

Can Clustering of Deep Image Representations Reveal Latent Breast Cancer Risk Subtypes in a Multiethnic Population?

Lydia Sollis, Arianna Bunnell, Thomas K. Wolfgruber¹, Brandon Quon¹, Cade Kane², Dustin Valdez¹, Brenda Y. Hernandez¹, Peter Sadowski², and John A. Shepherd¹
¹University of Hawai'i Cancer Center, ²University of Hawai'i at Mānoa



Introduction

Deep learning breast cancer (BC) risk models, such as Mirai [1], offer improved prediction over standard models but remain opaque in clinical decision-making. Clustering has been used to characterize radiomic and clinical risk subtypes [2, 3]. Here, we compare three clustering methods—k-means, Gaussian Mixture Models (GMM), and Leiden community detection [4]—to identify BC risk subtypes from features extracted at two levels of the Mirai model: (1) post-ResNet-18 [5] and (2) post-Transformer. We hypothesized that these feature representations may yield subgroups aligned with clinical attributes.

Methods

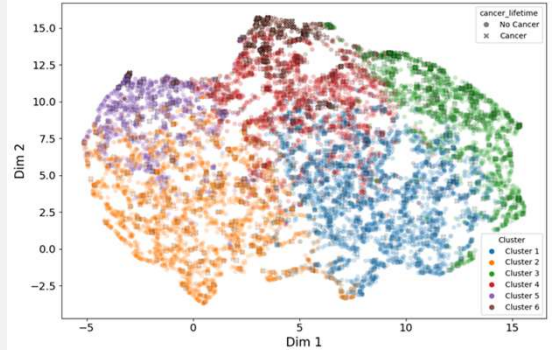
We analyzed 4,442 sequential mammograms from women aged 40–79 in the Hawai'i and Pacific Islands Mammography Registry (HIPIMR; NIH R01CA263491, U54CA143728). For each exam, we extracted 512-dimensional embeddings from (1) the ResNet-18 image encoder (post-global max pooling) and (2) the post-Transformer layer prior to attention pooling. After mean pooling across views, we applied k-means, GMM, and Leiden clustering to each feature space. Cluster structure was visualized using UMAP. Clinical differences were assessed using chi-square and Kruskal-Wallis [6] tests with Bonferroni correction ($\alpha = 0.00625$).

Table 1: Patient demographic and clinical characteristics.

	Cases	Controls
Women, N	1,283	3,159
Age at First Exam		
40–49	324	720
50–59	418	975
60–69	395	1,062
70–79	146	402
Race/Ethnicity		
NHPI	272	508
White	182	138
Chinese	99	178
Filipino	161	476
Japanese	444	174
Other Asian	62	869
Other	16	43
Breast Density		
Fatty (A)	35	189
Scattered (B)	620	1,438
Heterogeneous (C)	524	1,272
Extremely dense (D)	128	260
Menopause Status		
Menopausal	904	2,314
Pre-menopausal	379	845

Results

Figure 1: UMAP of GMM clusters from transformer-based mammogram representations. Each point is a patient exam. Colors show cluster assignments (1–6), ordered by ascending Mirai 5-year breast cancer risk. Shapes indicate cancer status (o = no cancer, x = cancer), and transparency reflects predicted risk (lighter = lower). Clusters are partially separable with mixed cancer outcomes.



A total of 1,283 cases and 3,159 matched controls met our inclusion criteria. Table 1 shows the clinical and demographic characteristics of the study population. We extracted six distinct subgroup clusters with significant differences in age, BI-RADS density, BI-RADS assessment, menopausal status, and prior incidence of breast cancer, as shown in Table 2. Cluster 6 in the post-Transformer GMM space—located at the top of Figure 1—had the highest average Mirai 5-year risk (0.083 ± 0.003) vs an overall mean of 0.023 ± 0.000 and Cluster 1 mean of 0.014 ± 0.000 . Post-screening cancer rates in Cluster 6 were $18.6\% \pm 2.1\%$, compared to $10.6\% \pm 0.5\%$ overall and $9.7\% \pm 0.9\%$ in Cluster 1. Cluster 6 included older, mostly postmenopausal women with moderate density and was not recoverable by training a Decision Tree on clinical variables, suggesting transformer-derived features captured risk-relevant imaging patterns. The most predictive features were breast density, age at mammogram, and BI-RADS assessment, consistent with their Mean Decrease in Impurity values of 0.56, 0.32, and 0.10, respectively.

Conclusion

Clustering of features from mammographic deep learning models reveal phenotypes that are statistically and clinically distinct, suggesting potential utility for personalized screening strategies. Cluster 6 from post-Transformer GMM emerged as a distinct, high-risk subgroup. These findings support the use of Transformer embeddings with probabilistic and graph-based clustering for explainable medical AI. Future work will explore end-to-end deep clustering approaches, incorporate additional clinical variables, and apply advanced explainability techniques across external cohorts to further validate and interpret subgroup differences.

References

[1] A. Yala et al., "Toward robust mammography-based models for breast cancer risk," Science Translational Medicine, vol. 13, no. 578, p. eaba4373, 2021, doi:10.1126/scitranslmed.aba4373. [2] E. Kontos, R. Bakic, S. Kontos, and A. Maidment, "Radiomic Phenotypes of Mammographic Parenchymal Complexity: Toward Augmenting Breast Density in Breast Cancer Risk Assessment," Radiology, vol. 290, no. 1, pp. 41–49, Jan 2019, doi: 10.1148/radiol.2018180551. [3] Y. Gu et al., "Unveiling Breast Cancer Risk Profiles: A Survival Clustering Analysis Empowered by an Online Web Application," Future Oncology, vol. 19, no. 40, pp. 2651–2667, Oct 2023, doi: 10.2217/fon-2023-0736. [4] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," Scientific Reports, vol. 9, article 5233, Mar 26 2019. [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770–778, 2016. [6] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," Journal of the American Statistical Association, vol. 47, no. 260, pp. 583–621, 1952, doi:10.1080/01621459.1952.10483441.

Figure 2: Variation in clinical and demographic characteristics by GMM cluster



Table 2: Mean values for clinical variables per cluster, with clusters ordered from low to high Mirai risk values

Clinical Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
age at mammogram	58.493 ± 9.04	62.975 ± 8.33	51.228 ± 8.267	61.772 ± 8.183	65.805 ± 7.637	60.018 ± 8.367
BI-RADS assessment	1.301 ± 0.591	1.381 ± 0.592	1.306 ± 0.644	1.45 ± 0.679	1.452 ± 0.688	1.436 ± 0.77
breast density	2.739 ± 0.593	1.993 ± 0.544	3.171 ± 0.663	2.685 ± 0.568	2.101 ± 0.527	2.736 ± 0.608
menopause status	0.757 ± 0.428	0.876 ± 0.329	0.483 ± 0.499	0.853 ± 0.354	0.923 ± 0.287	0.783 ± 0.413
prior BC diagnosis	0.007 ± 0.085	0.006 ± 0.076	0.011 ± 0.104	0.023 ± 0.151	0.013 ± 0.113	0.013 ± 0.115
post-screening cancer dx	0.097 ± 0.296	0.063 ± 0.243	0.155 ± 0.362	0.131 ± 0.337	0.081 ± 0.273	0.186 ± 0.389
Mirai 5-year risk assessment	0.014 ± 0.003	0.015 ± 0.004	0.02 ± 0.012	0.025 ± 0.009	0.037 ± 0.024	0.083 ± 0.042

Figure 3: Decision splits for Decision Tree trained on Mirai deep feature cluster assignments

