Addressing Data Scarcity in Women's Health with Generative Autoencoders

Women's health research continues to face a persistent gender data gap, driven by two systemic issues: underrepresentation of women in clinical trials and the historical use of male symptoms as the diagnostic default [1]. These biases lead to missing or incomplete evidence for conditions that disproportionately affect women, making it difficult to train accurate and inclusive machine learning (ML) models. Generative AI offers a promising direction by learning from small, imbalanced datasets [2]. In particular, generative models can reduce dependence on expensive clinical trials, preserve privacy through synthetic data generation, and lower barriers to entry for ML research in women's health [3]. However, several challenges remain including the risk of bias amplification and difficulties with label fidelity in unsupervised settings.

We present a study using generative augmentation on a constrained subset of the Wisconsin Breast Cancer Dataset (N = 569). After a standard 70/30 train-test split, only 10% of the training partition (~40 samples) was used to simulate data scarcity, while retraining the full test set (171 samples) for evaluation. A variational autoencoders (VAE) was trained on these 40 samples to generate an additional 40 synthetic samples, resulting in an augmented training set of 80 samples. VAEs learn the underlying structure of the input distribution by encoding it into a compressed latent space.

We trained logistic regression classifiers under two configurations: (1) using only real data and (2) using both real and VAE-generated data. This design simulates real-world constraints common in women's health, where labeled data is limited, data collection is costly, and generative models may provide critical value. Our findings show that the real only model achieved the highest accuracy of 92% with F1 scores of 0.86 and 0.95 for benign and malignant classes respectively. The model augmented with VAE data performed comparably with 85.7% accuracy and F1 scores of 0.71 and 0.91 for benign and malignant classes respectively. These results suggest that generative augmentation can maintain diagnostic performance even in severely limited data regimes. We also evaluated a conditional variational autoencoder (CVAE), but its performance was similar to the real data only baseline. This may indicate that while conditional models offer more control, they may require more training data or stronger regularization to outperform simpler unsupervised approaches in low data regimes.

Beyond breast cancer, this generative approach can be extended to other underrepresented conditions that affect women, including osteoporosis, pregnancy related conditions, hormonal disorders and chronic pain conditions. Future work will explore diffusion models, improve conditioning on clinical labels, and benchmark performance on real-world datasets. This work contributes a reproducible, lightweight pipeline for addressing gender based data gaps in clinical ML. It supports the broader vision of equitable AI for healthcare by demonstrating how generative methods can bridge the gap between data scarcity and scalable diagnostics.

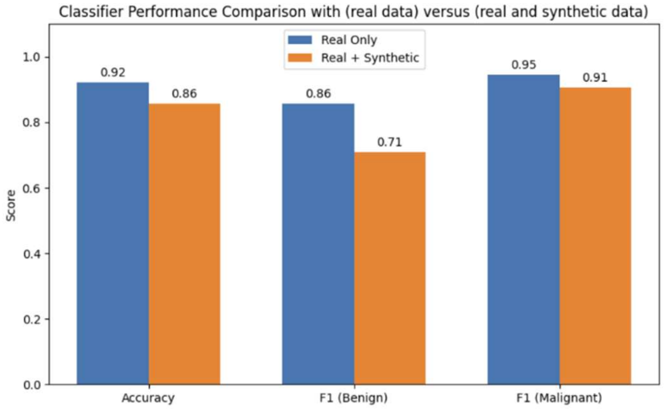

Figure 1. Model performance comparing metrics of accuracy, F1 scores for benign and malignant classes using Logistic Regression Model with real data only versus Logistic Regression Model with real and synthetic data.

[1] Lego, V. D. (2023). Uncovering the gender health data gap. *Cadernos de Saúde Pública*, *39*(7), e00065423.

[2] Ibrahim, M., Al Khalil, Y., Amirrajab, S., Sun, C., Breeuwer, M., Pluim, J., ... & Dumontier, M. (2025). Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in biology and medicine*, *189*, 109834.

[3] Rouzrokh, P., Khosravi, B., Faghani, S., Moassefi, M., Shariatnia, M. M., Rouzrokh, P., & Erickson, B. (2025). A Current Review of Generative AI in Medicine: Core Concepts, Applications, and Current Limitations. *Current Reviews in Musculoskeletal Medicine*, 1-21.