Simple 2-Step Process for Removing Unwanted Images in Large Mammography Dataset
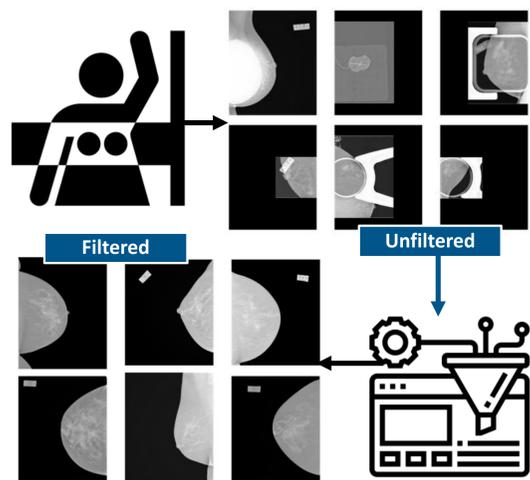
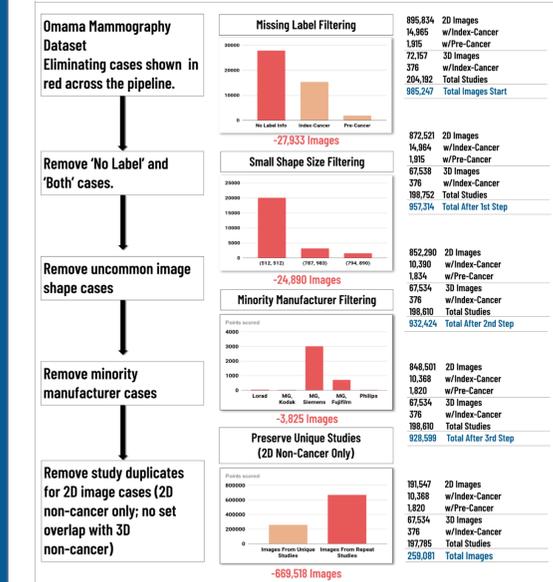Ryan Zurrin, Neha Goyal, Pablo Bendiksen, Muskaan Manocha, Daniel Haehn

#67

Abstract

We propose an automated pipeline designed to filter low-quality mammogram images..

Filtered

Unfiltered

Introduction

Omama Mammography Dataset Eliminating cases shown in red across the pipeline.

Missing Label Filtering

Remove 'No Label' and 'Both' cases.

Small Shape Size Filtering

Remove uncommon image shape cases

Minority Manufacturer Filtering

Remove minority manufacturer cases

Preserve Unique Studies (2D Non-Cancer Only)

Remove study duplicates for 2D image cases (2D non-cancer only; no set overlap with 3D non-cancer)

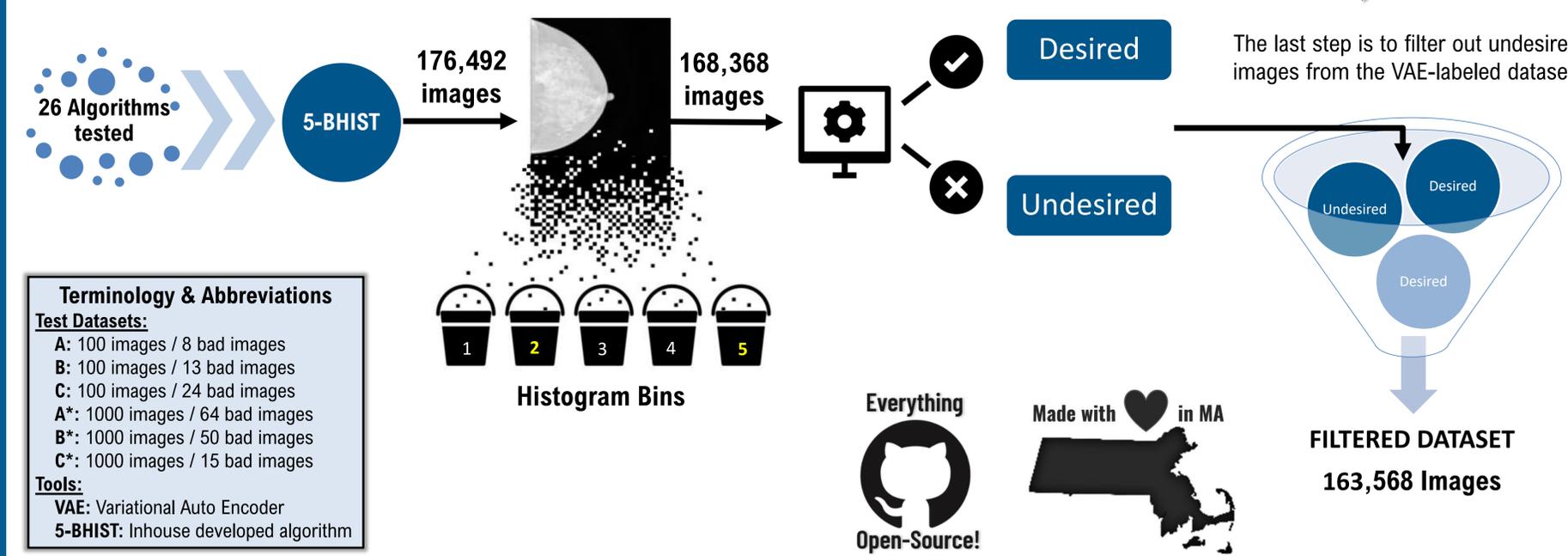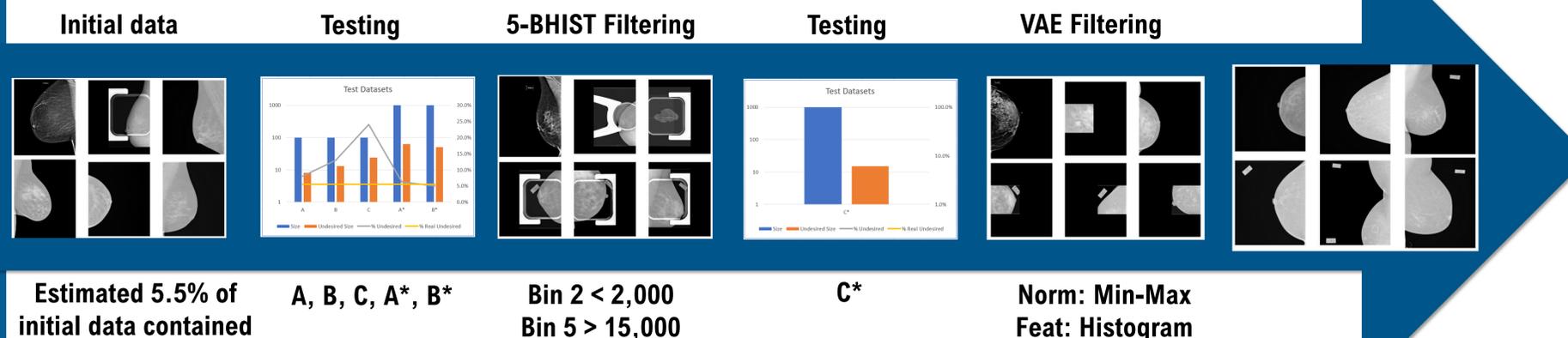## Methodology

This poster showcases a two-stage approach for removing undesirable images (outliers) from a large collection of mammograms. Stage 1 uses *threshold-based 5-bin histogram filtering* **(5-BHIST)** to eliminate unwanted images, and then for stage 2 it employs a *variational autoencoder* **(VAE)** to further refine and remove any remaining subpar scans. This innovative method was devised after rigorously testing 26 unsupervised outlier detection algorithms.

Initial data	Testing	5-BHIST Filtering	Testing	VAE Filtering

Estimated 5.5% of initial data contained unwanted images

A, B, C, A*, B*

Bin 2 < 2,000
Bin 5 > 15,000

C*

Norm: Min-Max
Feat: Histogram

26 Algorithms tested → 5-BHIST → **176,492 images** → **168,368 images** →

Desired

Undesired

The last step is to filter out undesired images from the VAE-labeled dataset.

Undesired   Desired   Desired

FILTERED DATASET
163,568 Images

Terminology & Abbreviations

Test Datasets:
- **A:** 100 images / 8 bad images
- **B:** 100 images / 13 bad images
- **C:** 100 images / 24 bad images
- **A*:** 1000 images / 64 bad images
- **B*:** 1000 images / 50 bad images
- **C*:** 1000 images / 15 bad images

Tools:
- **VAE:** Variational Auto Encoder
- **5-BHIST:** Inhouse developed algorithm

Histogram Bins
1 2 3 4 5

Everything
Open-Source!

Made with ♥ in MA

Results

F1 Score: A measure of an algorithm's accuracy, combining precision (true positives) and recall (true positives among all relevant elements). Ranging from 0 (worst) to 1 (best), it helps evaluate the effectiveness of outlier detection methods

Cleaned Dataset

- 163,568 images
- From 5.5 % to 1% unwanted images
- Total reduction by factor 5.93x
- Average F1 score of .8862
- < 6 hours to process over 176,000 images

Conclusion

Utilizing min-max normalized histogram binning with a variational autoencoder, we've reduced low-quality scans in our dataset by nearly 6x, achieving an average F1 Score of 0.8862. Consequently, only 1% of images in our final dataset are unwanted, verified by manual inspection.

Acknowledgments