Outlier Detection for Mammograms

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Mammograms are vital for detecting breast cancer, the most common cancer among women in the US. However, low-quality scans and imaging artifacts can compromise their efficacy. We introduce an automated pipeline to filter low-quality mammograms from large datasets. Our initial dataset of $176,492$ mammograms contained an estimated 5.5% lower quality scans with issues like metal coil frames, wire clamps, and breast implants. Manually removing these images is tedious and error-prone. Our two-stage process first uses threshold-based 5-bin histogram filtering to eliminate undesirable images, followed by a variational autoencoder to remove remaining low-quality scans. Our method detects such scans with an F1 Score of 0.8862 and preserves $163,568$ high-quality mammograms. We provide results and tools publicly available as open-source software.

Keywords: anomaly detection, outlier detection, mammograms, unsupervised learning

1. Introduction

Breast cancer, a prevalent cause of death among women (Yusuf et al., 2021; Lei et al., 2021), can be better managed with early detection and advanced machine-learning tools (Lotter et al., 2021). For a robust machine learning classifier, one strategy is to unify the quality and content of training data by removing low-quality images and outliers (Chandola et al., 2009; Smiti, 2020; Shvetsova et al., 2021). We are building an extensive, publicly available mammography database from which we began with $967,991$ mammograms acquired by our collaborators. Through data cleaning using metadata such as small dimensions and manufacturer, we reduced the number of images to $176,492$ mammograms, but an estimated 5.5% remained low-quality. Manually selecting these images would be infeasible, prompting us to evaluate 26 unsupervised outlier detection algorithms, including traditional and deep learning-based approaches (Section 2). Based on various experiments, we introduce 5-BHIST, a thresholded histogram-binning method paired with a variational autoencoder. This two-stage outlier detection pipeline significantly outperforms other unsupervised machine learning algorithms in detecting low-quality mammograms.

2. Experimental Setup

Test Datasets. We initially created the five representative test datasets A, B, C, A*, and B* with varying proportions of unwanted images (between 5 and 24%) by randomly sampling 100 and 1000 mammograms from our original collection. We manually selected undesired images through multiple consensus-driven user studies with 9 participants. After our initial experiments, we filtered our large collection of mammograms with the best approach (5-BHIST). We then randomly sampled dataset C* for additional testing to identify the optimal algorithm for the second filtering stage (Figure 1).

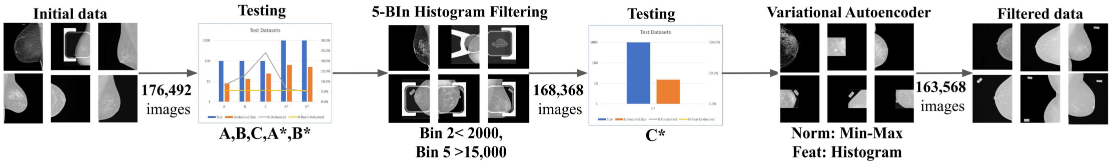

Figure 1: **Two-stage Outlier Detection.** Our method combines a 5-bin histogram filtration technique (5-BHIST) with a variational autoencoder (VAE) to automatically eliminate undesirable images. We perform experiments on a total of 6 test datasets from our initial collection of mammograms. With optimized parameters and normalization methods, we reduce the amount of low-quality mammograms by 83.15%.

Normalization. We applied various normalization methods to ensure comparable pixel intensities across different device manufacturers (Patro and Sahu, 2015). *Max:* Re-scale intensities between -1 and 1: $x_{\text{scaled}} = x/max(|x|)$. *Min-Max:* Re-scale intensities to the fixed range [0, 1]: $x_{\text{norm}} = (x_{\text{i}} - x_{\text{min}})/(x_{\text{max}} - x_{\text{min}})$. *Gaussian:* Introduce a blur: $x_{\text{gaussian}} = (x_{\text{gaussian\_filter(sigma=20)}})/x_{\text{max}}$. *zscore:* Standardize across a normal distribution: $x_{\text{gaussian}} = (x_i - \mu)/\sigma$. *Robust:* Scale data using median subtraction and IQR division: $x_{\text{robust}} = (x_i - \mu)/(IQR)$.
Image Features. We utilize image feature descriptors to reduce the number of data points per mammogram. *Full-intensity histograms*, with bin sizes selected automatically based on pixel ranges; *Downsampling*, which reduces the spatial resolution via stretching (without anti-aliasing); *Scale-invariant feature transforms (SIFT)*, used to create keypoints (Lowe, 2004); *Oriented FAST and rotated BRIEF (ORB)*, similar to *SIFT* (Rublee et al., 2011).
Algorithms. We carried out unsupervised outlier detection on all our test datasets using 26 distinct algorithms from the PyOD[1] software package (Zhao et al., 2019, 2021; Han et al., 2022), with a total number of 340 configured experiments across all tests.
Evaluation Metric. To quantify outlier detection success, we measure the F1 Score as *F1 = 2 * (precision * recall) / (precision + recall)* (Powers, 2011).

3. Results

We fully tuned the 26 anomaly detection algorithms available in PyOD for comparative analysis and evaluated the best-performing configurations on our representative test datasets (Table 1). Initial results indicated a preference for a specific normalization and feature descriptor configuration: Histogram binning after Gaussian normalization. We then performed ablation studies regarding the number of histogram bins. We compared different bin configurations ($b = 2, 5, 10$), optional Gaussian blur with varying sigma ($\sigma = 5, 10, 20$), and all normalization techniques with a 2-bin limitation based on previous explorations. Min-max normalization outperformed Gaussian, highlighting bin size as a critical factor for optimal algorithm performance. However, a 2-bin approach contributed to significant false positive classifications. Further ablation studies and consensus-driven inspection con-

[1] Python Outlier Detection (PyOD) available at https://github.com/yzhao062/pyod

Table 1: **Outlier Detection Results.** F1 Scores for the evaluated algorithms with best-performing normalization and features (G: Gaussian, M: Max, MM: Min-Max, R: Robust, Z: Z-Score, H: Histogram, S: SIFT, O: ORB) on our test datasets with varying properties of unwanted images. Our method 5-BHIST performs best overall, with an average F1 Score of 0.8772. With a variational autoencoder (VAE) as a second stage algorithm, we can further improve these results to an F1 Score of 0.8862.

Algorithm	**A (n=100,** 8%)		**B (n=100,** 13%)		**C (n=100,** 24%)		**A* (n=1000,** 6.3%)		**B* (n=1000,** 5.0%)		**C* (n=1000,** 1.5%)	
AE	M + H	0.2500	M + S	0.3077	M + S	0.4917 ± 0.0167	M + H	0.1270	M + H	0.1200	MM + H	0.1391
AvgKNN	G + H	0.6250	G + H	0.6923	G + H	0.8333	G + H	0.7460	G + H	0.6600	Z + S	0.0522
VAE	MM + H	0.2500	M + S	0.3077	MM + S	0.6000 ± 0.0333	MM + H	0.1111	MM + H	0.0940 ± 0.0092	**MM + H**	**0.1530±0.0070**
SOGAAL	M + S	0.0250 ± 0.0500	G + O	0.0000	M + H	0.0000	M + S	0.0000	M + S	0.0000	M + S	0.0124 ± 0.0247
DeepSVDD	G + H	0.6750 ± 0.0612	G + H	0.6978 ± 0.0111	G + H	0.2913 ± 0.2867	G + H	0.5322 ± 0.1676	G + H	0.4292 ± 0.0619	MM + H	0.1009 ± 0.0403
AnoGAN	G + H	0.0000	M + S	0.0769	M + O	0.2083	G + H	0.0000	G + H	0.0000	Z + S	0.1043
HBOS	G + H	0.6250	M + H	0.4615	G + H	0.8261	G + H	0.7885	G + H	0.7805	M + H	0.1217
LOF	MM + S	0.1750 ± 0.0612	MM + S	0.3077	MM + S	0.6000 ± 0.0333	MM + S	0.5095 ± 0.0321	MM + S	0.6100 ± 0.0257	M + S	0.1391
OCSVM	G + H	0.0000	G + H	0.0000	G + H	0.0000	G + H	0.0000	G + H	0.0000	G + O	0.0696
IForest	G + H	0.5000	G + H	0.6154	G + H	0.5833	G + H	0.6739 ± 0.0328	G + H	0.6473 ± 0.0101	M + H	0.1148 ± 0.0260
CBLOF	G + H	0.6250	G + H	0.6923	G + H	0.8333	G + H	0.7492 ± 0.0063	G + H	0.0202	Z + S	0.0452 ± 0.0085
COPOD	G + H	0.3750	G + H	0.3846	G + H	0.6250	G + H	0.3651	G + H	0.4583	R + H	0.1217
SOS	M + S	0.4750 ± 0.0500	M + S	0.5385 ± 0.0973	MM + S	0.7167 ± 0.0312	M + S	0.2159 ± 0.0384	M + S	0.5240 ± 0.0265	M + S	0.1217
KDE	G + H	0.0000	G + H	0.0000	G + H	0.0000	G + H	0.0000	G + H	0.0000	M + O	0.0000
Sampling	G + H	0.5750 ± 0.0612	G + H	0.5077 ± 0.0377	G + H	0.6500 ± 0.1007	G + H	0.5508 ± 0.2622	G + H	0.3341 ± 0.3221	Z + S	0.0417 ± 0.0085
PCA	G + H	0.3750	G + H	0.4800	G + H	0.5366	G + H	0.3651	G + H	0.4783	MM + H	0.1391
LMDD	G + H	0.0000	M + S	0.1692 ± 0.0897	MM + S	0.2250 ± 0.1225	G + H	0.0000	G + H	0.0000	M + O	0.1217
COF	G + H	0.6250	MM + S	0.3077	M + S	0.6250	G + H	0.1746	G + H	0.1000	M + S	0.1217
ECOD	G + H	0.5333	G + H	0.6154	G + H	0.6250	G + H	0.7097	G + H	0.6600	R + H	0.1217
KNN	G + H	0.6250	G + H	0.6400	G + H	0.8085	G + H	0.7460	G + H	0.6600	M + S	0.0522
MedKNN	G + H	0.6250	G + H	0.6923	G + H	0.8333	G + H	0.7460	G + H	0.6600	Z + S	0.0522
SOD	MM + S	0.3500 ± 0.0935	MM + S	0.4308 ± 0.0615	MM + S	0.6167 ± 0.0167	MM + S	0.2714 ± 0.0404	MM + S	0.2000 ± 0.0346	MM + S	0.0870
INNE	M + S	0.5500 ± 0.0612	MM + S	0.6308 ± 0.0308	MM + S	0.7833 ± 0.0312	M + S	0.3444 ± 0.0471	M + S	0.4280 ± 0.0431	MM + S	0.1530 ± 0.0170
FB	M + S	0.2500	G + H	0.3077	G + H	0.6250	M + S	0.4476 ± 0.0525	M + S	0.5900 ± 0.0392	MM + S	0.1496 ± 0.0085
LODA	G + H	0.3800 ± 0.1122	G + H	0.4017 ± 0.1585	G + H	0.4167	G + H	0.3312 ± 0.3241	G + H	0.5019 ± 0.3246	Z + H	0.0522 ± 0.0156
SUOD	G + H	0.5000	G + H	0.5742 ± 0.0444	G + H	0.6583 ± 0.0408	G + H	0.6926 ± 0.0104	G + H	0.6446 ± 0.0079	M + H	0.0939 ± 0.0085
5-BHIST	**G + H**	**0.8571**	**G + H**	**0.8696**	**G + H**	**0.9333**	**G + H**	**0.8908**	**G + H**	**0.8352**	N/A	N/A

firmed a setting of 5-bins with a bi-conditional thresholding operation (bins $b_2 < 2000$ and $b_5 > 15,000$) for high F1 scores. We report the performance of 5-BHIST in Table 1.

Limitations. We tuned all algorithms based on representative subsets of a large mammogram collection. While consensus-driven user studies validated the data selection steps, evaluation of normalization, feature calculation, and algorithms based on partial data can only yield estimated performance. A future release of our mammogram collection to the community will allow large-scale expert validation.

4. Conclusions

We evaluate 26 unsupervised algorithms for filtering low-quality mammograms in large data collections. Our findings indicate that a combination of min-max normalized histogram binning paired with a variational autoencoder can detect unwanted images with an average F1 Score of 0.8862. This reduces the number of unwanted images in our collection by factor 5.93x, from an estimated 9,708 low-quality scans to 1,636. Our final dataset now contains 1% unwanted images as validated by manual inspection. All code, data, and experiments are available at *url redacted*.

References

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 35:32142–32159, 2022.

Shaoyuan Lei, Rongshou Zheng, Siwei Zhang, Shaoming Wang, Ru Chen, Kexin Sun, Hongmei Zeng, Jiachen Zhou, and Wenqiang Wei. Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. *Cancer Communications*, 41(11):1183–1194, 2021.

William Lotter, Abdul Rahman Diab, Bryan Haslam, Jiye G Kim, Giorgia Grisot, Eric Wu, Kevin Wu, Jorge Onieva Onieva, Yun Boyer, Jerrold L Boxerman, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature Medicine*, 27(2):244–249, 2021.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

S Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

D. M. W. Powers. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

Nina Shvetsova, Bart Bakker, Irina Fedulova, Heinrich Schulz, and Dmitry V Dylov. Anomaly detection in medical imaging with deep perceptual autoencoders. *IEEE Access*, 9:118571–118583, 2021.

Abir Smiti. A critical overview of outlier detection methods. *Computer Science Review*, 38: 100306, 2020.

AB Yusuf, RM Dima, and SK Aina. Optimized breast cancer classification using feature selection and outliers detection. *Journal of the Nigerian Society of Physical Sciences*, pages 298–307, 2021.

Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019.

Yue Zhao, Xiyang Hu, Cheng Cheng, Cong Wang, Changlin Wan, Wen Wang, Jianing Yang, Haoping Bai, Zheng Li, Cao Xiao, et al. Suod: Accelerating large-scale unsupervised heterogeneous outlier detection. *Proceedings of Machine Learning and Systems*, 3:463–478, 2021.