Handling Uncertainty in Health Data using Generative Algorithms

Mahdi Arab Loodaricheh[1], **Neh Majmudar**[1], **Anita Raja**[1,2] **Ansaf Salleb-Aouissi**[3]

[1]City University of New York, Graduate Center
[2]City University of New York, Hunter College
[3]Columbia University
marabloodaricheh@gradcenter.cuny.edu, neh.majmudar@gradcenter.cuny.edu, anita.raja@hunter.cuny.edu,
ansafsalleb@columbia.edu

Abstract

Understanding and managing uncertainty is crucial in machine learning, especially in high-stakes domains like healthcare, where class imbalance can impact predictions. This paper introduces RIGA, a novel pipeline that mitigates class imbalance using generative AI. By converting tabular healthcare data into images, RIGA leverages models like cGAN, VQ-VAE, and VQGAN to generate balanced samples, improving classification performance. These representations are processed by CNNs and later transformed back into tabular format for seamless integration. This approach enhances traditional classifiers like XGBoost, improves Bayesian structure learning, and strengthens ML model robustness by generating realistic synthetic data for underrepresented classes.

Introduction

Decision-making in complex clinical settings often involves high-stakes choices regarding patient care under conditions of incomplete information. Numerous clinical decision support (CDS) models have been proposed to assist clinicians in navigating these scenarios. Bayesian Networks (BNs) (Koller and Friedman 2009) represent a particularly promising approach, as they meet essential requirements for CDS in complex, real-world datasets: (1) uncovering novel patterns and (2) enhancing explainability. The capability of BNs to visually represent probabilistic dependencies between features makes them especially beneficial for clinicians who need to understand intricate feature interactions when complete relationships are not well established. BNs learned in an unsupervised manner provide one of the few modeling approaches capable of simultaneously addressing these dual objectives.

Furthermore, class imbalance is a pervasive challenge in health datasets, where an uneven distribution of outcome classes complicates the application of machine learning techniques. For instance, we seek to determine the risk of adverse pregnancy outcomes, specifically preterm birth (PTB) (Purisch and Gyamfi-Bannerman 2017) among nulliparous women (i.e., those with no prior births). PTB, defined as delivery before 37 weeks of gestation, represents a minority outcome, affecting approximately 10% of pregnancies in the Nulliparous Pregnancy Outcomes Study (nuMoM2b) dataset (Haas et al. 2015), which comprises observational data from over 10,000 pregnancies, encompassing more than 4,600 features tracked across four prenatal visits. The inherent class imbalance necessitates sophisticated modeling approaches to accurately identify risk factors and enhance prediction efficacy. Although a prior history of PTB is a predominant risk factor for subsequent preterm deliveries, a substantial number of PTBs occur in nulliparous women with no previous childbirth history (Koullali et al. 2016), and many occur without discernible symptoms or established clinical predictors (Iams et al. 2001; Huang et al. 2018; Manuck 2017).

This paper introduces RIGA (Robustness using Imbalance-Resilient Generative Augmentation), a four-phase approach to mitigate class imbalance via data augmentation (Mumuni and Mumuni 2022; Maharana, Mondal, and Nemade 2022; Mikołajczyk and Grochowski 2018; Nanni et al. 2021). Generative modeling, particularly Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), is a notable method for creating synthetic data that resembles the original dataset, proving highly effective in data augmentation tasks. GANs excel with image data (Aggarwal 2018), thus transforming the data into image representations (Sharma et al. 2019; Zhu et al. 2021) before augmentation is advantageous.

The core idea of this work is to transform non-image data into image-like formats for input into a Convolutional Neural Network (CNN). Methods like similarity measurement and dimensionality reduction (e.g., t-SNE, kPCA) help create a 2D representation capturing feature relationships. The nuMoM2b dataset exhibits a complex hierarchical structure, with both temporal and non-temporal feature dependencies. By converting such tabular data into images, generative models can leverage hidden structures, leading to more realistic and useful samples.

The **key contributions** of this work include using a data-to-image conversion method to transform rows of the nuMoM2b dataset into images, providing researchers with a novel perspective on data structure. Next, generative models are employed to augment these transformed medical datasets. Following augmentation, a lossless inverse transformation restores the data to its original form, enabling

Copyright © 2025, GenAI4Health Workshop @ Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

the use of traditional classification algorithms. Finally, a Bayesian Network pipeline is applied to evaluate the augmentation's impact on determining feature relationships within the nuMoM2b dataset.

Related Work

Augmentation of Tabular Data: Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2002) is a widely-used data augmentation method for tabular data, generating new minority class instances using a nearest neighbor approach. However, SMOTE struggles with complex, heterogeneous datasets like nuMoM2b, where understanding the underlying structure of the minority class is crucial. Adaptive Synthetic (ADASYN) (He et al. 2008), which generates synthetic samples based on the difficulty of classification, also faces similar limitations in such contexts.

Deep neural networks have been used to augment tabular data, but their popularity is limited due to challenges with mixed numerical and categorical features. Conditional Tabular GAN (CTGAN) (Xu et al. 2019) adapts to tabular data by normalizing mixed features and has been found to outperform other GAN-based methods. However, our approach outperforms CTGAN, suggesting that converting tabular data to images can produce better synthetic data.

Generative models for Images: GANs (Goodfellow et al. 2014) are deep neural architectures successful in generating synthetic data, particularly images. GANs consist of a generator and a discriminator trained in tandem—the generator creates novel data while the discriminator distinguishes real from synthetic instances. This competition improves both networks over time. GANs are widely used for augmenting image datasets (Xu et al. 2023). In our work, we utilized three distinct generative models to compare their performance and evaluate their effectiveness across different underlying data structures. Conditional GAN (cGAN) (Mirza and Osindero 2014) variant effectively handles data imbalances.

VQVAE (Vector Quantized Variational Autoencoder) (Van Den Oord and Vinyals 2017) learns discrete latent representations for high-quality reconstructions, reducing input dimensionality. The VQ process involves finding the nearest codebook entry to a given pixel or region. The quantization is represented as $q(z) = \arg\min_{c \in C} \|z - c\|^2$, where C is the codebook. VQVAE reduces blurry outputs compared to traditional VAEs, but generated quality depends on codebook size. VQGAN (Esser, Rombach, and Ommer 2021) combines VQVAE with GANs to improve training efficiency and image quality. However, it requires careful tuning due to challenges like mode collapse.

Tabular to image transformation: *DeepInsight* (Sharma et al. 2019) converts non-image data into images by recovering local structure with a similarity-based method like KNN and mapping features to pixels based on similarity. Pixel locations correspond to features, and pixel intensities represent feature values, allowing CNN application to non-image data. A limitation is that many pixels do not represent features (Zhu et al. 2021), but our approach addresses this by using an appropriate image size with one pixel per feature, making the transformation invertible. We utilize this to augment data in image form and convert it back to tabular format for traditional classification. (Andresini et al. 2021) introduced MAGNETO, adapting *DeepInsight* and GANs for intrusion detection in traffic data using a CNN, but did not explore inverse transformation to tabular format for use with traditional models, which could have allowed for a direct comparison.

Approach: The RIGA pipeline

Phase 1: Data Transformation

During the initial stage, we apply the *DeepInsight* method (Sharma et al. 2019) to convert our tabular data into images. This transformation allows us to subsequently leverage generative models. Each row (feature vector) within our dataset undergoes a conversion process, resulting in images of dimensions 28x28. The feature value is mapped to the feature's location in the image which in turn is determined by its similarity. The data matrix, initially consisting of n rows and d features, is transposed to have d rows and n columns. Subsequently, this transposed matrix undergoes dimensionality reduction techniques such as t-SNE, pinpointing each feature's location in 2D space to derive a 2D plane. As a result, every row in our dataset is transformed into an image. Using this method, along with ensuring the correct image size (28x28 in our case), guarantees that each feature is assigned a specific location in the 2D space, allowing for a fully invertible transformation without any loss of information.

Phase 2: Augmentation

In the second phase, we implement three different generative models to generate additional data for the underrepresented class in our dataset. This approach effectively expands the dataset by introducing synthetic data points, helping to address class imbalance. During this phase, we conditionally train images using these models, based on the distinctions between the two classes in our dataset. Once trained, we use the generator components of each model to produce synthetic images for the class with fewer samples. The augmentation phase workflow in our method is illustrated in Figure 1.

cGAN conditions the generator and discriminator on binary labels to create images matching the labels and generates new images after training. **VQ-VAE** uses an encoder-decoder architecture to learn discrete latent representations, enabling the decoder to generate synthetic images for the minority class by sampling from the latent space. **VQ-GAN** combines an encoder-decoder setup with adversarial training, compressing images into discrete latent codes and reconstructing them to produce high-quality synthetic images for the minority class.

Phase 3: Classification

We leverage two distinct classification strategies: One is *Direct Image Classification* which is a straightforward technique that employs standard image classification techniques on both generated and real images to make predictions.

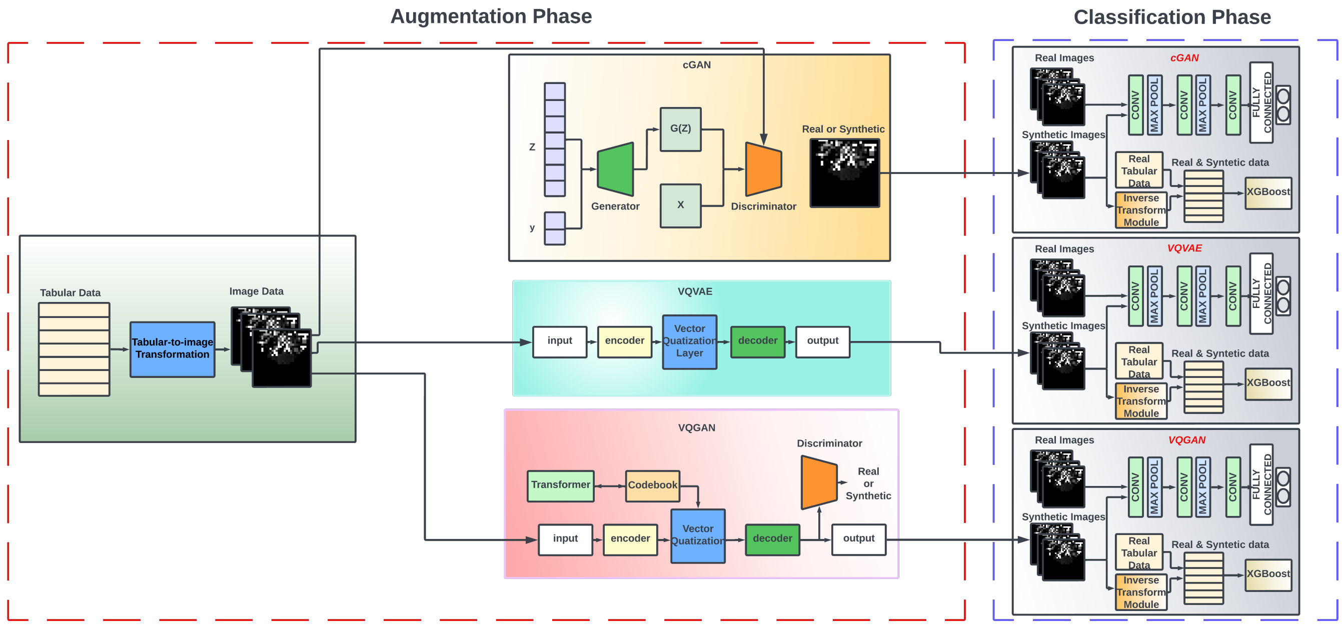

Figure 1: RIGA Pipeline of Augmentation and Classification Method: Left - Tabular-to-Image Conversion with cGAN, VQVAE, and VQGAN Training. (X: Data sample, y: Condition , and Z: Latent Noise Vector) Right - Dual Classification Paths: Top - Real and Generated Images Classified via CNN; Bottom - Synthetic Images Converted Back to Tabular and Classified with XGBoost.

The second is *Leveraging Tabular Space* which can exploit *DeepInsight*'s inverse transformation function to "decompose" augmented images back into tabular format. This allows application of powerful traditional machine learning algorithms on the familiar tabular data, even if they cannot handle image inputs directly, thereby unlocking the potential of superior performance from these algorithms.

In the first strategy, we employ CNNs to perform image classification. All real and synthetic images are input into the CNN model, and it undergoes training to make accurate classifications. In the second approach, we start by changing the synthetic images back into tabular data. Then, we use machine learning methods like XGBoost to analyze and classify this tabular data. The classification phase workflow in our approach is illustrated in Figure 1.

We convert tabular data into images where each feature has a dedicated pixel using the inverse transformation function in the *DeepInsight* method taking advantage of the following: 1. *Pixel-Perfect Mapping:* With an image size exceeding the number of features, each feature directly aligns with a unique pixel, preventing any aggregation or information loss. 2. *Lossless Recovery:* DeepInsight's inverse function helps us retrieve the original value of each feature using the pixel value it corresponds to after transformation. *3.Faithful Reconstruction:* In the final step, this allows for a perfect reconstruction of the original data row, preserving all information without compromise.

Phase 4: Bayesian Network Learning

The fourth phase builds on the Unsupervised Learning under Uncertainty (U2) Pipeline, based on the methodology in (Mallia 2023). In real-world datasets like nuMoM2b, uncertainty arises from noisy data and class imbalance, obscuring meaningful patterns. High-quality synthetic data improves data-driven discovery, including feature correlation. The U2 pipeline emphasizes (a) Uncertainty: noisy data and imbalance contribute to uncertainty, and (b) Unsupervised learning: models adjust to target outcomes with minimal expert input. The pipeline includes data transformation, augmentation, classification, discretization, visualization, structure learning, and analysis. Structure learning uses the Bayesian Information Criterion (BIC), calculated as $BIC = \log P(D|G) + \frac{d}{2}\log(N)$, to construct graphical models representing data distribution. Here, D is the data, G is the structure, d is the number of free parameters, and N is the dataset size. Tabu search addresses local optima, enhancing the likelihood of finding a global optimum. The learned graphical model maps the total distribution, supporting prediction and analysis.

Experimental Setup

Our proposed pipeline, *RIGA*, tackles the problem of class imbalance in tabular datasets to enhance Bayesian structure learning and improve prediction results. In this and the following section, we empirically evaluate the predictive performance of *RIGA* as compared to state-of-the-art approaches in the context of both simulated and real datasets and discuss the strengths and limitations of our approach under various conditions. This section describes the experimental setup while the next section describes the experiments and the results.

Datasets

The first dataset, nuMoM2b (Haas et al. 2015), is a real-world medical dataset that initially sparked our interest due to its class imbalance. The second, Madelon (Guyon et al. 2006), is a synthetic dataset modified to introduce imbalance for testing in non-real-world scenarios. The third, Myocardial Infarction (Golovenkin and Voino-Yasenetsky 2020), is a real-world dataset with significant class imbalance. The fourth, DARWIN (Cilia et al. 2022), is designed for early detection of Alzheimer's disease.

nuMom2b The Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be (nuMoM2b) dataset, introduced in (Haas et al. 2015), includes over 10,000 singleton pregnancies of nulliparous mothers. Each pregnancy has details from up to four visits: V1-V3 represent each trimester, while V4 captures delivery. The primary focus was to better understand adverse pregnancy outcomes (APOs), particularly PTB, defined as delivery before 37 weeks and affecting around 10% of pregnancies. Predicting PTB is challenging due to its imbalanced nature and lack of prior birth indicators for nulliparous mothers. The study gathered over 4600 features per pregnancy.

We selected PTB versus Full Term Birth (FTB) as class labels. The dataset includes 7,923 FTB and 780 PTB cases, reflecting an imbalance. After preprocessing, as outlined by (Goretsky et al. 2021), and incorporating Layer 2, Visit 3 features (360 features), each patient is represented as an image for further analysis.

Madelon Madelon (Guyon et al. 2006) is a synthetic dataset created for the NIPS 2003 feature selection challenge. It involves a binary classification task with continuous input features and a multivariate, highly non-linear nature. The dataset contains samples with 500 attributes. Initially balanced, it was made imbalanced in the experiment by randomly removing samples, resulting in a 10% class one and 90% class zero distribution.

Myocardial Infarction complications The Myocardial Infarction Complications dataset (Golovenkin and Voino-Yasenetsky 2020) contains 1700 records from Krasnoyarsk Interdistrict Clinical Hospital, Russia, released in August 2020 via the UCI Machine Learning Repository. It includes 124 attributes, with 111 detailing patient demographics, medical history, hospital complications, ECG results, and clinical interventions, while 12 describe complications across four time stages. This study focuses on Myocardial Rupture, selected for its class imbalance.

DARWIN The DARWIN dataset (Cilia et al. 2022) comprises handwriting data from 174 participants (89 with Alzheimer's and 85 healthy) collected via a graphic tablet across 25 tasks. It captures motor skill impairments indicative of neurodegenerative diseases. With 450 attributes per sample, the dataset aids in developing machine learning models for Alzheimer's diagnosis. Initially balanced, it was deliberately imbalanced in the first phase of the experiment.

Method

When evaluating our *RIGA* approach, we use separate training folds for image transformation and generative models, with all experiments conducted using 5-fold cross-validation from data transformation to final classification.

Data Transformation Employing the DeepInsight method for transforming tabular data into image data involved specifying the image size as 28x28 to encompass all features and facilitate subsequent use with generative models. Additionally, we utilized t-SNE as a dimensionality reduction technique for this transformation.The process began with normalizing the data, followed by feeding it into the method to obtain 28x28 images from the original tabular data.

Augmentation During the training phase, the training dataset is utilized for both image transformation and generative models training.

The cGAN is trained for 50 epochs, with the generator taking a 100-dimensional random noise vector and labels as input. During evaluation, it generates synthetic data, which is added to the training set to improve CNN or other machine learning models. The VQ-VAE model, also trained for 50 epochs, uses an encoder to quantize images with 128 embeddings and a decoder to reconstruct them. After training, PixelCNN generates priors that are decoded into synthetic images, which are added to the training set for enhancing CNNs or other models. Similarly, the VQ-GAN model, trained for 50 epochs, employs an encoder with 128 embeddings, a transformer for long-range dependencies, and a decoder refined by a discriminator. Synthetic data is generated from random latent vectors during evaluation and incorporated into the training dataset for further analysis.

Figure 4 in the appendix describes snapshots of real and synthetic images generated by the generative models trained for the four transformed datasets.

Classification In the classification phase, we initially generated a set of images from the underrepresented class to create a balanced dataset. At this phase, we had two alternatives: (a) Employing a CNN with the entire image dataset; (b) Employing the inverse transform to revert the data to tabular form and utilizing traditional machine learning techniques. For conventional machine learning, XGBoost was employed.

Through initial experiments with augmented images, we selected the CNN architecture by performing hyperparameter tuning using grid search. This involved optimizing batch size, the number of layers, and the number of neurons to identify the best configuration.

We use AUC (Area Under the Curve) as the performance metric because it is robust to class imbalance and independent of threshold selection. Unlike accuracy, AUC remains reliable regardless of class ratios and evaluates model performance across all thresholds, providing a comprehensive view of discrimination power. This made AUC ideal for assessing our models on imbalanced data.

Bayesian Network Learning The experiments conducted on the U2 pipeline focused on comparing the scoring function changes. Initially, we analyzed a dataset with imbalanced class distributions. The U2 process on the original dataset produced a BIC score reflecting the model's performance. Using the RIGA pipeline, the augmented dataset achieved a more balanced distribution, enabling improved evaluation. The scoring function change was analyzed alongside the visualization of the learned Bayesian Network using the Markov Blanket. Once trained, the Markov Blanket aids in predicting relevant information for any variable. As class balance improves, the representation of underlying data relationships becomes clearer, enhancing the reliability of the Bayesian Network.

Empirical Results

In this section, we describe the experiments conducted on each dataset to evaluate the pipeline's effectiveness. A 5-fold cross-validation was applied from image transformation and generative model training to image generation and classification.

nuMoM2b Results

In the nuMoM2b dataset, binary classification distinguishes PTB from FTB cases. ADASYN and SMOTE decreased XGBoost AUC, while CTGAN provided minimal improvement. For CNN classification, RIGA with cGAN, VQVAE, and VQGAN augmentation improved performance by 0.5%, 2.23%, and 2.82%, respectively. After converting to tabular data, RIGA with cGAN, VQVAE, and VQGAN enhanced XGBoost results by 1.82%, 1.88%, and 2.14%, respectively.

The performance difference observed in using VQGAN with XGBoost over other methods, such as VQVAE and cGAN, can be attributed to the nature of the numom2b dataset and the task at hand. With 360 features and an imbalanced binary classification problem (7,923 FTB vs. 780 PTB cases), capturing the nuanced patterns of the minority class (PTB) is critical. VQGAN's ability to leverage VQVAE's discrete latent space modeling enhances the representation of complex features, allowing it to generate synthetic samples that better reflect the distribution and variability of PTB cases. This improves the classifier's ability to distinguish minority class instances. In contrast, VQVAE alone, while effective at data representation, lacks the adversarial training that refines the synthetic outputs for higher fidelity, which explains its slightly lower performance. Meanwhile, cGAN, which depends on a continuous latent space, might struggle to model the discrete and intricate feature distributions inherent in the data, leading to relatively poorer representation and classification performance. These insights highlight the importance of aligning data generation methods with the structural properties of the dataset to improve model performance, offering generalizable lessons for other imbalanced datasets.

Madelon Results

In the Madelon dataset, a binary classification task with 2600 samples and 500 features, imbalance was induced by removing 1150 class 1 samples. Table 1 summarizes the results. ADASYN, SMOTE, and CTGAN reduced AUC scores. RIGA with cGAN, VQVAE, and VQGAN improved CNN AUC by 0.9%, 2.62%, and 4.4%, respectively. For XGBoost with tabular transformations, RIGA with cGAN, VQVAE, and VQGAN enhanced AUC by 1.63%, 3.28%, and 6.17%, respectively. VQGAN outperformed others by combining vector quantization with adversarial training for diverse, high-quality synthetic data. VQVAE, lacking adversarial training, ranked second, while cGAN, lacking vector quantization, was less effective.

Myocardial Infarction Results

In the Myocardial dataset, a binary classification problem, features with over 150 missing samples were excluded, and remaining samples with missing values were removed, resulting in 1024 samples and 94 features (1050 class zero, 24 class one). Table 1 summarizes the results.Table 1 summarizes the experimental results for the Myocardial dataset across different methods. ADASYN, SMOTE, and CTGAN reduced AUC scores, and CNN performed poorly, likely due to sparse feature representation in pixels. For XGBoost with RIGA, cGAN improved AUC by 0.1%, VQVAE by 0.6%, and VQGAN decreased AUC. These outcomes reflect the limited ability of models to capture meaningful structures due to sparse feature representation during transformation.

DARWIN Results

In the DARWIN dataset, we address a binary classification problem involving 451 features, one of which was an ID column removed prior to analysis. To induce class imbalance, we employed the same sample removal strategy outlined in Section , ensuring consistency across datasets. The final dataset contains 174 samples, a notably small size, which poses challenges for training and evaluating machine learning models. Table 1 presents the experimental results for the DARWIN dataset.

For this dataset, both ADASYN and SMOTE reduced AUC scores, highlighting their inability to generate synthetic samples that effectively represent the feature distributions in such a small dataset. CTGAN, by leveraging adversarial training, provided a modest improvement of 0.87%, showcasing its ability to capture more meaningful feature distributions. RIGA, combined with cGAN, achieved a slightly higher improvement of 0.98%, likely benefiting from its adversarial training tailored to specific transformations.

Interestingly, RIGA with VQVAE delivered the highest improvement, boosting AUC by 3.34%. This can be attributed to VQVAE's simpler architecture, which is well-suited for small datasets like DARWIN, as it avoids the overfitting risks associated with more complex models like VQGAN. In comparison, RIGA with VQGAN yielded a 1.87% improvement, falling short of VQVAE due to its complexity, which can struggle in scenarios with limited data. These

Method	nuMoM2b	Madelon	MI	DARWIN
CNN w/o Augmentation	0.7086 ± 0.0105	0.6072 ± 0.0110	0.6551 ± 0.0849	0.8152 ± 0033
CNN & cGAN	0.7136 ± 0.0220	0.6162 ± 0.0168	0.6139 ± 0.0793	0.8324 ± 0027
CNN & VQVAE	0.7309 ± 0.0140	0.6334 ± 0.0215	0.6398 ± 0.0832	0.8575 ± 0021
CNN & VQGAN	0.7368 ± 0.115	0.6512 ± 0.0173	0.6032 ± 0.0856	0.8324 ± 0042
XGBoost w/o Augmentation	0.7384 ± 0.0114	0.6738 ± 0.0490	0.8554 ± 0.0695	0.9252 ± 0010
XGBoost & cGAN	0.7566 ± 0.0126	0.6901 ± 0.0341	0.8565 ± 0.0693	0.9350 ± 0029
XGBoost & VQVAE	0.7572 ± 0.0135	0.7066 ± 0.0345	**0.8615 ± 0.0828**	**0.9586 ± 0.0046**
XGBoost & VQGAN	**0.7598 ± 0.0118**	**0.7355 ± 0.0505**	0.8459 ± 0.0784	0.9439 ± 0.0047
XGBoost & ADASYN	0.7162 ± 0.0060	0.6058 ± 0.0520	0.8342 ± 0.0990	0.9240 ± 0.0035
XGBoost & SMOTE	0.7183 ± 0.0030	0.6187 ± 0.0335	0.8366 ± 0.0812	0.9240 ± 0.0045
XGBoost & CTGAN	0.7442 ± 0.0138	0.6543 ± 0.0101	0.8321 ± 0.0937	0.9339 ± 0023

Table 1: AUC scores for different methods across datasets

results underscore the importance of selecting augmentation techniques that align with the size and structure of the dataset to maximize performance gains.

Bayesian Network Learning

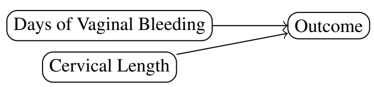

Figure 2: Markov Blanket without RIGA

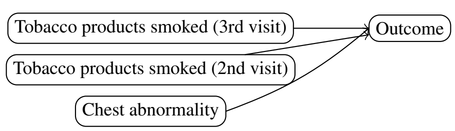

Figure 3: Markov Blanket with RIGA

As mentioned earlier, we compare the effect of RIGA on our U2 pipeline. Initially, the BIC score for the model trained on the imbalanced dataset was *-4.618e5*. After applying the RIGA pipeline, which augmented the dataset for a more balanced class distribution, the BIC score improved to *-4.855e5*. This decrease indicates that the Bayesian network fitted the augmented data more effectively, enhancing the representation of underlying relationships. Visualization of the learned Bayesian Network through the Markov Blanket revealed expanded variable interactions. Initially, the Markov blanket of the target variable included two parent features, which increased to three after augmentation, as shown in Figures 2 and 3. This expansion highlights improved clarity in variable interactions as class balance improved. Notably, the absence of certain features in the improved Markov Blanket doesn't imply their insignificance. These results underscore the importance of addressing class imbalances, enhancing both predictive capabilities and the understanding of data structure, thereby improving real-world applications.

Summary of results

Based on our results, VQGAN within the RIGA pipeline performed better on the nuMoM2b and Madelon datasets, which have larger sample sizes (360 and 500 features). Conversely, VQVAE excelled on the myocardial infarction and DARWIN datasets, which have smaller sample sizes (1,024 and 174 samples) and fewer features (94 and 450). This indicates that VQGAN's complex architecture benefits larger datasets, while VQVAE's simpler design suits smaller datasets, reducing overfitting. For instance, VQGAN effectively captures pixel-feature correlations in the nuMoM2b dataset but fails in the MI dataset, generating unrealistic single high-intensity pixel images, as shown in Figure 5 in the appendix.

Conclusions and Future Work

This work introduces RIGA, a novel pipeline addressing uncertainty and class imbalance in high-stakes AI applications like healthcare. By transforming tabular data into images, RIGA leverages generative models—cGAN, VQVAE, and VQGAN—to create balanced synthetic samples, improving classification performance. RIGA's inverse transformation enables generated images to be reverted back to their original tabular form, allowing seamless integration into downstream tasks. Experiments show RIGA enhances traditional classifiers like XGBoost, with VQGAN improving performance by 6.17% on Madelon and 2.14% on nuMoM2b, while VQVAE achieved 3.34% improvement on DARWIN. RIGA also strengthens Bayesian structure learning, enhancing robustness in imbalanced health datasets. Future work will incorporate an Error Feedback Loop to refine generative models and explore data augmentation's role in Bayesian learning to uncover hidden feature correlations.

Acknowledgements

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM013327 and a PSC-CUNY 2023 . The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Aggarwal, C. C. 2018. *Neural Networks and Deep Learning*, volume 10. Springer.

Andresini, G.; Appice, A.; Rose, L. D.; and Malerba, D. 2021. GAN augmentation to deal with imbalance in imaging-based intrusion detection. *Future Generation Computer Systems*, 123: 108–127.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Cilia, N. D.; De Gregorio, G.; De Stefano, C.; Fontanella, F.; Marcelli, A.; and Parziale, A. 2022. Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking. *Engineering Applications of Artificial Intelligence*, 111: 104822.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Golovenkin, S. V. R. D. S. P. N. S. O. Y., S.E.; and Voino-Yasenetsky, V. 2020. Myocardial infarction complications. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53P5M.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, volume 27.

Goretsky, A.; Dmitrienko, A.; Tang, I.; Lari, N.; Kunhardt, O.; Khan, R. R.; and Gyamfi-Bannerman, C. 2021. Data Preparation of the nuMoM2b Dataset. *medRxiv*. Preprint.

Guyon, I.; Gunn, S.; Nikravesh, M.; and Zadeh, L. A. 2006. *Feature Extraction: Foundations and Applications*. Springer Science & Business Media.

Haas, D. M.; Parker, C. B.; Wing, D. A.; Parry, S.; Grobman, W. A.; Mercer, B. M.; Simhan, H. N.; Hoffman, M. K.; Silver, R. M.; Wadhwa, P.; Iams, J. D.; Koch, M. A.; Caritis, S. N.; Wapner, R. J.; Esplin, M. S.; Elovitz, M. A.; Foroud, T.; Peaceman, A. M.; Saade, G. R.; Willinger, M.; and Reddy, U. M. 2015. A description of the methods of the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b). *American Journal of Obstetrics and Gynecology*, 212(4): 539.e1–539.e24.

He, H.; Bai, Y.; Garcia, E. A.; and Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning.

Huang, H.; Woodruff, T. J.; Baer, R. J.; Bangia, K.; August, L. M.; Jellife-Palowski, L. L.; ...; and Sirota, M. 2018. Investigation of association between environmental and socioeconomic factors and preterm birth in California. *Environment international*, 121: 1066–1078.

Iams, J. D.; Goldenberg, R. L.; Mercer, B. M.; Moawad, A. H.; Meis, P. J.; Das, A. F.; and Roberts, J. H. 2001. The preterm prediction study: Can low-risk women destined for spontaneous preterm birth be identified? *American journal of obstetrics and gynecology*, 184(4): 652–655.

Koller, D.; and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge Massachusetts: The MIT Press.

Koullali, B.; Oudijk, M. A.; Nijman, T. A.; Mol, B. W. J.; and Pajkrt, E. 2016. Risk assessment and management to prevent preterm birth. *Seminars in Fetal and Neonatal Medicine*, 21(2): 80–88.

Maharana, K.; Mondal, S.; and Nemade, B. 2022. A Review: Data Pre-processing and Data Augmentation Techniques. *Global Transitions Proceedings*, 3(1): 91–99.

Mallia, D. 2023. *Towards an unsupervised Bayesian network pipeline for explainable prediction, decision making and discovery*. Ph.D. thesis, City University of New York

Manuck, T. A. 2017. Racial and ethnic differences in preterm birth: a complex, multifactorial problem. *Seminars in perinatology*, 41(8): 511–518.

Mikołajczyk, A.; and Grochowski, M. 2018. Data Augmentation for Improving Deep Learning in Image Classification Problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 117–122. IEEE.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. *CoRR*, abs/1411.1784.

Mumuni, A.; and Mumuni, F. 2022. Data augmentation: A comprehensive survey of modern approaches. *Array*, 100258.

Nanni, L.; Paci, M.; Brahnam, S.; and Lumini, A. 2021. Comparison of Different Image Data Augmentation Approaches. *Journal of Imaging*, 7(12): 254.

Purisch, S. E.; and Gyamfi-Bannerman, C. 2017. Epidemiology of preterm birth. *Seminars in perinatology*, 41(7): 387–391.

Sharma, A.; Vans, E.; Shigemizu, D.; Boroevich, K. A.; and Tsunoda, T. 2019. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific Reports*, 9(1): 11399.

Van Den Oord, A.; and Vinyals, O. 2017. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30.

Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling Tabular data using Conditional GAN.

Xu, M.; Yoon, S.; Fuentes, A.; and Park, D. S. 2023. A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recognition*, 137: 109347.

Zhu, Y.; Brettin, T.; Xia, F.; et al. 2021. Converting tabular data into images for deep learning with convolutional neural networks. *Scientific Reports*, 11: 11325.

Appendix

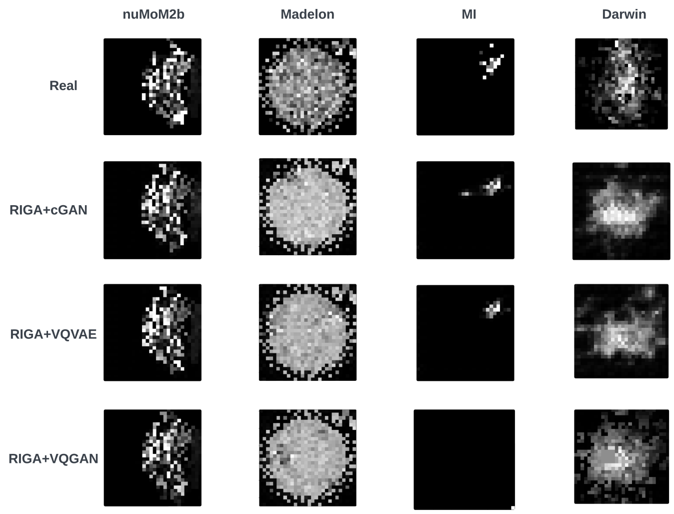

Figure 4: Real and Fake Images of four datasets.

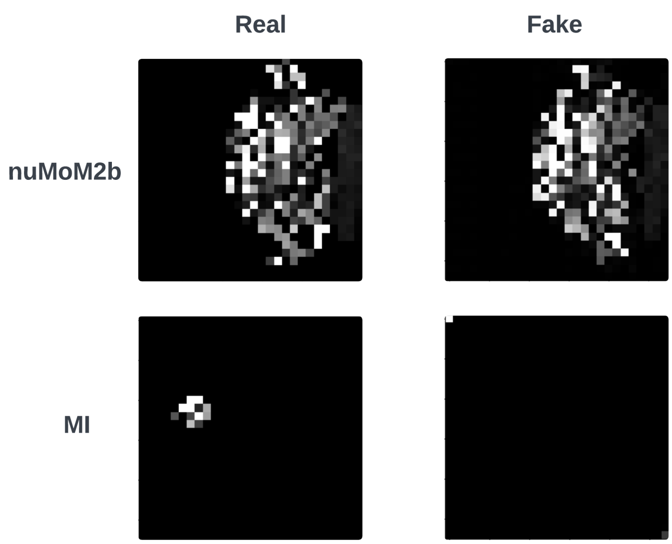

Figure 5: Real and Synthetic Images Generated by RIGA-VQGAN for nuMoM2b and MI Datasets.