Advancing Radiogenomics in Prostate Cancer Research via new data science methods

Kevin Fee[a,b,c,d,*]

[a]Queen's University Belfast
[b]School of Electronics, Electrical Engineering and Computer Science (EEECS)
[c]Patrick G. Johnstone Center for Cancer Research (PGJCCR)
[d]Prostate Cancer Centre of Excellence (ProEx)
ORCID (Kevin Fee): https://orcid.org/0009-0001-5062-4247

Abstract. Machine learning (ML) techniques are progressively being used in biomedical research to improve diagnostic and prognostic accuracy when used in conjunction with a clinician as an AI-powered decision support system. However, many datasets used to develop these systems often suffer from severe class imbalance due to small population sizes. This leads to a ML/deep learning (DL) model to become biased to majority class samples. Current oversampling methods focus primarily on balancing datasets without adequately validating the biological relevance of synthetic minority class data, risking the clinical applicability of downstream model predictions. In addition, in multimodal scenarios, current oversampling methods do not consider cross-modal alignment of synthetic data. Consequently, these methods do not explicitly allow for cross-modal generation of missing modality information. Subsequently, due to these problems, this PhD will focus on three broad research areas: the generation of biologically feasible synthetic gene expression Data; the alignment of synthetic data for multiple modalities, and the generation of synthetic modalities where modality information may be missing for minority class samples.

1 Introduction

I am in the second year of my PhD, supervised by a diverse team of computer scientists, clinicians, and bioinformaticians. My research focuses on the complexities of biomedical datasets. These datasets are often characterised by small population sizes and class imbalances, where samples representing the disease of interest are limited compared to many samples from controlled patients[3, 11, 26, 28]. This is particularly concerning in disease diagnosis, where the minority class often represents critical conditions [2, 30]. Such imbalances undermine clinicians' trust in AI-driven models using this data [4].

Oversampling addresses this issue by generating synthetic samples of the minority class to equal the number of majority class samples [29]. This technique helps mitigate bias towards majority-class samples in ML/DL models. Biomedical research increasingly utilizes multimodal datasets, which can encompass clinical data, imaging data like MRI and CT scans, and genomics data that influences disease susceptibility and treatment response [22].

However, cross-modality imbalances necessitate oversampling to reduce bias. The quality of synthetic samples is crucial for clinical reliability but can be compromised by four data-imbalance types: (i) class, (ii) dimensionality, (iii) performance, and (iv) modality imbalance.

The following research questions will guide this work:

- (RQ1) Optimize the biological feasibility of synthetic gene expression samples to improve clinical reliability and downstream classification accuracy in imbalanced datasets.
- (RQ2) Improve cross-modality alignment of synthetic data to ensure synthetic fusion data equitably represent each modality, irrespective of dimensionality or performance imbalances, enhancing multimodal classification.
- (RQ3) Refine synthetic data generation to address missing modality information in multimodal datasets, enabling more comprehensive representations and biomarker discovery.

2 Related Work

Oversampling methods can be broadly classified into interpolation-based methods and deep generative methods (DGMs). Interpolation-based methods create synthetic samples by interpolating between minority class samples and the K-Nearest Neighbors (KNN). The Synthetic Minority Oversampling TEchnique (SMOTE) [5] generates synthetic samples, but may mislabel them by ignoring class labels. Borderline-SMOTE [13] addresses this by focusing on oversampling critical border points. ADAptive SYNthetic (ADASYN) [14] enhances oversampling by generating samples around points with greater impurity. Despite their utility, these methods can introduce biases [1] and may not completely eliminate the bias of the majority class samples [21]. A multi-method approach has been proposed to enhance dataset diversity [17] by oversampling a dataset with multiple interpolation-based methods and selecting the best synthetic samples to be included in a study.

DGMs, such as Variational Autoencoders (VAEs) [19] and Generative Adversarial Networks (GANs) [12], also generate synthetic data, but face problems such as posterior collapse in VAEs and mode collapse in GANs [25]. While Wasserstein GAN (WGAN) [10] can mitigate mode collapse, it requires large datasets with many samples. In addition, combining interpolation-based methods with DGMs may misrepresent minority classes.

Ensuring biological validity in gene expression data through differential expression analysis is essential when oversampling this

* Email: kfee04@qub.ac.uk

data[23]. Utilizing differentially expressed genes (DEGs) for oversampling supports robust modeling and preserving biological integrity. Which can be evaluated through signaling pathway [23] and gene co-expression analyses [9]. However, there is little consensus on what defines biologically feasible synthetic data [6].

Traditional data fusion methods such as early, late, and intermediate fusion are foundational for multimodal deep learning [15, 8, 27]. Recently, deep multimodal models have been used for oversampling in healthcare, such as conditional GANs that synthesize patient records [24] and radiogenomic profiles for cancer [7]. Data imbalance also arises when synthetic data misrepresents the majority class, leading to biological, dimensionality, and performance imbalances [23, 27, 7].

3 Methods

To address RQ1, we propose the Biological Evaluation Framework for Oversampling (BEFO) gene expression data, consisting of four key steps as illustrated in Figure 1: (1) Detecting Biological Patterns: We utilize Weighted Gene Co-expression Network Analysis (WGCNA) to identify gene co-expression clusters; (2) Generating Synthetic Samples: Various oversampling techniques are used to oversample the imbalanced dataset and independently generate synthetic samples; (3) Evaluating Biological Feasibility: We use a collection of random forests to assess the alignment of synthetic samples with the original clusters identified in step 1; (4) Self-Learning: We add biologically feasible samples to the training dataset and iteratively repeat steps 2-4 until the number of minority class samples equal the number of majority class samples. This procedure

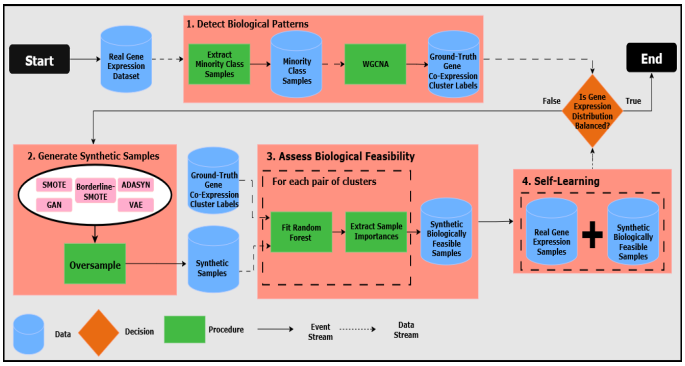

Figure 1. Proposed BEFO Methodology

rigorously evaluates each synthetic sample for biological feasibility/relevance by integrating gene co-expression analysis to retain only meaningful samples.

Gene co-expression analysis uses WGCNA to identify co-expression clusters of differentially expressed genes in minority-class samples. Co-expressed genes share expression patterns, indicating functional relationships and involved pathways [20]. WGCNA groups genes into clusters based on expression similarity, with non-significant genes forming their own seperate cluster [16].

We align synthetic samples with the WGCNA-defined ground-truth clusters using a collection of random forests. Random forests yield sample importance scores with high scores indicating fit to these clusters. Synthetic samples with scores higher than the original samples are kept, ensuring greater biological relevance.

Biologically feasible synthetic samples are then added to the original dataset for continued oversampling. This iterative process improves dataset quality by reinforcing alignment with gene co-expression patterns, preserving biology and boosting ML/DL model performance.

4 Results

To evaluate the performance of our proposed BEFO approach, we conduct our experiments on five publicly accessible, real-world datasets. In addition to this, we use a dataset (FASTMAN dataset) obtained from our affiliations at the University of Manchester [18]. The primary characteristics of each dataset can be seen in Table 1. The

Dataset	Samples	Class Imb.	Features	Clusters
Breast Cancer	78	0.38	142	12
Pancreatic Cancer	220	0.10	438	25
Prostate Cancer-Cambridge	66	0.38	377	33
Prostate Cancer-FASTMAN	184	0.11	416	22
Prostate Cancer-Michigan	88	0.44	400	24
Thyroid Cancer	167	0.49	142	9

Table 1. Characteristics of gene expression datasets. This table shows the total samples, class imbalance (minority/majority), the number of features, and the number of WGCNA-defined gene co-expression clusters.

Fowlkes-Mallows Index (FMI) assesses whether gene co-expression patterns are preserved after adding synthetic samples to an imbalanced dataset, measuring the similarity between original and oversampled WGCNA clusters [8]. The FMI ranges from 0 to 1, integrating precision and recall to evaluate clustering performance.

We ran experiments comparing average F1 scores and Area Under the Curve (AUC) for models trained on datasets oversampled using (base) state-of-the-art methods (SMOTE, Borderline-SMOTE, ADASYN, CTGAN, VAE) with and without our BEFO approach.

Results in Table 2 show that oversampling using our BEFO approach significantly enhances predictive performance.

Dataset	Base			BEFO		
	FMI	F1	AUC	FMI	F1	AUC
Breast Cancer	0.59	0.56	0.73	**0.75**	**0.61**	**0.74**
Pancreatic Cancer	0.24	0.48	0.77	**0.31**	**0.59**	**0.83**
Prostate Cancer-Cambridge	0.61	0.48	0.68	0.61	**0.54**	**0.75**
Prostate Cancer-FASTMAN	0.63	0.47	0.69	**0.69**	**0.56**	**0.77**
Prostate Cancer-Michigan	0.39	0.87	0.96	**0.47**	**0.93**	**0.97**
Thyroid Cancer	0.66	0.77	0.88	**0.83**	**0.84**	**0.91**

Table 2. Comparison of average FMI, F1, and AUC values for 5 state-of-the-art oversampling methods, with and without BEFO (Base). Bold scores indicate BEFO increases over the base methods.

5 Future Work

Currently our contribution to RQ1 is a paper titled "Towards a Biological Evaluation Framework for Oversampling (BEFO) gene expression data". This work is focused on improving the biological validity of synthetic gene expression data. Thereby improving the realism of the synthetic data and clinical reliability of downstream model diagnostic and prognostic performance. Additionally, our plan to address RQ2, is to develop a sophisticated multmodal GAN using a novel data fusion technique based on Dempster-Shafer theory, to improve synthetic data alignment between modalities. Our work on RQ3 on the generation of synthetic data for samples with missing modalities will complete a coherent and impactful thesis.

References

[1] M. M. Ahsan, S. Raman, and Z. Siddique. BSGAN: A Novel Oversampling Technique for Imbalanced Pattern Recognitions, May 2023. URL http://arxiv.org/abs/2305.09777. arXiv:2305.09777 [cs].

[2] F. Azuaje. Genomic data sampling and its effect on classification performance assessment. *BMC Bioinformatics*, 4(1):1–14, Jan. 2003. ISSN 1471-2105. URL https://link.springer.com/article/10.1186/1471-2105-4-5. ISBN: 9781471210549 Number: 1 Publisher: BioMed Central.

[3] A. J. Barton, J. Hill, A. J. Pollard, and C. J. Blohmke. Frontiers | Transcriptomics in Human Challenge Models. doi: 10.3389/fimmu.2017.01839. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2017.01839/full.

[4] A. Carriero, K. Luijken, A. de Hond, K. G. Moons, B. van Calster, and M. van Smeden. The harms of class imbalance corrections for machine learning based prediction models: a simulation study, Apr. 2024. URL http://arxiv.org/abs/2404.19494. arXiv:2404.19494 [stat].

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://arxiv.org/abs/1106.1813. arXiv:1106.1813 [cs].

[6] D. Chen, M. Oestreich, T. Afonja, R. Kerkouche, M. Becker, and M. Fritz. Towards Biologically Plausible and Private Gene Expression Data Generation, Feb. 2024. URL http://arxiv.org/abs/2402.04912. arXiv:2402.04912 [cs].

[7] L. Chen, Z. H. Huang, Y. Sun, M. Domaratzki, Q. Liu, and P. Hu. Conditional probabilistic diffusion model driven synthetic radiogenomic applications in breast cancer. *PLOS Computational Biology*, 20(10):e1012490, Oct. 2024. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1012490. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012490. Publisher: Public Library of Science.

[8] D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):1–13, Jan. 2020. ISSN 1471-2164. doi: 10.1186/s12864-019-6413-7. URL https://link.springer.com/article/10.1186/s12864-019-6413-7. Number: 1 Publisher: BioMed Central.

[9] J. H. Do and D.-K. Choi. Clustering Approaches to Identifying Gene Expression Patterns from DNA Microarray Data. *Molecules and Cells*, 25(2):279–288, Apr. 2008. ISSN 1016-8478. doi: 10.1016/S1016-8478(23)17582-0. URL https://www.sciencedirect.com/science/article/pii/S1016847823175820.

[10] J. Engelmann and S. Lessmann. Conditional Wasserstein GAN-based Oversampling of Tabular Data for Imbalanced Learning, Aug. 2020. URL http://arxiv.org/abs/2008.09202. arXiv:2008.09202 [cs].

[11] J. Fan, K. Slowikowski, and F. Zhang. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine*, 52(9):1452–1465, Sept. 2020. ISSN 2092-6413. doi: 10.1038/s12276-020-0422-0. URL https://www.nature.com/articles/s12276-020-0422-0. Publisher: Nature Publishing Group.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[13] H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In D.-S. Huang, X.-P. Zhang, and G.-B. Huang, editors, *Advances in Intelligent Computing*, Lecture Notes in Computer Science, pages 878–887, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31902-3. doi: 10.1007/11538059_91.

[14] H. He, Y. Bai, E. Garcia, and S. Li. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. pages 1322–1328, July 2008. doi: 10.1109/IJCNN.2008.4633969.

[15] I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, and A. Lerchner. Early Visual Concept Learning with Unsupervised Deep Learning, Sept. 2016. URL http://arxiv.org/abs/1606.05579. arXiv:1606.05579 [cs, q-bio, stat].

[16] J. Hou, X. Ye, W. Feng, Q. Zhang, Y. Han, Y. Liu, Y. Li, and Y. Wei. Distance correlation application to gene co-expression network analysis. *BMC Bioinformatics*, 23(1):81, Feb. 2022. ISSN 1471-2105. doi: 10.1186/s12859-022-04609-x. URL https://doi.org/10.1186/s12859-022-04609-x.

[17] A. Islam, S. B. Belhaouari, A. U. Rehman, and H. Bensmail. KNNOR: An oversampling technique for imbalanced datasets. *Applied Soft Computing*, 115:108288, Jan. 2022. ISSN 1568-4946. doi: 10.1016/j.asoc.2021.108288. URL https://www.sciencedirect.com/science/article/pii/S1568494621010942.

[18] S. Jain, C. A. Lyons, S. M. Walker, S. McQuaid, S. O. Hynes, D. M. Mitchell, B. Pang, G. E. Logan, A. M. McCavigan, D. O'Rourke, D. G. McArt, S. S. McDade, I. G. Mills, K. M. Prise, L. A. Knight, C. J. Steele, P. W. Medlow, V. Berge, B. Katz, D. A. Loblaw, D. P. Harkin, J. A. James, J. M. O'Sullivan, R. D. Kennedy, and D. J. Waugh. Validation of a Metastatic Assay using biopsies to improve risk stratification in patients with prostate cancer treated with radical radiation therapy. *Annals of Oncology*, 29(1):215–222, Jan. 2018. ISSN 0923-7534. doi: 10.1093/annonc/mdx637. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5834121/.

[19] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes, Dec. 2013. URL https://ui.adsabs.harvard.edu/abs/2013arXiv1312.6114K. Publication Title: arXiv e-prints ADS Bibcode: 2013arXiv1312.6114K.

[20] H.-J. Lee, Y. Chung, K. Y. Chung, Y.-K. Kim, J. H. Lee, Y. J. Koh, and S. H. Lee. Use of a graph neural network to the weighted gene co-expression network analysis of Korean native cattle. *Scientific Reports*, 12(1):9854, June 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-13796-9. URL https://www.nature.com/articles/s41598-022-13796-9. Publisher: Nature Publishing Group.

[21] J. Li, Q. Zhu, Q. Wu, and Z. Fan. A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Information Sciences*, 565:438–455, July 2021. ISSN 0020-0255. doi: 10.1016/j.ins.2021.03.041. URL https://www.sciencedirect.com/science/article/pii/S0020025521002863.

[22] H. Luo, J. Huang, H. Ju, T. Zhou, and W. Ding. Multimodal multi-instance evidence fusion neural networks for cancer survival prediction. *Scientific Reports*, 15(1):10470, Mar. 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-93770-3. URL https://www.nature.com/articles/s41598-025-93770-3. Publisher: Nature Publishing Group.

[23] F. Parish, W. P. Williams, G. L. Windham, and X. Shan. Differential Expression of Signaling Pathway Genes Associated With Aflatoxin Reduction Quantitative Trait Loci in Maize (Zea mays L.). *Frontiers in Microbiology*, 10:2683, Nov. 2019. ISSN 1664-302X. doi: 10.3389/fmicb.2019.02683. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6901933/.

[24] C. Sun, J. van Soest, and M. Dumontier. Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *Journal of Biomedical Informatics*, 143:104404, July 2023. ISSN 1532-0464. doi: 10.1016/j.jbi.2023.104404. URL https://www.sciencedirect.com/science/article/pii/S1532046423001259.

[25] Y. Wang, D. M. Blei, and J. P. Cunningham. Posterior Collapse and Latent Variable Non-identifiability, Jan. 2023. URL http://arxiv.org/abs/2301.00537. arXiv:2301.00537 [cs, stat].

[26] R. Yamada, D. Okada, J. Wang, T. Basak, and S. Koyama. Interpretation of omics data analyses. *Journal of Human Genetics*, 66(1):93–102, Jan. 2021. ISSN 1435-232X. doi: 10.1038/s10038-020-0763-5. URL https://www.nature.com/articles/s10038-020-0763-5. Number: 1 Publisher: Nature Publishing Group.

[27] R. Yan, F. Ren, X. Rao, B. Shi, T. Xiang, L. Zhang, Y. Liu, J. Liang, C. Zheng, and F. Zhang. Integration of Multimodal Data for Breast Cancer Classification Using a Hybrid Deep Learning Method. July 2019. doi: 10.1007/978-3-030-26763-6_44. URL https://link.springer.com/chapter/10.1007/978-3-030-26763-6_44.

[28] Y. Zhang, S. Shen, X. Li, S. Wang, Z. Xiao, J. Cheng, and R. Li. A multiclass extreme gradient boosting model for evaluation of transcriptomic biomarkers in Alzheimer's disease prediction. *Neuroscience Letters*, 821:137609, Jan. 2024. ISSN 0304-3940. doi: 10.1016/j.neulet.2023.137609. URL https://www.sciencedirect.com/science/article/pii/S0304394023005682.

[29] Y. Zhao, Z. S.-Y. Wong, and K. L. Tsui. A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection. *Journal of Healthcare Engineering*, 2018:6275435, May 2018. ISSN 2040-2295. doi: 10.1155/2018/6275435. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5987310/.

[30] T. Zhu, Y. Lin, and Y. Liu. Improving interpolation-based oversampling for imbalanced data learning. *Knowledge-Based Systems*, 187:104826, Jan. 2020. ISSN 0950-7051. doi: 10.1016/j.knosys.2019.06.034. URL https://www.sciencedirect.com/science/article/pii/S0950705119303016.