Journal Title Here, 2023, pp. 1–21

doi:

Signature Informed Sampling for Transcriptomic Data

Nikita Janakarajan,[1,2*] Mara Graziani[1] and María Rodríguez Martínez[1]

[1]Accelerated Discovery, IBM Research Europe, Zürich, Switzerland and [2]Institute of Machine Learning, D-INFK, ETH Zürich, Zürich, Switzerland

[*]Corresponding author: nja@zurich.ibm.com

Abstract

Working with transcriptomic data is challenging in deep learning applications due to its high dimensionality and low patient numbers. Deep learning models tend to overfit this data and do not generalize well on out-of-distribution samples and new cohorts. Data augmentation strategies help alleviate this problem by introducing synthetic data points and acting as regularisers. However, existing approaches are either computationally intensive or require parametric estimates. We introduce a new solution to an old problem - a simple, non-parametric, and novel data augmentation approach inspired by the phenomenon of chromosomal crossover. Based on the assumption that there exist non-overlapping gene signatures describing each phenotype of interest, we demonstrate how new synthetic data points can be generated by sampling gene signatures from different patients under certain phenotypic constraints. As a case study, we apply our method to transcriptomic data of colorectal cancer. Through discriminative and generative experiments on two different datasets, we show that our method improves patient stratification by generating samples that mirror biological variability as well as the models' robustness to overfitting and distribution shift. Our approach requires little to no computation, and outperforms, or at the very least matches, the performance of established augmentation methods.

Key words: Data Augmentation, Transcriptomics, Machine Learning

Introduction

The application of machine learning in biomedicine, particularly for classification tasks, has become increasingly popular over the last decade [1, 2]. High-throughput sequencing has provided access to transcriptomic data, catalyzing an increase in machine learning research aimed at pattern discovery, phenotype classification, disease subtyping, survival analysis, and more. However, the inherent challenges in this domain persist, notably the high dimensionality of data and limited sample numbers [3]. These challenges make machine learning models prone to overfitting, and reduces their ability to generalize to new datasets [4], even from different institutions [5].

To address these challenges, multiple techniques for dimensionality reduction and data augmentation have been developed. On one hand, dimensionality reduction methods help mitigate overfitting and enhance generalisation by only selecting highly informative features [6, 7], or combinations of features such as in Principal Component Analysis (PCA) [8], Non-Negative Matrix Factorisation (NMF) [9], and Autoencoders [10]. On the other hand, data augmentation generates new data points from existing ones, addressing dataset imbalances and under-representation. Common data augmentation techniques involve noise injection-based methods, particularly in the imaging domain [11, 12]. The noise perturbation prevents complex models

from memorizing the training data [13]. Applying noise injection to transcriptomic and biological data, however, requires caution since the noise may potentially disrupt the biological signal. For this, other augmentation approaches were proposed for transcriptomic data, including parametric and non-parametric strategies. Parametric approaches assume specific data distributions, such as negative binomial [14] or Poisson [15], to generate new data points. Unfortunately, these approaches can be compromised by extreme data support values or sensitivity to outliers. Non-parametric techniques, such as random oversampling [16, 17, 18], do not have this issue, as they bypass distribution assumptions but cannot generate new samples to enhance diversity and improve robustness. In this paper, we address some of these limitations by proposing novel non-parametric approaches for transcriptomic data augmentation.

Our method is adapted to the biology domain, and leverages phenotypic gene signatures – sets of genes with distinct expression patterns that hold prognostic, diagnostic or predictive value – to drive the augmentation. Gene signatures are invaluable for phenotypic predictions, and are especially useful in cancer subtyping [19, 20, 21, 22, 23]. They can be leveraged to reduce the dimensionality of the data, where each patient is represented by the set of all gene signatures associated with the phenotypes instead of the entire 60,000 genes.

Although gene signatures reduce the dimension of the feature space, their integration with computational methods such as machine learning models remains challenging due to limited patient samples. Here we present a solution to this problem by leveraging identified gene signatures to generate new data. Specifically, we make the following assumptions: (i) Highly predictive gene signatures are available for each variant of the phenotype of interest; (ii) Gene signatures of phenotype variants are non-overlapping; (iii) The genes associated with each gene signature can interact with other genes from the same signature (directly or indirectly).

Our approach is inspired by chromosomal crossover and involves performing crossovers between patients' gene signatures within and between phenotype variants. The study evaluates this method against standard techniques, including random oversampling, current state-of-the-art SMOTE [24], and parametric approaches like negative binomial and Poisson sampling. To make the latter a challenging baseline, we propose a modification of the original approaches with respect to the sampling step. Briefly, we use local parameter estimations to generate each new sample. The modifications are explained in detail in 3.2 and 3.3, respectively. To demonstrate the effectiveness of our sampling methods, the TCGA COADREAD dataset is used, and generalisation performance is assessed using the external validation of the CPTAC COAD dataset. The method successfully addresses class imbalance, improves model performance, and offers a promising solution for data augmentation in the biomedical domain.

Our contributions are as follows:

1. We introduce a novel, non-parametric data augmentation method and benchmark it against conventional techniques.
2. We demonstrate the utility of the negative binomial and Poisson distributions as data augmentation strategies that alleviate class imbalance and enhance the diversity of generated samples.
3. We highlight the influence of augmentation size on model performance through a series of experiments in both discriminative and generative settings.

Related Work

Transcriptomics data augmentation strategies vary in their approaches and effectiveness. One of the simplest methods is non-parametric random sampling with replacement [16]. While effective for constructing accurate classifiers [25], it lacks diversity in newly generated samples. To increase sample diversity, [26] proposed a weighted mixing of observations, inspired by the concept of mixing images [11]. Here, observations are multiplied by small random weights to create artificial samples. The advantage of this technique is that all observations contribute to the generation of new samples. However, assigning unique labels to such samples can be challenging, making this approach less suitable for supervised learning.

Parametric methods rely on population statistics and data models. The Poisson distribution is a widely used distribution to model RNA-Seq count data [15, 27, 28], but assumes equal variance and mean. This limits its effectiveness for count data with a large support [29] and fails to capture over-dispersion in the data. The negative binomial distribution, also known as the gamma-Poisson distribution, is another widely used distribution that captures the over-dispersion of RNA-Seq count data [30, 31].

To solve class imbalance, augmenting transcripomic data using parametric methods requires class-specific parameter estimation. Challenges arise when the data distribution support is too large, leading to low data density in some regions and potentially generating uninformative data distributions. Conversely, limited support can result in insufficient sample diversity. Additionally, out-of-distribution data points can also impact parameter inference, rendering these methods sensitive to outliers and measurement errors [32].

Oversampling by interpolation is a popular method of augmenting biological data. Synthetic Minority Oversampling Technique (SMOTE) [24] and its variants [25, 33, 34, 35, 36, 37] are prime examples of this sampling strategy. Primarily developed to target class imbalance, SMOTE generates a new sample on the line connecting a reference sample and a randomly chosen sample from its neighbourhood. This strategy has achieved state of the art performances on tasks affected by class imbalances. However, it has been found that for high-dimensional datasets SMOTE does not attenuate the bias towards the majority class [38]. Furthermore, when SMOTE is applied to high-dimensional data, it preserves the expected value of the minority class while decreasing its biological variability, which can be problematic for classifiers that rely on class-specific variances [38]. Moreover, SMOTE can be computationally demanding due to extensive pairwise distance calculations.

Alternative methods for data expansion include aggregation and the generation of synthetic samples *in silico* [39, 40, 41]. Generative adversarial network (GAN) [42] and its variants are used to create synthetic gene expression samples [43, 44, 45, 26], but are computationally intensive, difficult to train, and require detailed knowledge about gene regulatory networks.

In contrast, our proposed signature-based crossover sampling approach is computationally efficient and requires no distribution parameter inference. It offers versatility for various scenarios, making it a promising addition to transcriptomic data augmentation methods.

Proposed Augmentation Techniques

Crossover sampling

Inspired by chromosomal crossover during meiosis, phenotype-driven sampling aims to create new data points from existing data points while increasing diversity in the generated samples. Simply put, to generate new samples for a given phenotype, we leverage the high informativeness of the associated gene signatures for that particular phenotype, while assuming that the signatures for the other phenotypes in that sample are less informative for the given phenotype. We investigate two ways of augmenting the data samples:

1. Crossing over between samples having the same phenotypic variant.
2. Crossing over between samples across all phenotypic variants.

In the following sections, we describe the details of our method and then illustrate its applicability to the gene expression of patients with colorectal cancer from TCGA as the data source.

Intra-class crossover sampling

To limit the likelihood of mixing phenotypes, and hence introducing samples of dubious labels, we first perform crossing over between samples belonging to the same phenotype. Thus, in this setting, new samples are generated only from samples having the same phenotype and so naturally, the synthetic sample inherits that phenotype. The crossing-over between samples is done at the gene signature level. Consider a dataset with n samples x_i, where $i \in 1, ..., n$, and m phenotype variants P_j, where $j \in 1, ..., m$. Each sample x_i is of one phenotype P_j. Given each variant P_j is associated with a gene signature S_j, we represent each sample by the set of signatures associated with all possible variants, that is, $x_i = \{S_1, ..., S_m\}$. The number of genes in a signature is given by $|S_j|$. Therefore, each patient is represented by $\sum_{j=1}^{m} |S_j|$ genes. For the sake of simplicity, the genes are ordered by signature - all genes belonging to signature S_1 come first and so on, thus creating signature blocks, one for each phenotypic variant. To generate a new sample, x_{n+1} with phenotype $P_{j=1}$, we first subset all samples, belonging to this phenotype (Equation 1). This subset, $X_{j=1}$, is analogous to a bag of signature blocks (Equation 2).

$$X_{j=1} = \{x_i : x_i \in P_{j=1} \forall i \in 1, ..., n\} \quad (1)$$

$$X_{j=1} = \{S_{ij} : j \in 1, ..., m \forall i : x_i \in P_{j=1}\} \quad (2)$$

We randomly sample signature blocks from this subset (Equation 3) which are put together to create a new sample x_{n+1}.

$$x_{n+1} = \{S_j \in_R X_{j=1} \forall j \in 1, ..., m\} \quad (3)$$

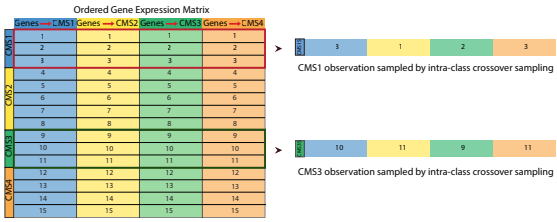

Fig. 1. An illustration of intra-class crossover sampling. A sample with class CMS1 can be generated by sampling gene signature blocks from the subset of CMS1 patients (surrounded by the red box). Similarly, an observation with class CMS3 can be generated by sampling gene signature blocks from the subset of CMS3 patients (surrounded by the dark green box).

We illustrate a practical example on the gene expression of patients with colorectal cancer, using the Consensus Molecular Subtype (CMS) as phenotype. This phenotype has four variants, namely, CMS1, CMS2, CMS3, and CMS4 [46]. A study conducted in 2020 found a set of 10 genes that are highly predictive for each CMS class [47]. This greatly reduces the dimensionality of our gene expression dataset and gives us a panel of 40 genes that can be used for classifying patients into the 4 colorectal cancer subtypes. Figure 1 illustrates the process of intra-class crossover sampling. Consider an ordered gene matrix where the patients and the genes are sorted according to their CMS class. To generate a new CMS1 sample, we first consider the subset of all CMS1 patients (highlighted by a red rectangle in Figure 1). To create a new CMS1 sample, we randomly and sequentially sample one gene signature block represented by the blue, yellow, green, and orange columns, each comprising genes predictive of the CMS1, CMS2, CMS3, and CMS4 classes.

Since we are sampling the blocks only from the subset of patients belonging to CMS1, the expression values of genes associated with the other CMS types are typical of what is observed in a CMS1 patient. Thus, we minimise the likelihood of creating ambiguous samples where the combined expression values indicate a different CMS than the one assigned. This process is repeated for the other CMS types, that is, a CMS2 patient is sampled only from the subset of patients belonging to CMS2, and so on.

Inter-class crossover sampling

Inter-class crossover sampling aims at utilising the entire dataset to create new samples, as opposed to just the phenotype specific subset in intra-class crossover sampling. In this method, we investigate the effect of sampling signature blocks across all samples as opposed to just a subset of samples. To avoid ambiguous labels, we add a constraint that prevents sampling the defining signature block from a sample unless it belongs to the phenotype to be sampled. To formalise this, consider a similar premise as the one described in subsubsection 3.1.1. To generate a synthetic sample of phenotype $P_{j=1}$, we sample $S_{j=1}$ only from the subset of patients, $X_{j=1}$, that have phenotype $P_{j=1}$, that is. $S_j \in_R X_j$. We then randomly sample the remainder signatures S_k for $k \in 1, ..., m$ and $k \neq j$, from all samples except those that belong to phenotype P_k as described in Equation 4. The sampled signatures are then put together to create a new synthetic sample, $x_{n+1} = \{S_j, S_k\}$ for $k \in 1, ..., m$ and $k \neq j$ and where $j = 1$.

$$S_k \in_R \{X - X_k\} \; \forall k \in 1, ..., m \text{ and } k \neq j \qquad (4)$$

This method is illustrated in Figure 2 in the context of CMS classes. When creating a CMS1 sample, the CMS1 gene block has to come from the set of patients belonging to CMS1, as this generates expression values typical of a patient with CMS1. However, the CMS2 block can come from all patients except those belonging to CMS2, since the expression values are in a range that is not predictive of CMS2. If it comes from a CMS2 patient, then the class label of the new sample being generated would have multiple labels, namely CMS1 and CMS2. Due to this ambiguity in label assignment, we avoid sampling the CMS2-associated gene block from a CMS2 patient. Similarly, the CMS3- and CMS4-associated gene blocks can come from all patients except those belonging to the CMS3 and CMS4 classes respectively.

Modified gamma-Poisson sampling

The negative-binomial (NB) distribution is a commonly used parametric method for augmenting RNA-Seq data for differential expression analysis due to it's ability to capture the overdispersion typically seen in these data. To model the RNA-Seq read frequency Y of a given gene j, the distribution relies on two parameters, namely the mean and dispersion of the counts, i.e, $Y_j = \text{NB}(\mu_j, \phi)$.

Here, ϕ is the dispersion parameter that controls the variance of the counts, $Var(Y_j) = \mu_j + \phi\mu_j^2$. The estimation of ϕ is, however, not trivial. The negative-binomial (NB) distribution is typically expressed as

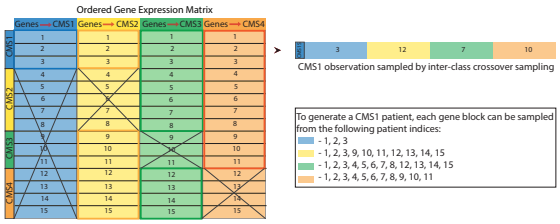

Fig. 2. An illustration of inter-class crossover sampling. A sample with class CMS1 can be generated by sampling the CMS1 gene signature block from the subset of CMS1 patients (surrounded by the blue box). The remaining gene blocks can be sampled from those patients that do not belong to the CMS class the gene block predicts.

a gamma-Poisson mixture [48, 49] to make it more tractable. In this formulation, the rate parameter λ of the Poisson distribution is a random variable sampled from a gamma distribution parameterised by the shape (α) and rate (β) parameters, which are estimated from the observed population, $\lambda \sim \Gamma(\alpha, \beta)$. The RNA-Seq read counts are then sampled from this initialised Poisson distribution, $Y_j \sim Poisson(\lambda_j)$.

To ensure that the newly generated samples add some diversity while maintaining the same distribution as the augmented class, we create a mixture of gamma-Poisson distributions. In this strategy, to generate every new observation, a subset S of samples is randomly chosen to estimate the α and β parameters from its mean μ and variance σ^2 as shown in Equation 5. These parameters initialise the gamma distribution. Since the augmentation is aimed at solving class imbalance, the subset is created from samples belonging to the same class. Thus, the newly generated observation will have the same label as the subset. The size of the subset $|S|$ is a hyperparameter defined by the user. The random variable sampled from this gamma distribution is used to initialise the Poisson distribution from which a new observation is sampled. This process is repeated n times to generate n new observations. We recommend a smaller subset size to maximise the variance in the subset. A larger variance would mean more scattered observations, thereby generating a diverse set of samples as opposed to dense local clusters with a smaller variance.

$$\mu = \frac{1}{|S|}\sum_{x \in S} x, \quad \sigma^2 = \frac{1}{|S|}\sum_{x \in S}(x-\mu)^2$$
$$\alpha = \frac{\mu^2}{\sigma^2}, \quad \beta = \frac{\mu}{\sigma^2} \tag{5}$$

Table 1. Sample counts of the different colorectal cancer subtypes in the TCGA and CPTAC gene expression datasets.

Datasets	CMS1	CMS2	CMS3	CMS4	Total
TCGA	74	216	74	142	506
CPTAC	14	33	16	22	85

Modified Poisson sampling

The Poisson distribution is another method commonly used to model RNA-Seq read counts. Similarly to the gamma-Poisson strategy, a subset is first sampled from the class to be augmented, after which the mean is computed and set as the rate parameter λ for the Poisson distribution as shown in Equation 6.

$$\lambda = \frac{1}{|S|}\sum_{x \in S} x\ , \quad Y_j \sim Poisson(\lambda_j) \tag{6}$$

However, unlike the gamma-Poisson, this distribution assumes the variance of the sample equals the mean and fails to capture overdispersion in the data. To generate new observations with this method, a larger subset size is recommended such that the variance is closer to the mean while still being high enough to generate variability in the new data.

Materials and Methods

Datasets

TCGA COADRED (Colon Adenocarcinoma and Rectal Adenocarcinoma)[50] and CPTAC[51] are used for all experiments. TCGA COADREAD is designated for training and in-domain testing, while the CPTAC Colon Cancer gene expression data are used for external validation. As an initial preprocessing step, we filter out replicates and only consider patients with primary tumours and available CMS labels. Summary of the class counts of the datasets are shown in Table 1. Additionally, since we are interested in colorectal cancer subtyping, we only consider the 40 signature genes of the subtypes [47]. Details on dataset creation are described in Supplementary A.3.

Models

For all discriminative tasks, we consider 6 different classifiers, namely, Logistic Regression (LR), K-Nearest Neighbours (KNN), Support Vector Machines with RBF kernel (SVM), Explainable Boosting Machines (EBM), Random Forests (RF) and Multi-Layer Perceptron (MLP). For the generative tasks, we

consider a variational autoencoder [52]. Implementation details for the discriminative and generative experiments are in Supplementary A.4 and A.5, respectively.

Evaluation Criteria

We report the balanced accuracy (Equation 7) for the discriminative tasks as it takes into account both the sensitivity and specificity of the model, and is particularly useful when dealing with imbalanced datasets. If the support samples for each class exhibit sufficient and meaningful diversity, the performance of a standard classification model is expected to outperform one trained on the unbalanced dataset. In the generative tasks, we evaluate the latent embeddings in terms of cluster purity (Equation 8), a K-means based metric that quantifies the homogeneity of clusters. The basis for this metric is that high quality latent embeddings will effectively separate data into distinct clusters. We perform a 5-fold stratified cross-validation repeated 5 times for each augmentation method and augmentation size. Performance metrics are reported as the average across all splits.

Results and Discussions

Tuning the parametric sample set size

As described in section 3, the gamma-Poisson and Poisson methods require a sample set S of a predefined size $r = |S|$ from which parameters are estimated and a new sample is generated. To set the hyperparameter r for the gamma-Poisson method, we explore a range of values and select the optimal size by testing the average performance of the augmented dataset in the CMS prediction task across various classification models (Supplementary A.2). On average, we find $r = 5$ to be the optimal value across all models (Supplementary Figure 6). For the modified Poisson sampling, we double the sample size to $r = 10$ to help mitigate the effects of over-dispersion while still ensuring diversity in terms of variability in the newly generated samples (Supplementary Figure 7).

Visualising the augmented data

A qualitative assessment of the augmentation methods in capturing trends in the data is shown in Figure 3. We report the results for class size 5000, as the trends are more evident due to the large size. Similar trends are observed in smaller class sizes although less distinctly (Supplementary Figure 4, Figure 5). Visually, both our crossover and gamma-Poisson sampling strategies capture data trends better than the other methods. We find that the intra-class crossover, inter-class crossover and gamma-Poisson sampling methods (Figure 3B,C,D) best capture both the global and local structure of the original unbalanced data Figure 3A. The real and synthetic samples are interspersed with the newly generated samples and in general, the synthetic samples appear well distributed over the latent space. The Poisson sampled data (Figure 3E) form unique clusters for each subtype, whereas in the unbalanced data visualisation Figure 3A, we see mixing of samples between clusters. These points could potentially correspond to samples of a mixed subtype or shared transcriptional states, which the data augmented by the Poisson distribution does not reflect. The data augmented by SMOTE are generated by interpolating between reference samples in close proximity to each other, and therefore fail to occupy the input space evenly, as shown in the inset of Figure 3F.

Discriminative modeling of augmented data

We conduct quantitative experiments to determine if our data augmentation is representative of the original data and contributes to improving the accuracy on downstream tasks. The aim of this task is to predict the colorectal cancer subtype (CMS), which is challenging due to high class imbalance in the data (Table 1). We use the 6 data augmentation methods described in Section A.3 to balance classes during training, and evaluate various augmentation sizes (subsection 4.1) using the classifiers mentioned in subsection 4.2.

The results of this experiment are reported in Table 2 for three types of classifiers, namely SVM, EBM and RF, and in Supplementary Table 7 for the rest. While training on gamma-Poisson augmented data improves in-domain test performance, training on the inter-class crossover augmented data enables the models to generalise better to the CPTAC external dataset. Interestingly, as the class size increases, the in-domain performance associated with both cross-over methods decreases, whereas the performance on the external dataset increases. This suggests that the crossover methods generate new samples that are *out-of-distribution* with respect to the given dataset, which in turn provide better support for learning a robust decision boundary. Moreover, with the inter-class crossover method, augmenting only the minority classes (class size

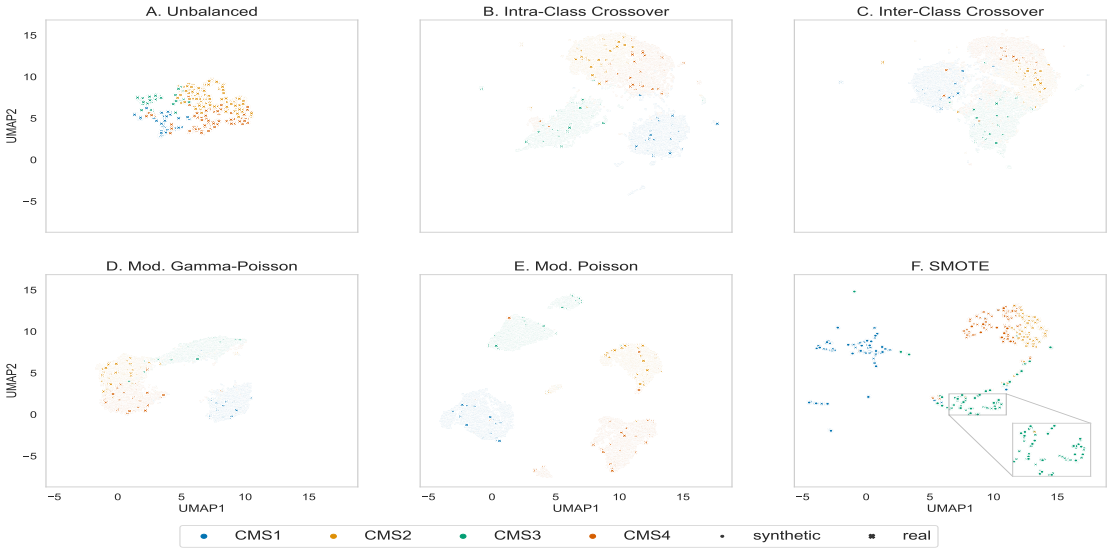

Fig. 3. UMAP [53] visualisations of datasets in the study. A. Unbalanced dataset. B. Intra-class crossover sampled dataset. C. Inter-class crossover sampled dataset. D. Gamma-Poisson sampled dataset. E. Poisson sampled dataset. F. SMOTE sampled dataset. All augmented datasets have 5000 samples per class.

"Max") is sufficient for an in-domain performance that is as good as that with the gamma-Poisson at higher class sizes. For instance, the SVM achieves a balanced accuracy score of 89.6% for data augmented to class size "Max" by the by inter-class crossover method, whereas the modified gamma-Poisson needs to generate more data for the SVM to achieve similar performances – 90.1% at class size 500 and 89.7% at class size 5000. Therefore, inter-class crossover proves to be a suitable method for oversampling minority classes in in-domain data. It creates a variety of samples that allow models to learn robust decision boundaries, which can generalize well to external data, particularly when augmentation sizes are increased.

Generative modeling of augmented data

We now consider generative modeling and evaluate how effectively the augmented data contributes to estimating a parametric representation of the data. We expect greater diversity in the input samples to lead to more accurate parametric estimates of the latent space distribution.

We train a variational autoencoder (VAE) on the augmented data with the ELBO loss as the objective [52] to learn a compressed 4-D representation of the data. The cluster purity of the test data in the latent space is reported in Table 3. We observe that the inter-class crossover method has the highest cluster purity score for both TCGA and CPTAC. When data is augmented to a class size of 5000, it leads to significantly better (p-value¡0.02) performance than modified gamma-Poisson, modified Poisson, SMOTE and unbalanced datasets. The modified Poisson method is associated with the best performance at lower class sizes but it is only significantly better for TCGA data with class size 500 (p-value ¡ 0.05). The performance of modified Poisson can be explained by its tendency to create distinct clusters (Figure 3, Figure 5 and Figure 4), which are, however, not reflective of the original distribution. When increasing in the class size, the method generates artefacts that cause the drop in latent cluster purity, suggesting lower embedding quality. Conversely, the inter-class crossover method consistently maintains a high performance on both TCGA and CPTAC data across all augmentation sizes. This shows that our data augmentation method facilitates the learning of a latent space with high fidelity to the original data.

Next, we investigate if the learned 4-D compression of the data is informative for the prediction of

Table 2. 5x5 cross-validation results for supervised classification (CMS prediction) on augmented data with various classifiers. Each classifier is trained with data augmented by different sampling methods. Each class is augmented to a different class-size - Max, 500, and 5000. The TCGA test set indicates in-domain performance. The CPTAC dataset serves as external validation. We report the average balanced accuracy scores and their standard deviation, with the best highlighted in bold. Although in-domain performances vary depending on the classifier and class-size, models trained with the inter-class crossover augmented data generalise best in 6 out of the 9 *out-of-distribution* settings (i.e, CPTAC).

Class-Size	Sampling Methods	SVM		EBM		RF	
		TCGA	CPTAC	TCGA	CPTAC	TCGA	CPTAC
Max	Intra-Crossover (Ours)	0.890 ± 0.029	0.594 ± 0.033	0.888 ± 0.026	0.647 ± 0.035	0.876 ± 0.030	0.643 ± 0.021
	Inter-Crossover (Ours)	0.896 ± 0.028	$\mathbf{0.608 \pm 0.030}$	$\mathbf{0.893 \pm 0.032}$	$\mathbf{0.670 \pm 0.026}$	$\mathbf{0.878 \pm 0.039}$	0.625 ± 0.023
	Mod. Gamma-Poisson	0.892 ± 0.028	0.600 ± 0.033	0.882 ± 0.032	0.647 ± 0.029	0.876 ± 0.035	0.645 ± 0.026
	Mod. Poisson	0.882 ± 0.036	0.602 ± 0.031	0.875 ± 0.029	0.623 ± 0.031	0.865 ± 0.034	0.627 ± 0.038
	SMOTE	0.894 ± 0.025	0.584 ± 0.030	0.88 ± 0.03	0.633 ± 0.030	0.871 ± 0.034	$\mathbf{0.653 \pm 0.028}$
	Resampling	$\mathbf{0.897 \pm 0.031}$	0.596 ± 0.032	0.871 ± 0.030	0.628 ± 0.026	0.875 ± 0.032	0.632 ± 0.038
500	Intra-Crossover (Ours)	0.885 ± 0.023	0.592 ± 0.041	0.884 ± 0.032	0.683 ± 0.029	0.868 ± 0.036	$\mathbf{0.673 \pm 0.020}$
	Inter-Crossover (Ours)	0.881 ± 0.037	$\mathbf{0.643 \pm 0.028}$	0.877 ± 0.039	$\mathbf{0.708 \pm 0.023}$	0.873 ± 0.038	0.669 ± 0.027
	Mod. Gamma-Poisson	$\mathbf{0.901 \pm 0.023}$	0.608 ± 0.031	$\mathbf{0.895 \pm 0.026}$	0.669 ± 0.025	0.876 ± 0.040	0.658 ± 0.016
	Mod. Poisson	0.878 ± 0.034	0.601 ± 0.035	0.877 ± 0.029	0.615 ± 0.021	0.862 ± 0.038	0.627 ± 0.039
	SMOTE	0.894 ± 0.032	0.583 ± 0.032	0.879 ± 0.028	0.642 ± 0.031	$\mathbf{0.876 \pm 0.030}$	0.665 ± 0.032
	Resampling	0.889 ± 0.028	0.563 ± 0.038	0.862 ± 0.032	0.618 ± 0.040	0.872 ± 0.035	0.623 ± 0.029
5000	Intra-Crossover (Ours)	0.876 ± 0.028	0.613 ± 0.024	0.871 ± 0.033	0.659 ± 0.040	0.861 ± 0.029	0.677 ± 0.028
	Inter-Crossover (Ours)	0.870 ± 0.033	$\mathbf{0.662 \pm 0.014}$	0.846 ± 0.043	0.654 ± 0.023	0.849 ± 0.031	$\mathbf{0.685 \pm 0.019}$
	Mod. Gamma-Poisson	$\mathbf{0.897 \pm 0.030}$	0.614 ± 0.021	$\mathbf{0.893 \pm 0.026}$	$\mathbf{0.674 \pm 0.033}$	0.879 ± 0.032	0.677 ± 0.013
	Mod. Poisson	0.876 ± 0.035	0.600 ± 0.033	0.882 ± 0.031	0.61 ± 0.03	0.871 ± 0.037	0.634 ± 0.043
	SMOTE	0.893 ± 0.027	0.589 ± 0.032	0.883 ± 0.027	0.637 ± 0.031	$\mathbf{0.879 \pm 0.025}$	0.661 ± 0.024
	Resampling	0.886 ± 0.028	0.580 ± 0.038	0.866 ± 0.033	0.621 ± 0.032	0.863 ± 0.034	0.627 ± 0.026
	Unbalanced	0.886 ± 0.032	0.606 ± 0.029	0.867 ± 0.027	0.637 ± 0.026	0.862 ± 0.034	0.626 ± 0.036

Table 3. 5x5 cross-validation results of latent space cluster purity. VAEs are trained with data augmented by different sampling methods and to different class-sizes to learn a 4D latent representation. The latent spaces are quantitatively assessed by cluster purity. The best scores are highlighted in bold. The inter-class crossover method when augmented to class-size 5000 has the highest in-domain and external purity scores.

Sampling Methods	Max		500		5000	
	TCGA	CPTAC	TCGA	CPTAC	TCGA	CPTAC
Intra-Crossover (Ours)	0.739 ± 0.052	0.618 ± 0.051	0.757 ± 0.050	0.635 ± 0.060	0.762 ± 0.041	0.634 ± 0.042
Inter-Crossover (Ours)	0.753 ± 0.053	$\mathbf{0.624 \pm 0.065}$	0.739 ± 0.050	0.678 ± 0.062	$\mathbf{0.772 \pm 0.032}$	$\mathbf{0.680 \pm 0.038}$
Mod. Gamma-Poisson	0.735 ± 0.055	0.605 ± 0.046	0.732 ± 0.068	0.639 ± 0.053	0.719 ± 0.034	0.677 ± 0.040
Mod. Poisson	$\mathbf{0.772 \pm 0.070}$	0.609 ± 0.048	$\mathbf{0.791 \pm 0.041}$	$\mathbf{0.683 \pm 0.040}$	0.733 ± 0.041	0.636 ± 0.036
SMOTE	0.731 ± 0.094	0.586 ± 0.048	0.735 ± 0.065	0.605 ± 0.038	0.531 ± 0.109	0.484 ± 0.104
Resampling	0.724 ± 0.069	0.605 ± 0.045	0.760 ± 0.053	0.590 ± 0.038	0.731 ± 0.052	0.588 ± 0.042
Unbalanced			0.740 ± 0.058	0.637 ± 0.054		

MSI, CIMP and OS, which are key clinical variables of colorectal cancer phenotypes. Predicting these variables, among others such as demographic factors, genomic markers and treatment response, is relevant for patient prognostics, disease monitoring and disease subtyping. Here, we discuss explicitly MSI and CIMP prediction, and report the survival analysis task results in Supplementary subsubsection A.5.1. The implementation details of all tasks are also described in Supplementary A.5.

Predicting MSI: MSI is a valuable marker for prognosis [54, 55], treatment choice [56, 57] and colorectal cancer subtyping. CMS1 shows a high level of microsatellite instability (MSI-H), whereas CMS2 and CMS4 are microsatellite stable. CMS3, on the other hand, shows a mixed MSI status (MSI-Low). We train both an SVM and an EBM to predict the MSI status from the 4D latent embeddings of the VAE models.

The balanced accuracy from a 5-fold stratified 5 times repeated cross-validation for the class size

Table 4. 5x5 cross-validation results for MSI and CIMP Classification from VAEs trained with data augmented by different sampling methods. We report the averaged balanced accuracy for class-size 5000. Results for other class sizes are shown in Supplementary Table 11 and Table 10, respectively. The best scores are highlighted in bold. For MSI classification, all data augmentation models – except for SMOTE – achieved comparable balanced accuracy with both SVM and EBM in both in-domain and external validation. For CIMP classification, the inter-class crossover method is significantly better than all the other augmentation methods for the SVM model in in-domain testing, and significantly better than Poisson, SMOTE and unbalanced (p-value ≤ 0.05) for the EBM.

Clinical Variable	Sampling Methods	SVM		EBM	
		TCGA	CPTAC	TCGA	CPTAC
MSI	Intra-Crossover (Ours)	0.570 ± 0.034	0.884 ± 0.046	$\mathbf{0.555 \pm 0.047}$	0.863 ± 0.047
	Inter-Crossover (Ours)	0.569 ± 0.041	0.889 ± 0.044	0.552 ± 0.042	$\mathbf{0.867 \pm 0.042}$
	Mod. Gamma-Poisson	0.566 ± 0.040	0.845 ± 0.051	0.550 ± 0.039	0.761 ± 0.067
	Mod. Poisson	0.557 ± 0.032	0.848 ± 0.062	0.542 ± 0.049	0.794 ± 0.051
	SMOTE	0.390 ± 0.088	0.578 ± 0.130	0.373 ± 0.067	0.558 ± 0.109
	Resampling	$\mathbf{0.578 \pm 0.033}$	$\mathbf{0.891 \pm 0.048}$	0.548 ± 0.042	0.814 ± 0.045
	Unbalanced	0.565 ± 0.037	0.838 ± 0.062	0.527 ± 0.045	0.795 ± 0.071
CIMP	Intra-Crossover (Ours)	0.552 ± 0.052	—	0.528 ± 0.043	—
	Inter-Crossover (Ours)	$\mathbf{0.576 \pm 0.043}$	—	$\mathbf{0.538 \pm 0.060}$	—
	Mod. Gamma-Poisson	0.542 ± 0.049	—	0.521 ± 0.044	—
	Mod. Poisson	0.542 ± 0.038	—	0.502 ± 0.036	—
	SMOTE	0.389 ± 0.081	—	0.372 ± 0.064	—
	Resampling	0.539 ± 0.039	—	0.524 ± 0.043	—
	Unbalanced	0.539 ± 0.044	—	0.500 ± 0.045	—

'5000' setting is reported in Table 4. All the SVM models – except for those trained on SMOTE – achieved comparable balanced accuracy in both in-domain and external validation. Resampling-based augmentation led to the highest average balanced accuracy (57% on TCGA and 89% on CPTAC). Nonetheless, both crossover methods, the modified gamma-Poisson, and modified Poisson models, resulted in accuracies that fell within the confidence interval of this method. Particularly, the inter-class crossover method is associated with a better external performance (88.9%), which is significantly better than the one obtained on the gamma-Poisson (p-value = 0.002), Poisson (p-value = 0.01), SMOTE (p-value = 2.98e-8) and unbalanced (p-value = 0.0008) datasets. Interestingly, SMOTE-derived representations deteriorate in quality as the augmentation class size increases. In the case of "Max" class size, they provide a strong signal resulting in on-par performance with the other models (Supplementary Table 11. As the augmentation class size increases, the performance with SMOTE-augmented data drops significantly (Supplementary Table 11).

With the EBMs, the crossover methods perform better than the other augmentation methods in terms of averaged balanced accuracy. As before, all models but those trained with SMOTE achieve statistically similar accuracies on in-domain TCGA data. On external data, however, both cross-over methods lead to significantly better performance than the rest, showing that these data augmentation methods foster model generalisation. As seen with SVMs, SMOTE-derived models perform poorly as they are trained on larger class sizes(Supplementary Table 11). In addition to reporting the balanced accuracy, we use the global explanation feature of EBMs to understand which features contributed the most to the predictions. The results are discussed in subsubsection A.5.2.

Predicting CIMP: The CIMP status is not only useful in characterising the subtypes of colorectal cancer but also in predicting response to therapy [58, 59]. For example, CMS1 shows a high level of CpG Island Methylator Phenotype (CIMP.High). CMS2 and CMS4 on the other hand are negative for CpG Island Methylator Phenotype (CIMP.Neg), and CMS3 shows a low level of CIMP (CIMP.Low) [46, 60]. We perform similar experiments as with the MSI, and report the balanced accuracy from a 5-fold stratified 5 times repeated cross-validation for the class size '5000' setting in Table 4. As the CIMP status is not available

for patients in the CPTAC cohort, we omit the external validation. Results on the TCGA test set are in favour of the inter-class crossover method, which leads to a significantly better SVM performance than intra-class crossover (p-value=0.006), gamma-Poisson (p-value = 0.0005), Poisson (p-value = 0.0008), SMOTE (p-value = 2.98e-8), resampling (p-value = 0.0006) augmentation methods and unbalanced data (p-value = 2.08e-6). With the EBM, our inter-class crossover augmentation method is significantly better than Poisson (p-value = 0.009), SMOTE (2.98e-8) and unbalanced (0.007). Similar to the findings from the MSI experiment, SMOTE-augmented representations perform the worst. This reinforces our understanding that the good performances attained with SMOTE are limited to minority oversampling, and do not extend to oversampling all classes. Results for CIMP prediction for all class sizes are reported in Supplementary Table 10.

Conclusion

We introduced a novel method for augmenting transcriptomic data using chromosomal crossover, which demands minimal computational resources. When distinct gene signatures represent target phenotypes, our approach allows us to generate synthetic data by sampling values across patients while adhering to specific phenotypic constraints.

Results show that the augmented datasets obtained by our method not only reflect biological variability but also enhance model robustness against overfitting and distribution shifts. The effectiveness of our approach is demonstrated by the higher performances on the classification experiments on TCGA COADREAD, and CPTAC COAD. Our data augmentation approaches, including the modified gamma-Poisson, excel in generalisation and were in most experiments superior to popular methods like SMOTE.

A limitation of the intra- and inter-class crossover approaches is the reliance on non-overlapping gene signatures predictive of each phenotype. This can be overcome, however, also in research areas lacking established signatures, by preliminary feature selection analyses that can identify distinct phenotype-predictive gene sets.

We envision extending our approach to gene signatures with correlated genes, as classifiers can disentangle existing correlations. Although augmenting datasets with correlated classes may introduce challenges, we intend to explore this further in future work. Additionally, we plan to investigate the impact of our method when applied to the entire genome without phenotype conditioning. This entails dividing the genome into chunks and performing crossovers. We expect potential spurious correlations and signal corruption.

Despite these limitations, our sampling methods hold promise for heterogeneity studies, particularly in disease analysis. By cross-mixing samples, we create a landscape of mixed phenotypes, the diversity of which depends on the dataset size, and offers potential applications in unsupervised or semi-supervised learning given fully unlabeled data.

Data and Code availability

The code is hosted at `https://github.com/PaccMann/transcriptomic_signature_sampling`, and includes a link to the augmented datasets for reproducibility.

Author contributions statement

N.J. and M.G. conceived the experiments, N.J. conducted the experiments, N.J created visualisations and summarised results, N.J. and M.G analysed the results, N.J., M.G. and M.R.M wrote and reviewed the manuscript.

Acknowledgments

This work is supported by funds from the Swiss National Science Foundation (Sinergia CRSII5_193832). The results shown here are in whole or part based upon data generated by the TCGA Research Network: `https://www.cancer.gov/tcga`. Data used in this publication were generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH).

References

1. Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.
2. Polina Mamoshina, Armando Vieira, Evgeny Putin, and Alex Zhavoronkov. Applications of deep learning in biomedicine. *Molecular pharmaceutics*, 13(5):1445–1454, 2016.
3. Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of discrimination methods for the classification of tumors using gene expression

data. *Journal of the American statistical association*, 97(457):77–87, 2002.
4. Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
5. Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.
6. Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156, 1997.
7. Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.
8. Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.
9. Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
10. Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
11. Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.
12. Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021.
13. Christopher M Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.
14. Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.
15. John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.
16. Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
17. Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.
18. Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM computing surveys (CSUR)*, 49(2):1–50, 2016.
19. Ozlem Yersal and Sabri Barutca. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World journal of clinical oncology*, 5(3):412, 2014.
20. Nuria Rodriguez-Salas, Gema Dominguez, Rodrigo Barderas, Marta Mendiola, Xabier García-Albéniz, Juan Maurel, and Jaime Feliu Batlle. Clinical relevance of colorectal cancer molecular subtypes. *Critical reviews in oncology/hematology*, 109:9–19, 2017.
21. Xu Zhou, Kai Hu, Peter Bailey, Christoph Springfeld, Susanne Roth, Roma Kurilov, Benedikt Brors, Thomas Gress, Malte Buchholz, Jingyu An, et al. Clinical impact of molecular subtyping of pancreatic cancer. *Frontiers in cell and developmental biology*, 9:743908, 2021.
22. Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.
23. Andrew I Su, John B Welsh, Lisa M Sapinoso, Suzanne G Kern, Petre Dimitrov, Hilmar Lapp, Peter G Schultz, Steven M Powell, Christopher A Moskaluk, Henry F Frierson Jr, et al. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer research*, 61(20):7388–7393, 2001.
24. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
25. Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.
26. Magdalena Kircher, Elisa Chludzinski, Jessica Krepel, Babak Saremi, Andreas Beineke, and Klaus Jung. Augmentation of transcriptomic data for improved classification of patients with respiratory diseases of viral origin. *International journal of molecular sciences*, 23(5):2481, 2022.
27. James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and

differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):1–13, 2010.
28. Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang. Degseq: an r package for identifying differentially expressed genes from rna-seq data. *Bioinformatics*, 26(1):136–138, 2010.
29. Radhika Khetani Meeta Mistry. Gene-level differential expression analysis, May 2020.
30. Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.
31. Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.
32. Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
33. H He, Y Bai, EA Garcia, and S ADASYN Li. adaptive synthetic sampling approach for imbalanced learning. ieee international joint conference on neural networks. In *2008 (IEEE World Congress On Computational Intelligence)*, 2008.
34. Felix Last, Georgios Douzas, and Fernando Bacao. Oversampling for imbalanced learning based on k-means and smote. *arXiv preprint arXiv:1711.00837*, 2017.
35. Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.
36. Qikang Wan, Xiongshi Deng, Min Li, and Haotian Yang. Sddsmote: Synthetic minority oversampling technique based on sample density distribution for enhanced classification on imbalanced microarray data. In *2022 The 6th International Conference on Compute and Data Analysis*, pages 35–42, 2022.
37. Junnan Li, Qingsheng Zhu, Quanwang Wu, and Zhu Fan. A novel oversampling technique for class-imbalanced learning based on smote and natural neighbors. *Information Sciences*, 565:438–455, 2021.
38. Rok Blagus and Lara Lusa. Smote for high-dimensional class-imbalanced data. *BMC bioinformatics*, 14(1):1–16, 2013.
39. Hai-Hui Huang, Hao Rao, Rui Miao, and Yong Liang. A novel meta-analysis based on data augmentation and elastic data shared lasso regularization for gene expression. *BMC bioinformatics*, 23(10):1–24, 2022.
40. Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7(1):1–12, 2006.
41. Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.
42. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
43. Ramon Viñas, Helena Andrés-Terré, Pietro Liò, and Kevin Bryson. Adversarial generation of gene expression data. *Bioinformatics*, 38(3):730–737, 2022.
44. Poonam Chaudhari, Himanshu Agrawal, and Ketan Kotecha. Data augmentation using mg-gan for improved cancer classification on gene expression data. *Soft Computing*, 24(15):11381–11391, 2020.
45. Mohamed Marouf, Pierre Machart, Vikas Bansal, Christoph Kilian, Daniel S Magruder, Christian F Krebs, and Stefan Bonn. Realistic in silico generation and augmentation of single-cell rna-seq data using generative adversarial networks. *Nature communications*, 11(1):1–12, 2020.
46. Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien De Reynies, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, et al. The consensus molecular subtypes of colorectal cancer. *Nature medicine*, 21(11):1350–1356, 2015.
47. Steven A Buechler, Melissa T Stephens, Amanda B Hummon, Katelyn Ludwig, Emily Cannon, Tonia C Carter, Jeffrey Resnick, Yesim Gökmen-Polar, and Sunil S Badve. Colotype: a forty gene signature for consensus molecular subtyping of colorectal cancer tumors using whole-genome assay or targeted rna-sequencing.

Scientific reports, 10(1):1–13, 2020.

48. Major Greenwood and G Udny Yule. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal statistical society*, 83(2):255–279, 1920.
49. Tim Barry. Tim barry: Gamma, poisson, and negative binomial distributions, 2020.
50. Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.
51. Nathan J Edwards, Mauricio Oberti, Ratna R Thangudu, Shuang Cai, Peter B McGarvey, Shine Jacob, Subha Madhavan, and Karen A Ketchum. The cptac data portal: a resource for cancer proteomics research. *Journal of proteome research*, 14(6):2707–2713, 2015.
52. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.
53. Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
54. Sanghee Kang, Younghyun Na, Sung Yup Joung, Sun Il Lee, Sang Cheul Oh, and Byung Wook Min. The significance of microsatellite instability in colorectal cancer after controlling for clinicopathological factors. *Medicine*, 97(9), 2018.
55. S Popat, R Hubner, and RS Houlston. Systematic review of microsatellite instability and colorectal cancer prognosis. *Journal of clinical oncology*, 23(3):609–618, 2005.
56. Zhenli Diao, Yanxi Han, Yuqing Chen, Rui Zhang, and Jinming Li. The clinical utility of microsatellite instability in colorectal cancer. *Critical reviews in oncology/hematology*, 157:103171, 2021.
57. TO Sillo, AD Beggs, DG Morton, and G Middleton. Mechanisms of immunogenicity in colorectal cancer. *Journal of British Surgery*, 106(10):1283–1297, 2019.
58. Byung-Hoon Min, Jeong Mo Bae, Eui Jin Lee, Hong Suk Yu, Young-Ho Kim, Dong Kyung Chang, Hee Cheol Kim, Cheol Keun Park, Suk-Hee Lee, Kyoung-Mee Kim, et al. The cpg island methylator phenotype may confer a survival benefit in patients with stage ii or iii colorectal carcinomas receiving fluoropyrimidine-based adjuvant chemotherapy. *BMC cancer*, 11(1):1–10, 2011.
59. Stacey Shiovitz, Monica M Bertagnolli, Lindsay A Renfro, Eunmi Nam, Nathan R Foster, Slavomir Dzieciatkowski, Yanxin Luo, Victoria Valinluck Lao, Raymond J Monnat Jr, Mary J Emond, et al. Cpg island methylator phenotype is associated with response to adjuvant irinotecan-based therapy for stage iii colon cancer. *Gastroenterology*, 147(3):637–645, 2014.
60. Guillermo Valenzuela, Joaquín Canepa, Carolina Simonetti, Loreto Solo de Zaldívar, Katherine Marcelain, and Jaime González-Montero. Consensus molecular subtypes of colorectal cancer in clinical practice: A translational approach. *World journal of clinical oncology*, 12(11):1000, 2021.
61. Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.
62. Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1–19, 2016.
63. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
64. Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.
65. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
66. Rachel V Purcell, Sebastian Schmeier, Yee Chen Lau, John F Pearson, and Francis A Frizelle. Molecular subtyping improves prognostication of stage 2 colorectal cancer. *BMC cancer*, 19(1):1–9, 2019.
67. Moritz Herrmann, Philipp Probst, Roman Hornung, Vindi Jurinovic, and Anne-Laure Boulesteix. Large-scale benchmark study of survival prediction methods using multi-omics

data. *Briefings in bioinformatics*, 22(3):bbaa167, 2021.

Supplementary

Visualising augmented data for other class sizes

Figure 4 and Figure 5 illustrate the UMAP projected distributions of the datasets augmented by various methods to class sizes "Max" and 500 respectively. These visualisations are compared to the UMAP projection of the unbalanced dataset to ascertain visually if global and local trends are fairly captured. We see trends similar to that observed for class size 5000, although less distinctly. The crossover and gamma-Poisson augmentation methods appear to mimic the distribution of the unbalanced data, whereas the Poisson method creates well separated clusters which could possibly lead to spurious inference on possible shared subtypes. The SMOTE augmentation method captures the distribution well in the "Max" sampling case, that is, minority oversampling case for which it was designed. The distribution of the unbalanced data is still reflected when the class size is increased to 500 per class, however the augmented data tend to become sparsely distributed as the class size increases due to the interpolation between reference samples (section 3, Figure 3).

Distribution size effect

In this experiment, we test various reference set sizes to understand its impact on CMS classification performance. The reference set sizes considered are $r = 3, 5, 10, 25, 50$. We choose $class_size = 5000$ so that there are enough newly sampled observations to draw a conclusion about the effect of reference set size. The study is conducted on a classification task with various classifiers and the 5-fold cross validation balanced accuracy is reported in **??**. The reference set size with the highest accuracy on average, $r = 5$, is chosen for the gamma-Poisson sampling method. Figure 6 illustrates what the distribution looks like with different reference set sizes. We observe that for $r = 5$, the real and synthetic samples are well interspersed with each other showing qualitatively that the distribution of synthetic samples is similar to that of the real samples. As r increases, new samples are created that seemingly fall away from its member class resulting in new sub-clusters as seen in Figure 6 c and d, for CMS1 and CMS2 classes.

For the Poisson distribution, we double the reference set size to $r = 10$ such that the variance is more stable and closer to the mean. Figure 7 illustrates the Poisson augmented data at

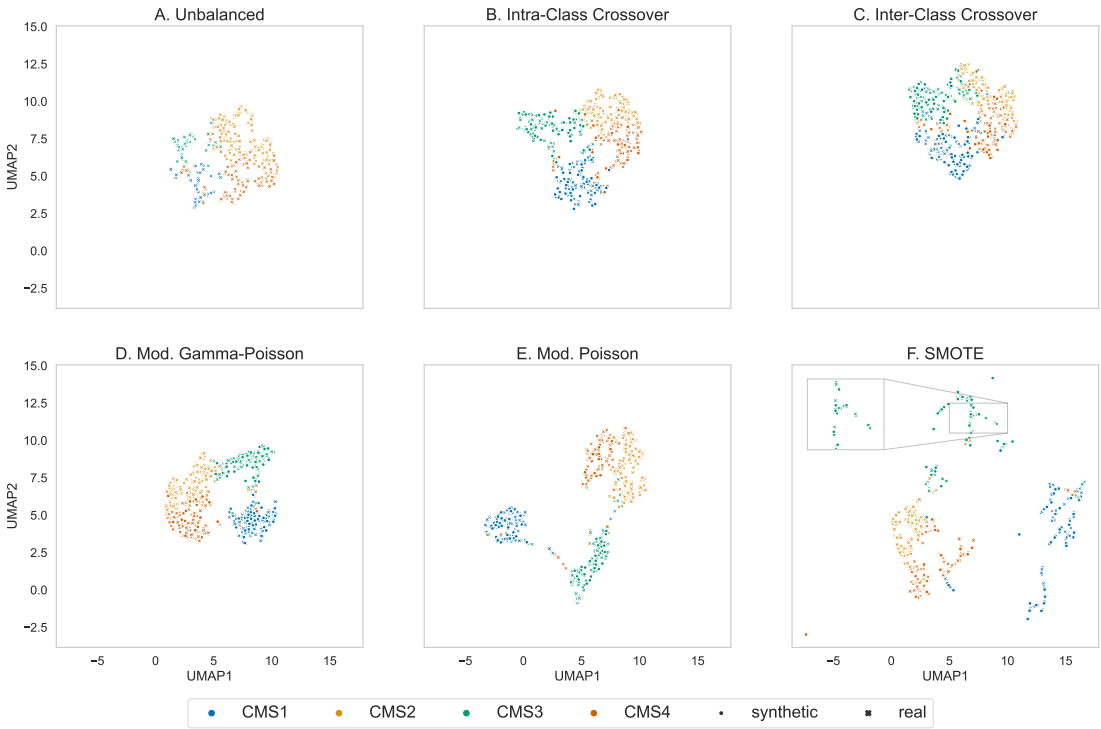

Fig. 4. UMAP [53] visualisations of datasets in the study. A. Unbalanced dataset. B. Intra-class crossover sampled dataset. C. Inter-class crossover sampled dataset. D. Gamma-Poisson sampled dataset. E. Poisson sampled dataset. F. SMOTE sampled dataset. Only minority classes are augmented to match the majority class size.

different reference set sizes. Here, we make a similar observation that as r increases, sub-clusters begin to form due to synthetic samples being generated well outside its member class distribution. On comparing the Poisson augmented data to the gamma-Poisson augmented data, we observe that the CMS clusters appear to be well separated with the Poisson. While this may appear to be beneficial for certain tasks that are strictly associated with the these types, it does not capture the distribution of the real data well (Figure 3a). This could introduce spurious associations and attributions.

While we acknowledge that conclusions drawn from UMAP visualisations have to be taken with a grain of salt, we note that all UMAPs were generated with the same hyperparameter choice, differing only in their input data. This gives us a qualitative insight into what these differently augmented datasets with various reference set sizes look like and whether or not the newly generated data improve the support of a given class.

We omit results for $r = 3$ since it has high failure rates. Since the parametric distributions depend on the mean and variance of the subsample, it can sometimes happen that when $r = 3$ the variance associated with the subsample is too small, resulting in NaNs being sampled from the gamma distribution due to an invalid argument (negative integers or 0 possibly due to a possible overflow error).

Data Augmentation Process

We compare our two crossover sampling methods against four different data augmentation strategies - (i) Modified gamma-Poisson sampling, (ii) Modified Poisson sampling, (iii) Synthetic Minority Oversampling Technique (SMOTE) [61], and (iv) Random oversampling (bootstrapping with replacement). The train and test splits are created from the TCGA COADREAD gene

Fig. 5. UMAP [53] visualisations of datasets in the study. A. Unbalanced dataset. B. Intra-class crossover sampled dataset. C. Inter-class crossover sampled dataset. D. Gamma-Poisson sampled dataset. E. Poisson sampled dataset. F. SMOTE sampled dataset. All augmented datasets have 500 samples per class.

Table 5. Effect of reference sample size on classifier performance for parametric methods. The table shows 5-fold cross-validation results for different r sizes and classifiers. Majority of the best performances belong to $r = 5$, which also has the best performance on average.

Model	r=5	r=10	r=25	r=50
KNN	0.857 ± 0.035	0.836 ± 0.027	0.836 ± 0.036	$\mathbf{0.858 \pm 0.029}$
LR	$\mathbf{0.889 \pm 0.031}$	0.881 ± 0.040	0.886 ± 0.033	0.888 ± 0.049
SVM	0.886 ± 0.027	$\mathbf{0.893 \pm 0.033}$	0.889 ± 0.027	0.887 ± 0.036
RF	$\mathbf{0.881 \pm 0.029}$	0.865 ± 0.036	0.857 ± 0.037	0.859 ± 0.035
EBM	$\mathbf{0.905 \pm 0.011}$	0.892 ± 0.028	0.881 ± 0.032	0.882 ± 0.030
MLP	0.881 ± 0.020	0.889 ± 0.040	$\mathbf{0.889 \pm 0.025}$	0.883 ± 0.038

expression count data. Each training data split is augmented by the aforementioned methods, followed by FPKM normalisation (Fragment Per Kilobase per Million mapped reads) and log transformation [62]. All datasets are standardised based on the training split. We consider 3 class sizes for augmentation - Max, 500 and 5000. In the 'Max' setting, only the minority classes are oversampled to match the size of the majority class. For class size 500 or 5000, all classes are augmented until this size is reached for each class. In addition to testing the generalisability of the augmentation methods, the different sizes also provide insights into the effect of size on the quality of the augmented data. We also include the unaugmented log-FPKM normalised gene expression data from TCGA COADREAD as a baseline. This dataset is referred to as the "unbalanced" dataset

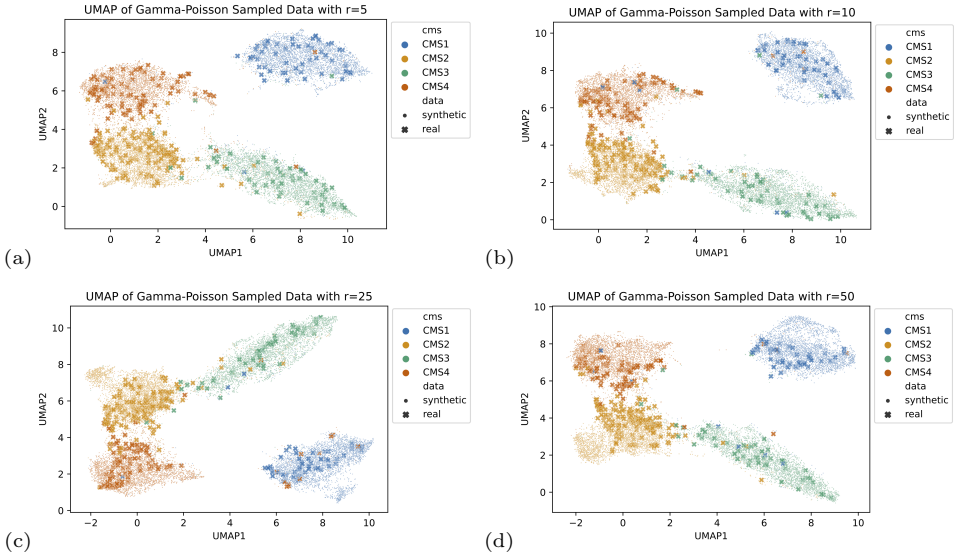

Fig. 6. UMAP [53] visualisations of gamma-Poisson augmented datasets in the study. (a)r=5. (b) r=10. (c) r=25. (d) r=50. All augmented datasets have 5000 samples per class. When r is set to 5, the new samples are more interspersed with the real data, as opposed to r=50, where new samples are generated "out of distribution" resulting in new clusters.

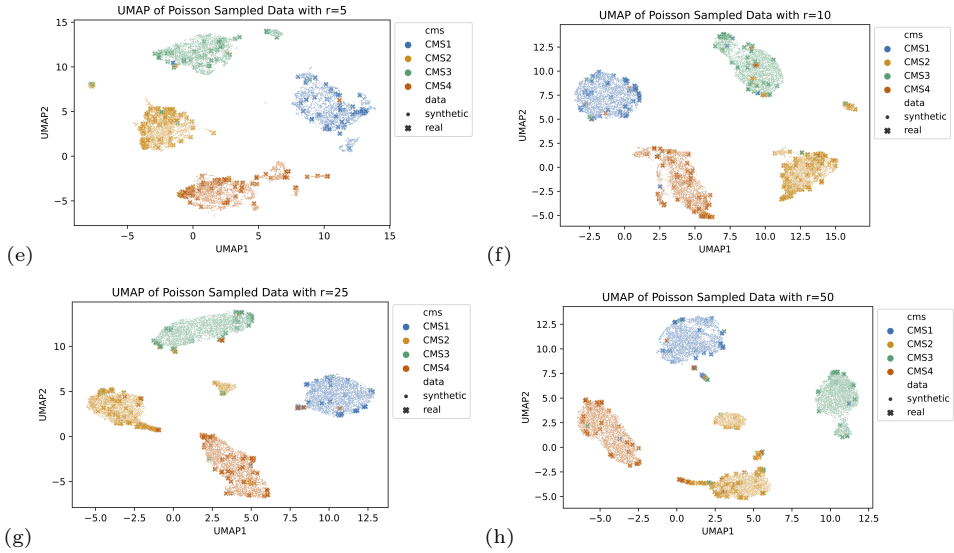

Fig. 7. UMAP [53] visualisations of Poisson augmented datasets in the study. (a)r=5. (b) r=10. (c) r=25. (d) r=50. All augmented datasets have 5000 samples per class. For r=10, the new samples appear better interspersed with the real data and there are fewer outlier clusters.

henceforth. All datasets are made publicly available through Box.

Discriminative Modelling

To demonstrate the utility of our augmentation methods in discriminative modelling, we make use of 6 classifiers, namely, Logistic Regression (LR), K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Explainable Boosting Machines (EBM), Random Forests (RF) and Multi-Layer Perceptron (MLP). For the standard classifiers, we make use of the `scikit-learn` [63] library implementations with the default parameters. We only increase the `max_iter` to 5000 for the Logistic Regression model. For the EBMs, we use the `InterpretML` library. We set a seed (42) for all models.

Complex non-linear models typically require a large amount of data to learn robust decision boundaries because they are easy to overfit. Generally, a regularisation method such as data augmentation is employed to counter model overfitting, where the quality of the augmented data plays a crucial role. For this reason we also include a Multi-Layer Perceptron (MLP) to predict the CMS types. For the MLP, we use our implementation in `PyTorch`. We construct a single layer network with 20 hidden neurons and apply the `ReLu` [64] activation followed by an output layer with 4 neurons, each predicting a probability for a CMS type.. The Cross-Entropy objective is optimised by `Adam`. The batch size and learning rate are adapted according to the size of the dataset as shown in Table 6. We apply `l2` regularisation with a rate of 0.01. The number of epochs are set to 200 with an early stopping wait of 5 steps. The seed is set to 42.

We report the mean balanced accuracy (Equation 7) on the TCGA standardised test sets and the CPTAC external set. The metric is computed using true positive (TP), true negative (TN), false positive (FP) and false negative (FN) counts and is averaged across the 25 splits (5-fold CV repeated 5 times). The CPTAC dataset is standardised according to the training data mean and standard deviation from each split after which it is passed to the trained model from that split. We drop patients for whom the target label is not available. Additionally, we perform significance testing with the Wilcoxon test. All experiments are run locally on a 2019 MacBook Pro.

$$balanced-accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (7)$$

In Table 7 we report the mean balanced accuracy for the Logistic Regression and K-Nearest Neighbour models. With the linear logistic regression, we observe that gamma-Poisson augmented data provide a better support than the other datasets as evidenced by a high balanced accuracy on the TCGA test data (~ 0.894). Similarly, the inter-class crossover method provides better support for generalisation compared to the the other augmentation methods. These results are in line with the general observations made with the other classifiers. Interestingly, with the KNN, the Poisson and SMOTE methods perform better in-domain. With the Poisson, this can be attributed to the way the classes cluster around a mean representation and are separated from each other. This would automatically assign that class label to any sample in that cluster. This may not be the best strategy when dealing with mixed subtypes as evidenced by the lower best accuracy (0.872) compared to that achieved by a linear model (0.894). Similarly, SMOTE generates new samples by interpolation, which means all the gaps in the input space are filled giving it an advantage when classifying with methods that assigns labels based on the neighbourhood. Again, the best accuracy with the KNN method is much lower than what is achieved with other classification methods, suggesting that neighbourhood based methods are not the right strategy when dealing with potential mixing of subtypes. For results of the other studied classifiers, please refer to Table 2.

Table 6. Batch size and learning rate values are adapted according to the size of the datasets. The batch size is reduced for smaller datasets and increased for larger datasets.

Class-Size	Batch Size	Learning Rate
Unbalanced	64	0.001
Max	64	0.001
500	128	0.001
5000	128	0.0001

Generative Modelling

In this task, we train a Variational Autoencoder (VAE). The encoder and decoder of the VAE are single-layer neural networks with 20 hidden units activated by the `ReLU` function [64]. The dimension of the latent representation is 4. The batch size and learning rate are adapted according to the size of the dataset. For the unbalanced dataset and for the augmented datasets with class sizes 'Max' and '500', we set the batch size to 64 and learning rate to 0.001. For all the other datasets, we set the batch size to 128 and learning rate to 0.0001. The models are trained on the sum of the L2 loss and KL-divergence. An L2

Table 7. 5x5 cross-validation results for supervised classification (CMS prediction) on augmented data with various classifiers. Each classifier is trained with data augmented by different sampling methods. Each class is augmented to a different class-size - Max, 500, and 5000. The TCGA test set indicates in-domain performance. The CPTAC dataset serves as external validation. We report the average balanced accuracy scores and their standard deviation, with the best highlighted in bold. Although in-domain peformances vary depending on the classifier and class-size, it is clear that models trained with the inter-class crossover augmented data generalise the best as the class-size increases.

Class-Size	Sampling Methods	Logistic Regression		KNN		MLP	
		TCGA	CPTAC	TCGA	CPTAC	TCGA	CPTAC
Max	Intra-Crossover	0.890 ± 0.031	0.515 ± 0.047	0.843 ± 0.035	0.590 ± 0.029	$\mathbf{0.899 \pm 0.048}$	0.552 ± 0.054
	Inter-Crossover	0.887 ± 0.026	0.517 ± 0.045	0.848 ± 0.038	0.596 ± 0.032	0.889 ± 0.063	0.595 ± 0.058
	Gamma-Poisson	$\mathbf{0.894 \pm 0.026}$	0.530 ± 0.041	0.840 ± 0.039	0.581 ± 0.034	0.885 ± 0.056	0.604 ± 0.027
	Poisson	0.887 ± 0.029	0.538 ± 0.040	$\mathbf{0.872 \pm 0.041}$	$\mathbf{0.620 \pm 0.028}$	0.879 ± 0.073	0.601 ± 0.034
	SMOTE	0.891 ± 0.034	0.540 ± 0.031	0.862 ± 0.044	0.591 ± 0.031	0.899 ± 0.054	$\mathbf{0.619 \pm 0.016}$
	Resampling	0.883 ± 0.031	$\mathbf{0.551 \pm 0.037}$	0.848 ± 0.034	0.578 ± 0.037	0.877 ± 0.074	0.591 ± 0.037
500	Intra-Crossover	0.874 ± 0.024	0.541 ± 0.052	0.850 ± 0.038	0.683 ± 0.029	0.885 ± 0.031	0.530 ± 0.041
	Inter-Crossover	0.875 ± 0.031	$\mathbf{0.616 \pm 0.045}$	0.610 ± 0.035	$\mathbf{0.708 \pm 0.023}$	0.884 ± 0.031	$\mathbf{0.650 \pm 0.021}$
	Gamma-Poisson	$\mathbf{0.894 \pm 0.024}$	0.567 ± 0.047	0.841 ± 0.038	0.588 ± 0.034	0.897 ± 0.020	0.609 ± 0.033
	Poisson	0.885 ± 0.022	0.532 ± 0.039	0.864 ± 0.036	0.636 ± 0.023	0.888 ± 0.032	0.604 ± 0.042
	SMOTE	0.887 ± 0.030	0.542 ± 0.033	$\mathbf{0.870 \pm 0.034}$	0.598 ± 0.026	$\mathbf{0.901 \pm 0.023}$	0.553 ± 0.019
	Resampling	0.879 ± 0.029	0.552 ± 0.041	0.814 ± 0.042	0.562 ± 0.037	0.900 ± 0.018	0.546 ± 0.031
5000	Intra-Crossover	0.863 ± 0.031	0.591 ± 0.043	0.853 ± 0.038	0.634 ± 0.036	0.877 ± 0.021	0.607 ± 0.037
	Inter-Crossover	0.867 ± 0.029	$\mathbf{0.652 \pm 0.020}$	0.844 ± 0.038	$\mathbf{0.685 \pm 0.020}$	0.871 ± 0.028	$\mathbf{0.663 \pm 0.009}$
	Gamma-Poisson	$\mathbf{0.895 \pm 0.027}$	0.627 ± 0.022	0.852 ± 0.037	0.624 ± 0.024	0.894 ± 0.023	0.627 ± 0.017
	Poisson	0.891 ± 0.028	0.535 ± 0.042	0.861 ± 0.045	0.642 ± 0.014	0.877 ± 0.033	0.663 ± 0.017
	SMOTE	0.889 ± 0.029	0.540 ± 0.036	$\mathbf{0.870 \pm 0.033}$	0.603 ± 0.028	0.883 ± 0.031	0.634 ± 0.043
	Resampling	0.868 ± 0.035	0.562 ± 0.053	0.803 ± 0.047	0.575 ± 0.030	$\mathbf{0.897 \pm 0.029}$	0.531 ± 0.030
	Unbalanced	0.889 ± 0.033	0.535 ± 0.034	0.841 ± 0.036	0.592 ± 0.026	0.898 ± 0.043	0.606 ± 0.027

penalty of 0.0001 is applied to all models, which are optimised by Adam [65] and trained for 300 epochs with an early stopping wait time of 8 steps. All experiments are run locally on a 2019 MacBook Pro. We report the latent cluster purity score Equation 8 on the test set for each sampling method and sampling averaged across the 5 repetitions. This score is computed by running a KMeans [63] algorithm on the latent embeddings of the test set, computing the cluster identity followed by computing the accuracy score. In addition to evaluating on the test set, we also check the generalisation power of the datasets by predicting CMS on unseen CPTAC COAD (Colon Adenocarcinoma) [51] gene expression data.

The cluster purity equation is formalised as:

$$purity(A, B) = \frac{1}{N} \sum_{k} \max_{j} |a_k \cap b_j| \tag{8}$$

where, A is the set of all clusters, A $= \mathrm{a}_1, ..., a_k, andBisthesetsetofallclasses, B = b_1, ..., b_j$.

Survival Analysis

In this task, we investigate whether an increase in sample diversity through data augmentation can enhance the predictive accuracy for survival outcomes across other CMS types. The latent representations of real data, derived from models trained on differently augmented data, are fed into the `scikit-survival` Cox Proportional Hazards model [] which predicts a risk score for each patient. For each augmentation method, we have 25 splits since we do a 5-fold stratified cross-validation repeated 5 times. This means there are 25 different models for each augmentation method and the unbalanced dataset. We perform training and testing with the same splits. First, we retrieve the latent representations of only the real train and test data. Next, we drop all those samples for which survival information, namely, 'Overall Survival' (OS) and 'Overall Survival Time' (OS.time), are not available. We then fit a `CoxPHSurvivalanalysis` model on the training data, and predict on the test data and external CPTAC data. This is repeated for every split of every augmentation method after which we compute the mean concordance-index of the predictions. A point to note here is that during testing, the model encounters singular matrices sometimes and hence fails. The splits that fail are excluded from the metric aggregation.

We use the concordance-index (C-index) to measure the alignment between the predicted risk

Table 8. 5x5 cross-validation results of survival prediction with the Cox Proportional Hazards model. Concordance index (C-index) values are shown. The input to the model is the 4D latent representation learnt by the VAEs trained with differently augmented data. We report the avergae concordence-index and its standard deviation. The best scores are highlighted in bold. The resampling method produces better representations for in-domain survival prediction whereas the inter-class crossover method produces better representations for generalisation.

Sampling Methods	Max		500		5000	
	TCGA	CPTAC	TCGA	CPTAC	TCGA	CPTAC
Intra-Crossover (Ours)	0.525 ± 0.056	0.635 ± 0.087	0.526 ± 0.057	0.645 ± 0.097	0.497 ± 0.072	0.716 ± 0.082
Inter-Crossover (Ours)	0.519 ± 0.054	0.677 ± 0.086	0.523 ± 0.068	$\mathbf{0.689 \pm 0.081}$	0.507 ± 0.057	$\mathbf{0.727 \pm 0.070}$
Mod. Gamma-Poisson	0.536 ± 0.062	0.674 ± 0.087	0.532 ± 0.048	0.639 ± 0.081	0.543 ± 0.069	0.671 ± 0.069
Mod. Poisson	0.530 ± 0.064	0.681 ± 0.093	0.537 ± 0.052	0.662 ± 0.069	0.522 ± 0.067	0.652 ± 0.070
SMOTE	$\mathbf{0.540 \pm 0.065}$	0.639 ± 0.080	0.534 ± 0.072	0.684 ± 0.085	0.497 ± 0.070	0.501 ± 0.141
Resampling	0.536 ± 0.065	$\mathbf{0.692 \pm 0.071}$	$\mathbf{0.556 \pm 0.061}$	0.645 ± 0.066	$\mathbf{0.545 \pm 0.061}$	0.643 ± 0.063
Unbalanced			0.529 ± 0.075	0.639 ± 0.081		

scores and the true event order (here, death) of the patients. These scores are reported in Table 8, where we observe that the inter-class crossover method, despite leading to a lower in-domain performance, generalises better, i.e. leads to better performance in the external dataset than the other augmentation methods at larger class sizes. The intra-class crossover method shows a similar trend although the decrease in C-index is much larger than for the inter-class method. The resampling method, on the other hand, is associated with a better in-domain performance, but a poorer performance on the external dataset as the class size increases. The gamma-Poisson sampling method provides the best trade-off between performance on in-domain and external data. Although its C-index scores are not the best for either dataset, the method provides a significant improvement over unbalanced data for both. Finally, the SMOTE method works well when the augmentation is done only in the minority classes. Despite having the best in-domain C-index score in the 'Max' class size setting, its generalisation is on par with that achieved by unbalanced data. Increasing the class size thereby augmenting even the majority class does not do the method any favours as seen by its deterioration in performance on both in-domain and external data.

In general, CMS has been found to be a poor predictive of survival [66]. Only CMS4 has a significantly different survival curve compared to other CMS types [46]. This is reflected in our results – for class size 500, the resampling method is the only method that performs significantly better than unbalanced data (p-value = 0.04), whereas for the "Max" and 5000 class sizes, none of the methods are significantly different in performance compared to the unbalanced case. With external testing, our inter-class crossover method performs significantly better than the unbalanced data at class-sizes 500 and 5000 (p-value = 0.01 and 0.0002 respectively). Additionally, due to the limited size of the latent space, low C-index scores are anticipated. For context, a multiomics model trained with BlockForests produces the best C-index score 0f 0.656 [67]. To truly examine the utility of our learned latent representations, we look into predicting other clinical variables associated with CMS (Section 5.4, 5.4).

Predicting Clinical Variables

For the classification tasks of predicting MSI and CIMP status, we use the SVM-RBF and EBM models from `scikit-learn` and `InterpretML` to run our experiments (with the default parameters and seed 42). Here the input to the models is a 4D latent representation extracted from the VAE for each patient. Similar to the survival analysis task, we perform training and testing with the same 5x5 stratified cross-validation splits, and all samples for which the MSI/CIMP status is not available are dropped for the MSI/CIMP prediction task respectively. The patient summary counts for both MSI and CIMP are reported in Table 9.

The MSI status prediction task has 3 classes - MSI-H(High), MSI-L(Low) and MSS(Stable). The mean balanced accuracy scores for TCGA test set are reported based on these 3 classes. However, the MSI status associated with the CPTAC dataset has only two classes - MSS-H and MSS. The mean balanced accuracy scores for CPTAC are based only on these two classes despite being trained on data containing all 3 classes. The results of the experiments for all three class sizes are reported in Table 11.

Table 9. Summary counts of patients with MSI and CIMP labels available in the TCGA COADREAD and CPTAC COAD gene expression datasets.

Clinical Variable	Categories	TCGA	CPTAC
MSI	MSI-H	80	24
	MSI-L	74	-
	MSS	339	81
	Total	493	105
CIMP	CIMP.High	66	-
	CIMP.Low	86	-
	CIMP.Neg	272	-
	Total	364	-

In addition to reporting the balanced accuracy, we generate global explanations from an EBM model trained on the best performing split of inter-class crossover augmented data for class-size '5000'. These explanations are in the form of bar charts as seen in Figure 8a which is a depiction of how much each latent feature contributes to the predictions. To identify which genes are behind each latent representation, we train another Explainable Boosting Machine on the regression task of predicting each latent feature from the 40-gene signature (Figure 8b-c shows the feature importance scores for the top two features). We compare these explanations with that of an EBM trained on latent representations learned by a VAE trained on unaugmented data Figure 8d-f and an EMB trained directly on the 40-gene signature to predict MSI (Figure 8g).

When the model is given the 40-gene signatures, it tends to assign higher feature importances to most of the CMS1 genes, which are known to be MSI-High, and a few CMS2 genes, which are associated with MSS. When trained on the 4D latent representations, the model assigns the highest importance to feature 4, followed by feature 1 for both settings - inter-class crossover augmented and unaugmented. On inspecting these features further, we find that the latent representations from inter-class crossover augmented data include genes from all four CMS with feature importance score ¿ 0.05 as shown in Figure 8b-c. On the other hand, latent representations from unaugmented data place a higher importance on CMS1 genes and its interaction with other genes, with a lot of overlap between the top two latent features suggesting that these features are not disentangled.

The CIMP prediction task on the other hand also has three classes - CIMP.Neg (Negative), CIMP.Low, and CIMP.High. Since there is no data about the CIMP status for patients in the CPTAC cohort, we only report the mean balanced accuracy over the TCGA test splits in Table 10.

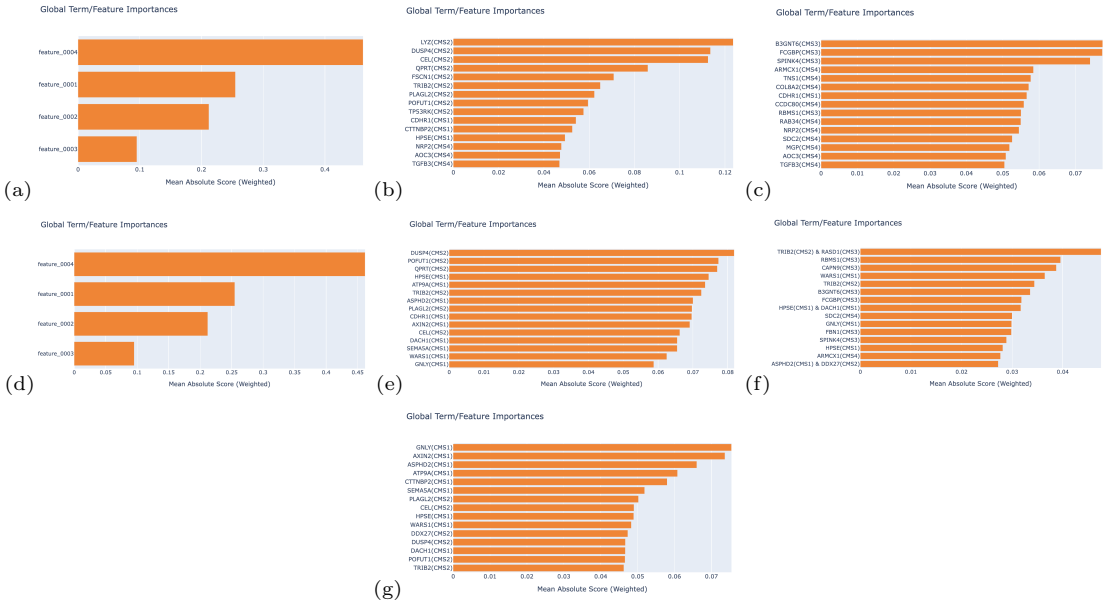

Fig. 8. Global explanations generated by EBM showing feature importance scores in predicting the MSI status. Feature importance scores as determined by an EBM model trained on a. 4D VAE latent representations from inter-class crossover data augmented to size 5000, b. 40-gene signatures to predict latent feature 4 learned from this data, c. 40-gene signatures to predict latent feature 1 learned from this data, d. 4D VAE latent representations from unaugmented data, e. 40-gene signatures to predict latent feature 4 learned from unaugmented data, f. 40-gene signatures to predict latent feature 1 learned from unaugmented data g. 40-gene signatures.

Table 10. 5x5 cross-validation results for CIMP Classification from VAEs trained with data augmented by different sampling methods. We report the averaged balanced accuracy for all class-sizes - Max, 500 and 5000. The best scores are highlighted in bold. The inter-class crossover method dominates in in-domain testing with a significant difference over all methods for the SVM model. With the EBM, the Poisson method outperforms in the class-size Max and 500 settings, with the inter-class method closely following in second place. For class-size 5000, the inter-class method is significantly better than Poisson, SMOTE and unbalanced (p-value ≤ 0.05).

Class-Size	Sampling Methods	SVM-RBF		EBM	
		TCGA	CPTAC	TCGA	CPTAC
Max	Intra-Crossover	0.540 ± 0.046	—	0.505 ± 0.048	—
	Inter-Crossover	$\mathbf{0.558 \pm 0.042}$	—	0.526 ± 0.038	—
	Gamma-Poisson	0.542 ± 0.035	—	0.493 ± 0.049	—
	Poisson	0.556 ± 0.036	—	$\mathbf{0.530 \pm 0.049}$	—
	SMOTE	0.546 ± 0.044	—	0.517 ± 0.052	—
	Resampling	0.542 ± 0.043	—	0.519 ± 0.057	—
500	Intra-Crossover	0.555 ± 0.045	—	0.526 ± 0.044	—
	Inter-Crossover	$\mathbf{0.556 \pm 0.055}$	—	0.527 ± 0.056	—
	Gamma-Poisson	0.526 ± 0.049	—	0.506 ± 0.044	—
	Poisson	0.551 ± 0.056	—	$\mathbf{0.536 \pm 0.057}$	—
	SMOTE	0.539 ± 0.057	—	0.524 ± 0.042	—
	Resampling	0.555 ± 0.043	—	0.528 ± 0.043	—
5000	Intra-Crossover	0.552 ± 0.052	—	0.528 ± 0.043	—
	Inter-Crossover	$\mathbf{0.576 \pm 0.043}$	—	$\mathbf{0.538 \pm 0.060}$	—
	Gamma-Poisson	0.542 ± 0.049	—	0.521 ± 0.044	—
	Poisson	0.542 ± 0.038	—	0.502 ± 0.036	—
	SMOTE	0.389 ± 0.081	—	0.372 ± 0.064	—
	Resampling	0.539 ± 0.039	—	0.524 ± 0.043	—
	Unbalanced	0.539 ± 0.044	—	0.500 ± 0.045	—

Table 11. 5x5 cross-validation results for MSI Classification from VAEs trained with data augmented by different sampling methods. We report the averaged balanced accuracy for class-size 5000. The best scores are highlighted in bold. With SVMs for in-domain testing, none of the methods provide significant improvement over a model trained on unbalanced data. For external validation, however, SVMs trained with the intra- and inter-crossover augmented data are significantly better than the SVM trained on unbalanced data (p-value = 0.003 and 0.01 respectively) for class-size Max and 500. In the 5000 setting, the resampling method leads to the best performance in MSI prediction with SVM, followed by the crossover methods for both in-domain (both not significantly better than unbalanced or each other) and external testing (both significantly better than unbalanced but not each other). With EBMs, the gamma-Poisson method leads to the best in-domain performance significantly better than the SVM trained with unbalanced data in both Max and 500 class-size settings (p-value =). However, the intra-class crossover method is associated with a better performance on external data, with the inter-class crossover method closely following behind. As the class-size is increased to 5000, SVMs trained with both the crossover methods outperform other models in-domain to be the top two performing, and display the best performance with external data, being significantly better than unbalanced data (p-value ≤ 0.05).

Class-Size	Sampling Methods	SVM-RBF		EBM	
		TCGA	CPTAC	TCGA	CPTAC
Max	Intra-Crossover	0.557 ± 0.046	0.858 ± 0.060	0.546 ± 0.044	$\mathbf{0.822 \pm 0.054}$
	Inter-Crossover	0.554 ± 0.036	$\mathbf{0.871 \pm 0.049}$	0.531 ± 0.039	0.811 ± 0.052
	Gamma-Poisson	0.561 ± 0.039	0.839 ± 0.057	$\mathbf{0.549 \pm 0.042}$	0.784 ± 0.053
	Poisson	$\mathbf{0.565 \pm 0.033}$	0.850 ± 0.068	0.539 ± 0.047	0.794 ± 0.068
	SMOTE	0.564 ± 0.031	0.849 ± 0.056	0.547 ± 0.044	0.786 ± 0.072
	Resampling	0.561 ± 0.038	0.831 ± 0.060	0.544 ± 0.045	0.793 ± 0.073
500	Intra-Crossover	0.573 ± 0.035	$\mathbf{0.889 \pm 0.058}$	0.551 ± 0.037	$\mathbf{0.839 \pm 0.059}$
	Inter-Crossover	0.567 ± 0.045	0.872 ± 0.050	0.537 ± 0.041	0.817 ± 0.062
	Gamma-Poisson	0.566 ± 0.028	0.843 ± 0.063	$\mathbf{0.551 \pm 0.025}$	0.801 ± 0.073
	Poisson	0.577 ± 0.035	0.861 ± 0.051	0.550 ± 0.042	0.796 ± 0.089
	SMOTE	0.561 ± 0.041	0.841 ± 0.061	0.538 ± 0.047	0.804 ± 0.066
	Resampling	$\mathbf{0.577 \pm 0.034}$	0.852 ± 0.058	0.550 ± 0.042	0.803 ± 0.063
5000	Intra-Crossover	0.570 ± 0.034	0.884 ± 0.046	$\mathbf{0.555 \pm 0.047}$	0.863 ± 0.047
	Inter-Crossover	0.569 ± 0.041	0.889 ± 0.044	0.552 ± 0.042	$\mathbf{0.867 \pm 0.042}$
	Gamma-Poisson	0.566 ± 0.040	0.845 ± 0.051	0.550 ± 0.039	0.761 ± 0.067
	Poisson	0.557 ± 0.032	0.848 ± 0.062	0.542 ± 0.049	0.794 ± 0.051
	SMOTE	0.390 ± 0.088	0.578 ± 0.130	0.373 ± 0.067	0.558 ± 0.109
	Resampling	$\mathbf{0.578 \pm 0.033}$	$\mathbf{0.891 \pm 0.048}$	0.548 ± 0.042	0.814 ± 0.045
	Unbalanced	0.565 ± 0.037	0.838 ± 0.062	0.527 ± 0.045	0.795 ± 0.071