumor tissue, and areas of overlap where the ECM and tumor intersected. Our analysis focused on characterizing positive green regions by quantifying the amount of tumor or ECM present in tissue samples across African American and Caucasian groups (Table 2). The percentage of LIME segments inside ECM regions was calculated by dividing the area of LIME segments that overlayed only with ECM regions by the total area of LIME segments. Similarly, the percentage of LIME segments inside tumor regions was calculated by dividing the area of LIME segments that overlayed only with tumor regions by the total area. Regions that contributed to the LIME model were ranked as the top five, ten, 20, 30, and 100 areas of interest to evaluate if the ECM and tumor values changed when applying additional selected segments for analysis. This data was further investigated in Table 3, indicating there was no statistical difference in the tumor areas. However, there was a statistically significant difference in the ECM for the top ten, 20, 30 and 100 regions evaluated.

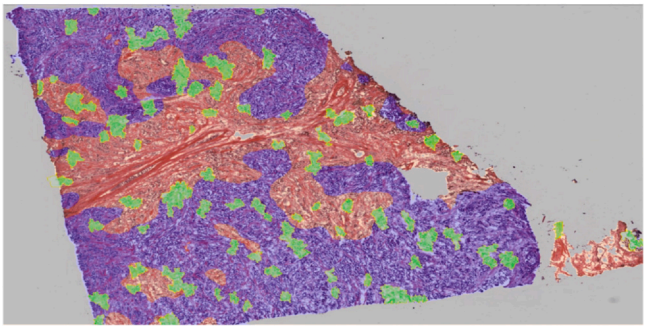

Fig. 2. Tumor areas are indicated in the lavender and dark blue regions, ECM areas are dark orange and red, and LIME segments are highlighted in green.

Table 2
LIME detected image segments from African American and Caucasian breast cancer images. Regions of the tumor and ECM were compared based on varying input regions determined by the LIME model ranked as the top 5, 10, 20, 30 and 100 areas of interest.

LIME regions	African American		Caucasian	
	Tumor	ECM	Tumor	ECM
Top 5	33.46%	27.40%	41.62%	33.80%
Top 10	34.47%	24.74%	42.15%	35.27%
Top 20	34.51%	23.50%	41.70%	33.93%
Top 30	35.67%	22.43%	41.05%	33.63%
Top 100	33.77%	20.52%	38.35%	32.36%

Table 3
T-statistic results between tumor and ECM regions across different racial groups. Regions of the tumor and ECM were compared based on varying input regions determined by the LIME model ranked as the top 5, 10, 20, 30 and 100 areas of interest.

LIME regions	Tumor		ECM	
	Statistics	p-value	Statistics	p-value
Top 5	-1.528855022	0.127888354	-1.355692603	0.176733512
Top 10	-1.512714885	0.131938581	-2.351583434	0.019671118
Top 20	-1.479693728	0.140536319	-2.439031638	0.015603469
Top 30	-1.107220064	0.269535351	-2.655137865	0.008569467
Top 100	-0.975779354	0.330364464	-2.891105852	0.004267577

3.3. Model interpretation with saliency maps and semantic segmentation

Saliency maps generated a pixel-based quantification of the importance of each pixel toward the model's outcome. This type of analysis is unique compared to the output generated by LIME as the ranking is local to each pixel, and the ranking is obtained from quantitative values. Therefore, a scalar value is defined for each pixel to quantify its importance.

We first defined different threshold levels, then categorized all pixels with a saliency value above each threshold as either tumor or ECM regions. Saliency maps for each image were saved for normalization to create binary mapping at pixel level values above or below chosen threshold values. Utilizing a segmentation model, pixels above threshold were multiplied by integers corresponding to segment labels resulting in a matrix with each pixel segment type defined. We were then able to aggregate this information and determine what segment types were most important to an image's racial group classification. Images evaluated for saliency mapping were based on biopsy tissue samples, as shown in Fig. 3a. A small region has been selected to show greater detail (Fig. 3b), highlighting tissue where pixels are mapped, and in Fig. 3c, how pixels are mapped to assigned regions of interest. Saliency map results were analyzed based on varying threshold levels set at five, ten, 20, 30, or 100 intervals that had the highest pixel value. These analyses defined which pixels were included in the computation where threshold varied based on the five different levels.

The results shown in Table 4 quantified the number of pixels in the tumor and the ECM across African American and Caucasian groups. This data was further investigated in Table 5, indicating there were no statistical differences in the tumor areas for the top five, ten and 100 thresholds of pixels, but the tumor regions were different with the top 20 and 30 threshold limits. The ECM area was

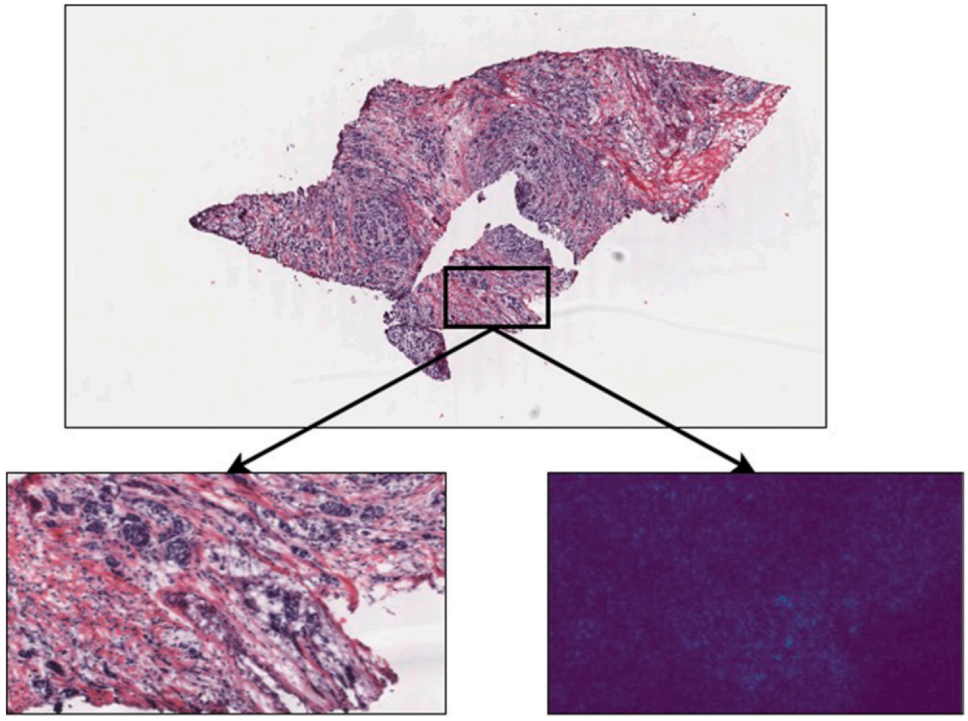

Fig. 3. Saliency maps were derived from each image by normalizing input prior to binary mapping of pixels. Varying threshold levels determined the pixels included for classification of racial groups: a) The entire tissue sample was analyzed for saliency mapping calculations; however, a smaller region has been selected by a black box to highlight for visualization, b) highlighted area used for saliency mapping where pixels were detected across different racial groups, c) representation of saliency map assigning pixels to region of interest.

Table 4
Saliency Maps detected pixels from African American and Caucasian breast cancer images. Regions of the tumor and ECM were compared based on pixel regions above threshold determined by the Saliency Maps ranked as the top 5, 10, 20, 30 and 100 most important pixel areas.

Saliency Maps	African American		Caucasian	
	Tumor	ECM	Tumor	ECM
Top 5	18.47%	26.75%	22.22%	42.48%
Top 10	18.34%	28.38%	23.79%	43.55%
Top 20	18.29%	28.99%	25.65%	47.12%
Top 30	18.95%	31.46%	29.39%	45.10%
Top 100	28.85%	23.39%	32.24%	33.18%

Table 5
T-statistic results between tumor and ECM regions across different racial groups. Regions of the tumor and ECM were compared based on varying input regions determined by Saliency Maps ranked as the top 5, 10, 20, 30 and 100 areas of pixels.

Saliency Maps	Tumor		ECM	
	Statistics	P-value	Statistics	P-value
Top 5	-0.9722766	0.33204228	-2.9253085	0.00382303
Top 10	-1.6299773	0.1046197	-3.2877874	0.00118548
Top 20	-2.6055837	0.00983453	-4.1807729	4.28E-05
Top 30	-2.8097948	0.00543031	-3.645485	0.00033721
Top 100	-1.6543695	0.09956068	-3.5349411	0.00050264

determined to be significantly different across racial groups using saliency mapping for all threshold limits reported.

4. Discussion

Racial disparities comprise one of the most distinct cancer-specific disparities [34,35]. Breast cancer disparities are most commonly associated with social determinants of health, including limited screening, increased risk factors and socioeconomic inequalities [36]. The racial disparities associated with the incidence of breast cancer is a complex issue that indicates multiple factors are potentially responsible for these differences. In addition, self-reported race or ethnicity is most often oversimplified and may not always be accurate. Emerging evidence suggests inherent biological differences in tumor microenvironment of breast cancer patients from different racial backgrounds [13]. Gene expression diversity in African American patients was observed in advanced stages of breast cancer, indicating increased genomic instability when compared to Caucasian patients [37]. The ECM involvement in breast cancer has been associated with migration, proliferation, metastasis, and reduced outcomes for breast cancer patients [38–41]. The microenvironment, also known as the ECM, has recently been addressed as a novel target for breast cancer treatment rather than focusing solely on the tumor [42]. Available literature has indicated evidence of additional factors, including the ECM that may contribute to breast cancer disparities [13,43]. Since tumor cells do not thrive in isolation, the bi-directional cross-talk between the tumor and the ECM is different in African American women due to the differential expression of cytokines and chemical messengers [13]. Gene expression analysis based on pathways related to chemotaxis and angiogenesis, and increased macrophages including tumor associated macrophages, all indicate differences in breast tumor biology of African American and Caucasian patients [43].

Advances in AI and digitization of pathology slides for diagnosis, detection and classification of tumors are becoming a promising approach. Computer scientists and pathologists have come together to apply AI techniques to improve efficiency of diagnosis. Tissue and cell level features can be identified using morphological structures which allows the computer program to discriminate between different histological components [44]. Our transfer learning model performed at a 92% level of accuracy, indicating our model could differentiate between African American and Caucasian breast cancer histology WSI. The two layers of the transfer learning model were

important for distinguishing areas that may deem morphologically different that were unable to be detected with the human eye. Not all models work with all datasets, therefore it was necessary to determine the most robust model for our collected WSI database. Different neural networks were analyzed using our dataset, and it was determined that AlexNet and ResNet50 produced the highest accuracy, F1 score and NMI score. Once the models were selected this study used two different explainers, and data was compared for the top five, ten-, 20-, 30- and 100-LIME regions or five, ten-, 20-, 30- and 100-pixel threshold values using saliency mapping. Regions of the tumor and ECM were analyzed for both African American and Caucasian patients and evaluated for differences in the tumor and ECM across racial groups.

For the explainers, the LIME program placed a green selection area over the image for local approximation, which indicated changes in tissue patterns from the two racial groups. The analyses compared the positive LIME regions, which resulted in no significant difference in the tumor regions, but statistical differences in the ECM based on the top ten, 20, 30 and 100 regions. Saliency mapping evaluated global pixel distribution across the tissue, and threshold values were determined from the top five, ten, 20, 30 and 100 most important pixel areas. There was a significant difference in the tumor tissue based on saliency mapping for the top 20, 30 and 100 thresholds, and all ECM regions were significantly different.

Based on emerging evidence, we hypothesized the ECM would be different between racial groups tested, and our results supported this hypothesis where the ECM was significant for image classification using AI and ML applications. Our model supports current literature suggesting African American and Caucasian breast cancer patients have significant differences in the ECM. This information indicates that tissue morphology variations may be one of the contributing intrinsic factors related to breast cancer health disparities for African American patients.

Regarding future work, our approach would benefit from a more articulated dataset for several reasons. First, a larger WSI dataset would allow us to test the generality of the results obtained and investigate the robustness of the model used. Second, by adding WSI from multiple racial groups, we could expand this analysis with significantly larger datasets. Such an investigation would allow us to collect additional evidence regarding the differences and similarities in ECM across racial groups. This current study was based on breast cancer tumors filtered by primary tumor location, sex of patient, race and AJCC classification due to limited resources for breast cancer tissue collected from African American patients. Future studies will benefit from the application of additional filters during the acquisition of tissue samples to include patient age at diagnosis, intake of exogenous hormones, menopause status and history of smoking for a more robust analysis once improved and expanded datasets are available.

Ethics approval and consent to participate

Human cancer tissue images used for this analysis were based upon data generated by the TCGA Research Network: https://www.cancer.goc/tcga. Approval and consent from humans to participate since all patient data in this study were downloaded from the public use TCGA database. Data processing met TCGA publication guidelines:

https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

Authors' contributions

AS assisted with study design, experimental studies, interpreted results, and manuscript preparation. CK and LL made substantial contributions to ML experimental studies, analysis of data, and manuscript preparation. NH made contributions to the study design, ML experimental studies, and manuscript preparation. FI assisted with study design, ML experimental studies, analysis of data, creation of figures, and manuscript preparation. HD assisted with study design, experimental studies, interpretation of the data, analysis, and substantial contributions in writing the manuscript. All authors read, provided edits, and approved the final manuscript.

Authors' information

Amber Stone, Department of Animal & Veterinary Sciences, Clemson University, Clemson SC, USA. Chris Kalahiki, Luyi Li, Nina Hubig, Federico Iuricich, School of Computing, Clemson University, Clemson SC, USA. Heather Dunn, Department of Bioengineering, Clemson University, Clemson SC, USA.

Declaration of Competing Interest

None.

Acknowledgements

The source of WSI from breast cancer patients used for this study were obtained from the TCGA Research Network: https://www.cancer.gov/tcga. This work was supported by the Dabo's All In Team Foundation and the Clemson University Creative Inquiry and Undergraduate Research Program.

References

[1] Stringer-Reasor EM, Elkhanany A, Khoury K, Simon MA, Newman LA. Disparities in breast cancer associated with african American identity. Am Soc Clin Oncol Educ Book 2021;41:e29–46.
[2] Danforth Jr DN. Disparities in breast cancer outcomes between Caucasian and African American women: a model for describing the relationship of biological and nonbiological factors. Breast Cancer Res 2013;15(3):208.
[3] American Cancer Society. Breast Cancer Facts & Figures 2019–2020 [Internet]. Atlanta: American Cancer Society, Inc.; 2019. Available from: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/⟨https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019–2020.pdf⟩.
[4] Yap YS. Outcomes in breast cancer—does ethnicity matter? ESMO Open 2023;8(3):101564.
[5] Islami F, Goding Sauer A, Miller KD, Siegel RL, Fedewa SA, Jacobs EJ, et al. Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States: Potentially Preventable Cancers in US. CA Cancer J Clin 2018;68(1):31–54.
[6] Santos Ramos F., Wons L., João Cavalli I, M.S.F. Ribeiro E. Epithelial-mesenchymal transition in cancer: An overview. Integr Cancer Sci Ther [Internet]. 2017 [cited 2022 Jun 22];4(3). Available from: ⟨http://www.oatext.com/epithelial-mesenchymal-transition-in-cancer-an-overview.php⟩.
[7] McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. Cell 2017;168(4):613–28.
[8] Hawkins ED, Duarte D, Akinduro O, Khorshed RA, Passaro D, Nowicka M, et al. T-cell acute leukaemia exhibits dynamic interactions with bone marrow microenvironments. Nature 2016;538(7626):518–22.
[9] Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, Merad M, et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat Med 2018;24(5):541–50.
[10] Lu P, Takai K, Weaver VM, Werb Z. Extracellular matrix degradation and remodeling in development and disease. Cold Spring Harb Perspect Biol 2011;3(12):a005058.
[11] Sun Y. Tumor microenvironment and cancer therapy resistance. Cancer Lett 2016;380(1):205–15.
[12] Oskarsson T. Extracellular matrix components in breast cancer progression and metastasis. Breast 2013;22:S66–72.
[13] Deshmukh SK, Srivastava SK, Tyagi N, Ahmad A, Singh AP, Ghadhban AAL, et al. Emerging evidence for the role of differential tumor microenvironment in breast cancer racial disparity: a closer look at the surroundings. Carcinogenesis 2017;38(8):757–65.
[14] Tang D, Zhao X, Zhang L, Wang Z, Wang C. Identification of hub genes to regulate breast cancer metastasis to brain by bioinformatics analyses. J Cell Biochem 2019;120(6):9522–31.
[15] Deshmukh SK, Srivastava SK, Tyagi N, Ahmad A, Singh AP, Ghadhban AAL, et al. Emerging evidence for the role of differential tumor microenvironment in breast cancer racial disparity: a closer look at the surroundings. Carcinogenesis 2017;38(8):757–65.

[16] Stewart PA, Luks J, Roycik MD, Sang QXA, Zhang J. Differentially expressed transcripts and dysregulated signaling pathways and networks in African American breast cancer. PloS One 2013;8(12):e82460.
[17] Martin D.N., Boersma B.J., Yi M., Reimers M., Howe T.M., Yfantis H.G., et al. Differences in the Tumor Microenvironment between African-American and European-American Breast Cancer Patients. Seoighe C., editor. PLoS ONE. 2009 Feb 19;4(2):e4531.
[18] Zhou X, Li C, Rahaman MM, Yao Y, Ai S, Sun C, et al. A comprehensive review for breast histopathology image analysis using classical and deep neural networks. IEEE Access 2020;8:90931–56.
[19] Gautam R, Kaur P, Sharma M. A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings. Prog Artif Intell 2019;8(4):401–24.
[20] Pang T, Wong JHD, Ng WL, Chan CS. Deep learning radiomics in breast cancer with different modalities: overview and future. Expert Syst Appl 2020;158:113501.
[21] The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013;45(10):1113–20.
[22] Lutnick B., Manthey D., Becker J.U., Ginley B., Moos K., Zuckerman J.E., et al. A user-friendly tool for cloud-based whole slide image segmentation, with examples from renal histopathology. bioRxiv. 2022 Jan 1;2021.08.16.456524.
[23] Paszke A. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Vancouver, Canada; 2019. p. 1–12. Available from: ⟨https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf⟩.
[24] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84–90.
[25] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. Las Vegas, NV, USA: IEEE,; 2016. p. 770–8. (Available from). ⟨http://ieeexplore.ieee.org/document/7780459/⟩.
[26] Szegedy C., Wei Liu, Yangqing Jia, Sermanet P., Reed S., Anguelov D., et al. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. Boston, MA, USA: IEEE; 2015 [cited 2022 Jun 22]. p. 1–9. Available from: ⟨http://ieeexplore.ieee.org/document/7298594/⟩.
[27] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-Lev Accuracy 50x fewer Parameters lt;0 5MB Model size 2016 [cited 2022 Jun 22]; Available from. ⟨https://arxiv.org/abs/1602.07360⟩.
[28] Deng J., Dong W., Socher R., Li L.J., Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition [Internet]. Miami, FL: IEEE; 2009 [cited 2022 Oct 5]. p. 248–55. Available from: ⟨https://ieeexplore.ieee.org/document/5206848/⟩.
[29] Torrey L, Shavlik J. Transfer Learning. In: Olivas ES, Guerrero JDM, Martinez-Sober M, Magdalena-Benedito JR, Serrano López AJ, editors. Handbook of Research on Machine Learning Applications and Trends [Internet]. IGI Global; 2010. p. 242–64. cited 2022 Oct 5 ⟨http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60566-766-9.ch011⟩. cited 2022 Oct 5.
[30] Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. San Francisco California USA: ACM; 2016 [cited 2022 Jun 23]. p. 1135–44. Available from: ⟨https://dl.acm.org/doi/10.1145/2939672.2939778⟩.
[31] Kadir T, Brady M. Saliency, Scale and Image Description. Int J Comput Vis 2001;45(2):83–105.
[32] Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 [Internet]. Cham: Springer International Publishing; 2015 [cited 2022 Aug 15]. p. 234–41. (Lecture Notes in Computer Science; vol. 9351). Available from: ⟨http://link.springer.com/10.1007/978–3-319–24574-4_28⟩.
[33] Amgad M, Elfandy H, Hussein H, Atteya LA, Elsebaie MAT, Abo Elnasr LS, et al. Structured crowdsourcing enables convolutional segmentation of histology images. In: Murphy R, editor. Bioinformatics. 35. 2019. p. 3461–7.
[34] Ellis L, Canchola AJ, Spiegel D, Ladabaum U, Haile R, Gomez SL. Racial and Ethnic Disparities in Cancer Survival: The Contribution of Tumor, Sociodemographic, Institutional, and Neighborhood Characteristics. J Clin Oncol 2018;36(1):25–33.
[35] Özdemir BC, Dotto GP. Racial Differences in Cancer Susceptibility and Survival: More Than the Color of the Skin? Trends Cancer 2017;3(3):181–97.
[36] Hirko KA, Rocque G, Reasor E, Taye A, Daly A, Cutress RI, et al. The impact of race and ethnicity in breast cancer—disparities and implications for precision oncology. BMC Med 2022;20(1):72.
[37] Stewart P.A., Luks J., Roycik M.D., Sang Q.X.A., Zhang J. Differentially Expressed Transcripts and Dysregulated Signaling Pathways and Networks in African American Breast Cancer. Cao J., editor. PLoS ONE. 2013 Dec 4;8(12):e82460.
[38] Levental KR, Yu H, Kass L, Lakins JN, Egeblad M, Erler JT, et al. Matrix crosslinking forces tumor progression by enhancing integrin signaling. Cell 2009;139(5):891–906.
[39] Winkler J, Abisoye-Ogunniyan A, Metcalf KJ, Werb Z. Concepts of extracellular matrix remodelling in tumour progression and metastasis. Nat Commun 2020;11(1):5120.
[40] Psaila B, Lyden D. The metastatic niche: adapting the foreign soil. Nat Rev Cancer 2009;9(4):285–93.
[41] Liverani C, Mercatali L, Cristofolini L, Giordano E, Minardi S, Porta GD, et al. Investigating the mechanobiology of cancer cell–ECM interaction through collagen-based 3D scaffolds. Cell Mol Bioeng 2017;10(3):223–34.
[42] Bahrami A, Hassanian SM, Khazaei M, Hasanzadeh M, Shahidsales S, Maftouh M, et al. The therapeutic potential of targeting tumor microenvironment in breast cancer: rational strategies and recent progress. J Cell Biochem 2018;119(1):111–22.
[43] Martin D.N., Boersma B.J., Yi M., Reimers M., Howe T.M., Yfantis H.G., et al. Differences in the Tumor Microenvironment between African-American and European-American Breast Cancer Patients. Seoighe C., editor. PLoS ONE. 2009 Feb 19;4(2):e4531.
[44] Ibrahim A, Gamble P, Jaroensri R, Abdelsamea MM, Mermel CH, Chen PHC, et al. Artificial intelligence in digital breast pathology: techniques and applications. Breast 2020;49:267–73.