METHOD

Open Access

SyntheVAEiser: augmenting traditional machine learning methods with VAE-based gene expression sample generation for improved cancer subtype predictions

Brian Karlberg[1], Raphael Kirchgaessner[1], Jordan Lee[1], Matthew Peterkort[1], Liam Beckman[1], Jeremy Goecks[1,2] and Kyle Ellrott[1*]

*Correspondence: ellrott@ohsu.edu

[1] Biomedical Engineering, Oregon Health and Science University, 3181 S.W. Sam Jackson Park Road, Portland, OR 97239-3098, USA
[2] Department of Machine Learning, Moffitt Cancer Center, Tampa, USA

Abstract

The accuracy of machine learning methods is often limited by the amount of training data that is available. We proposed to improve machine learning training regimes by augmenting datasets with synthetically generated samples. We present a method for synthesizing gene expression samples and test the system's capabilities for improving the accuracy of categorical prediction of cancer subtypes. We developed SyntheVAEiser, a variational autoencoder based tool that was trained and tested on over 8000 cancer samples. We have shown that this technique can be used to augment machine learning tasks and increase performance of recognition of underrepresented cohorts.

Keywords: Sample synthesis, Synthetic data, Data augmentation, Generative modeling, Feature engineering, Cancer subtyping, Molecular subtyping, Variational autoencoder, Transcriptomics, Gene expression

Background

Machine learning (ML) has become common in genomics as a means of modeling with complex biological data [1, 2]. Across numerous publications from The Cancer Genome Atlas (TCGA) [3], bulk RNA-sequencing has been shown as a robust way for defining cancer subtypes [4–8]. Bulk RNA-seq based signatures have been translated from basic research into FDA approved diagnosis used in the clinic [9, 10]. While this technique has found use in more common cancers, issues begin to arise with more rare cancer variants. Small sample counts within genomics datasets can impede model performance because of the high dimensionality of the feature space and imbalanced classes. In training performance analysis, we have found that about 120 samples are often needed before a machine learning recognizer can achieve best possible performance. For rare cancers, the resulting low sample counts of these omics datasets limit the capability of machine learning to improve patient outcomes. In this paper, we show that synthetic sample generation is one possible mechanism to mitigate these issues.

Synthetic data have been shown to improve the sample efficiency of learning across diverse domains such as image processing, physics modeling, and neuroscience [11]. We propose to apply data synthesis methods to augmenting transcriptomic data sets and improve the performance of a variety of prediction tasks. Neural networks with multiple hidden layers known as deep learning (DL) models combined with transfer learning techniques have demonstrated utility across a wide range of modeling applications within the rapidly evolving field of ML [12]. Generative deep modeling has emerged as a route to generate new samples and works by creating representations of complicated, high-dimensional probability distributions [13].

A variational autoencoder (VAE) is a feed-forward neural network that approximates a function for mapping high dimensional variables into representative, or latent, variables of a reduced dimension [14–16]. Continuous normalizing flows and generative adversarial networks (GANs) are similar generative models to VAEs [17]. VAE training is an unsupervised machine learning technique, and is unaware of any outside labels, such as cancer subtype, and is only concerned with organizing a low dimensional latent space based on the sample data. The defining characteristic of a VAE is stochastic backpropagation [14] which allows the model to overcome the accuracy and scalability challenges of modeling high-dimensional data.

The aims of this study were to (1) build a generative model for creating synthetic gene expression samples, (2) develop an algorithm for creating synthetic samples based on combining these latent representations of multiple parent samples with a labeled dataset, and (3) integrate this generative modeling framework with a traditional ML classifier to robustly quantify the improvement in predictive power from the addition of synthetic samples. This will demonstrate that VAEs can be trained on pan-cancer data and use that information to extrapolate into new tissue types. In these new cohorts, a minimal set of examples can be used to extrapolate a larger training set, and that extended training set can help to improve the performance of machine learning methods.

Traditional reasons for developing synthetic data sets for genomics and imaging include insufficient sample sizes, too many or too few features, disproportionate feature to sample size ratio, and the class imbalance problem [18]. Methods used to deal with class imbalance can be seen as analogous to synthetic sample generation methods. SMOTE [19] is the canonical method addressing the class imbalance problem. This method seeks to improve classifier performance by undersampling the majority class and oversampling the minority class. The minority samples are not directly sampled with replacement, rather the feature values of two or more samples are recombined with the feature value differences multiplied by a random number between zero and one to generate novel samples. However, in cases of high feature dimensionality and low signal-to-noise such as gene expression applications, the performance of SMOTE has been shown to both lack robust performance and be classifier dependent [20]. In cancer imaging, synthetic data have advanced to the point where a Synthesis Study Trustworthy Test (SynTRUST) has been proposed as a meta-analysis framework to address specific challenges across research and clinical care [21]. For computer vision tasks, there are a multitude of techniques for data augmentation [22] including skin lesion image synthesis

[23]. Generative methods have been shown to be robust across multiple data types, and as our research shows, this trend continues with transcriptomic data.

In the area of transcriptomic sample generation, there are previous publications outlining the use of GANs to create synthetic mRNA samples and improve prediction tasks [24]. These methods utilize noise or alternate omics inputs to generate new synthetic samples. Our method differs from these approaches in how the basis for new samples are seeded. Rather than utilizing random noise for permuting existing models, our model mixes features of multiple samples in latent space before reconstructing a new synthetic sample. Importantly, the mixing of features in the low dimensional latent space occurs between samples of the same target label. This ensures that each synthetic sample is effectively a high dimensional average of similar elements and avoids mixing samples from different classes.

When compared to other machine learning methods, deep learning methods are viewed as "black boxes" that produce predictions based on uninterpretable methods. Many times, especially when thinking about clinically oriented tasks, non-DL machine learning methods can provide interpretable models that can be connected to specific biological elements. These more interpretable models may be seen favorably for translational use cases, but may lack the ability to extract additional information from large sample populations in the same way that deep learning methods are able. For this study, we demonstrated that traditional ML can benefit from the addition of synthetic data generated by a VAE. By combining the pan-cancer training set, the VAE model is able to learn common patterns seen across multiple cancer types, and use that information to enrich a traditional machine learning task, even if that problem is only specific to a single cancer type. Because these performance gains are seen in methods, such as random forest (RF) based models, that are commonly viewed as being interpretable, the results of this technique can be interrogated.

Results

Generative model overview

A new method combining a VAE with a RF classifier and a corresponding software tool for sample synthesis was developed for applications in ML applied to gene expression data. Our dataset, based on samples from the TCGA, was structured for supervised categorical prediction where each sample was labeled with a cancer subtype within 25 primary tumor types based on gene expression profiles. In total, the 25 different tumor types were segmented into 99 molecular subtypes. For example, breast cancer (TCGA code BRCA), is subdivided into luminal A, luminal B, basal, and HER2 [25]. A transfer learning framework was applied for training the VAE on a sample set composed of all TCGA samples using a tumor sample holdout strategy (Fig. 1A). This involved a sequence of training and fine-tuning a VAE and using a RF classifier to compare the predictive accuracy of the data modes. The VAE was never trained on or received any information about tissue type or cancer subtype. So in the case of the BRCA cohort, the trained VAE was not presented with any BRCA samples, but rather learned the patterns from all other available cancer types. Thus in that experiment, BRCA could be viewed as a rare cancer that had never been encountered. A VAE model is trained to compress gene expression data into a latent space and then decompress a faithful copy of

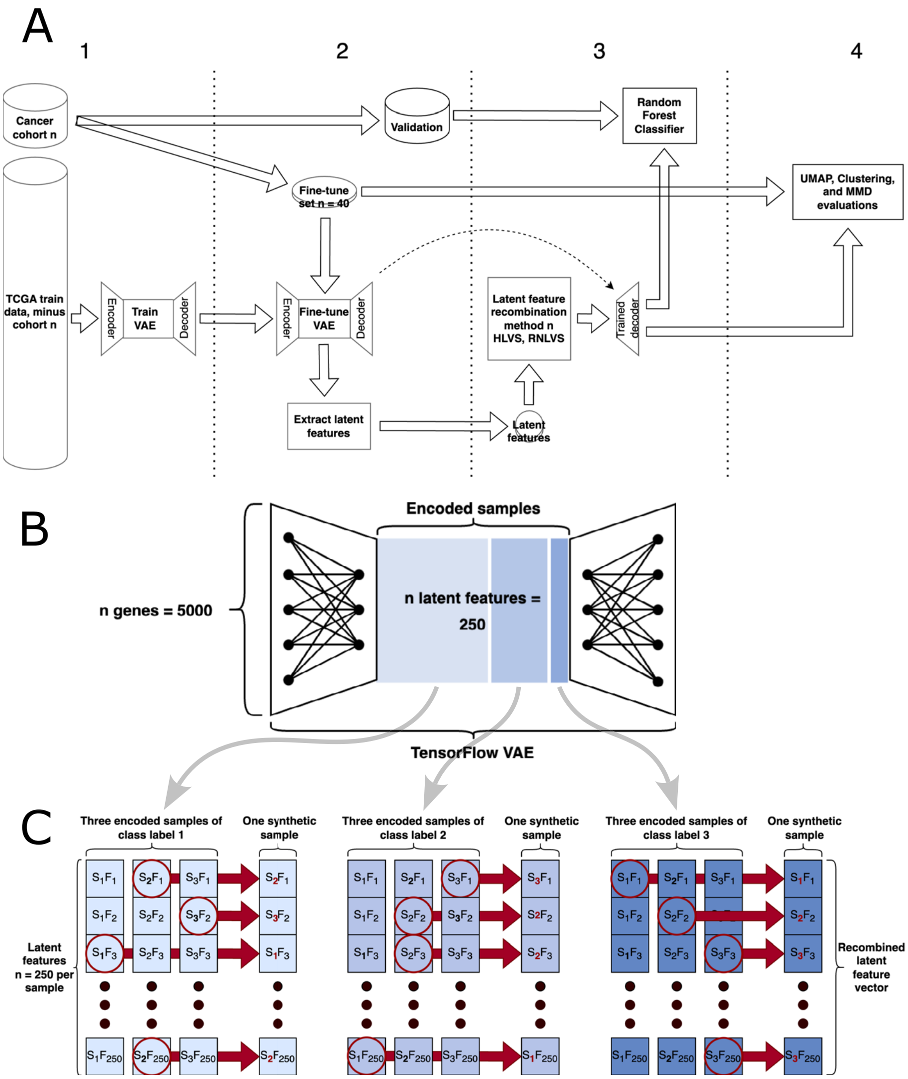
A
1
2
3
4
Cancer cohort n
Validation
Random Forest Classifier
Fine-tune set n = 40
UMAP, Clustering, and MMD evaluations
TCGA train data, minus cohort n
Encoder
Train VAE
Decoder
Encoder
Fine-tune VAE
Decoder
Latent feature recombination method n HLVS, RNLVS
Trained decoder
Extract latent features
Latent features
B
Encoded samples
n genes = 5000
n latent features = 250
TensorFlow VAE
C
Three encoded samples of class label 1
One synthetic sample
Three encoded samples of class label 2
One synthetic sample
Three encoded samples of class label 3
One synthetic sample
Latent features n = 250 per sample
Recombined latent feature vector
S_1F_1 S_2F_1 S_3F_1 S_2F_1
S_1F_2 S_2F_2 S_3F_2 S_3F_2
S_1F_3 S_2F_3 S_3F_3 S_1F_3
S_1F_{250} S_2F_{250} S_3F_{250} S_2F_{250}
S_1F_1 S_2F_1 S_3F_1 S_3F_1
S_1F_2 S_2F_2 S_3F_2 S_2F_2
S_1F_3 S_2F_3 S_3F_3 S_2F_3
S_1F_{250} S_2F_{250} S_3F_{250} S_1F_{250}
S_1F_1 S_2F_1 S_3F_1 S_1F_1
S_1F_2 S_2F_2 S_2F_2
S_1F_3 S_2F_3 S_3F_3 S_3F_3
S_1F_{250} S_2F_{250} S_3F_{250} S_3F_{250}

Fig. 1 Overview of the synthetic TCGA gene expression sample generation pipeline. **A** One cancer cohort at a time is designated for sample generation and removed from the TCGA sample set. The Tybalt VAE adapted from Way and Greene [26] is trained on these TCGA samples and then fine-tuned on 40 samples from the designated cohort_n. The remaining samples from cohort_n are used as validation. The latent feature values of three randomly selected samples from within each subtype are randomly recombined to form a latent sample feature vector which is then decoded with the trained decoder to generate a synthetic sample with feature dimensionality restored to that of the 5000 input genes. This latent feature value recombination and decoding process is repeated to generate 200 samples per subtype per validation split. The random forest classifier is trained five times, each time predicting on the entire held-out validation set to return a subtype prediction accuracy with quantified error. The train-validation split point at cohort_n and ensuing processes comprise a single experimental replicate which is repeated 25 times per cancer cohort. **B** Input gene expression features and latent dimension of the Tybalt VAE component of the pipeline. **C** Depiction of the three-sample version of the HLVS algorithm operating within each labeled class

the original signal. This encoder/decoder pair is then used to translate data into a "latent space" where values can be altered and decompressed back into "normal space" to create new samples. For our cross fold experiment, we produced 25 separate encoder/decoder pairs that each ignored a single cancer type. The sample generation pipeline was built around the Tybalt VAE [26] (Fig. 1B). The corresponding feature engineering pipeline takes the intersection of genes across cancer types and reduces the feature counts with mean absolute deviation. Original gene counts varied by primary tumor type are shown in Supplemental Table 1.

Using our hybrid DL/traditional ML synthesis and analysis pipeline, we analyzed the effect on subtype prediction performance with the RF classifier for 25 cancer types, using the cohort holdout strategy, where specific cancers were limited to 40 samples for training the RF classifier with all other samples from that cancer type used for performance validation. Effectively, our protocol simulated 25 separate rare cancer cases by restricting the RF training set to 40 samples. This process was repeated across these 25 cancer types, generating 200 additional samples per subtype to augment the 40 original samples. Thus, the number of synthetic samples generated varied for each primary tumor type, varying from 400 for the binary cancers up to 1400 for gastroesophageal (GEA) with seven subtypes. Using the validation sets, we measured F1 score performance improvement on the prediction of held out samples by a mean of 6.85% and a maximum improvement of 13.2% in lung squamous cell carcinoma (LUSC).

The transfer learning strategy involved first training the VAE on the gene expression data for approximately 8000 samples from the TCGA dataset, holding out one specific cancer type for testing. After the initial training, the VAE was fine-tuned on a subset of 40 randomly selected samples from the testing cancer type. The rationale for using this threshold of 40 samples for fine-tuning and sample generation across the 25 cancers was to balance a simulated reduced sample set with diminished accuracy while still having enough samples with which to generate quality synthetic samples. Reducing the batch size parameter of the VAE when transferring the model from training on a relatively large dataset to fine-tuning on a smaller dataset was identified as an important factor in learning a model capable of generating samples that improved predictive accuracy.

The effect of the quantity of training varied by cancer and could be inferred by the shape of the learning curves. In these data, the ratio of sample sizes in the training sets to fine-tuning sets was approximately two orders of magnitude and the number of epochs utilized in the training phase was observed to be a primary parameter in controlling the performance results of the generated synthetic data. This can be approached in absolute terms of training and fine-tuning epoch counts as well as from a ratio perspective. To investigate these effects, the quantity of TCGA training epochs was varied while holding the fine-tuning epochs constant at 150. The proportion of pan-TCGA training epochs to fine-tuning epochs on the cohort targeted for sample generation was observed to affect model performance asymmetrically across cohorts thus is a key point of consideration for generalizing this model to data with other distributional characteristics.

Synthetic sample generation

We tested two methods for synthetic sample generation: Random Noise Latent Variable Samples (RNLVS) and Hybrid Latent Variable Samples (HLVS). For a baseline, we deployed RNLVS which modulates samples with random noise in the latent space to create synthetic samples that are slightly perturbed from their original parent sample. We contrasted that method against HLVS which is designed to generate a synthetic sample of a specific subtype. It does this by randomly recombining the latent feature values of two or three samples from the same subtype into a novel latent feature vector (Fig. 1C).

Both two- and three-sample versions of HLVS were tested. The rationale for using three samples was to balance a generalized subtype representation based on a greater number of samples with the fact that for cancers with many subtypes, random samplings would begin to return one or zero samples of the rare subtypes as test set sizes decreased which negated the possibility of latent feature recombination. The decoder component of the VAE was then used to project each HLVS vector back into gene expression space. To validate the performance of RNLVS vs. HLVS derived synthetic samples, we tested machine learning models derived from cohorts generated using the two methods. We noted a marked improvement in performance using HLVS derived samples, as shown in Fig. 2.

For both the RNLVS and the HLVS sample generation methods and for each set of the experimental replicates, 200 samples were generated within each subtype for each of 25 replicates of 40 randomly selected training samples for a total of 5000 synthetic samples per subtype per replicate set. The trained decoder contained both pan-TCGA information as well as information from all subtypes via the 40 samples selected from within the cohort designated for sample generation. This was the result of the transfer learning design of the experiment in leveraging the combined learned representation of what a molecular cancer subtype is in general, with how molecular subtypes within a primary tumor cohort differed from each other.

After the synthetic samples were generated, they were mixed with the original training samples and then used to train a traditional ML RF classifier to predict on a validation set to assess performance of the sample generation. Across the 25 cancer subtype

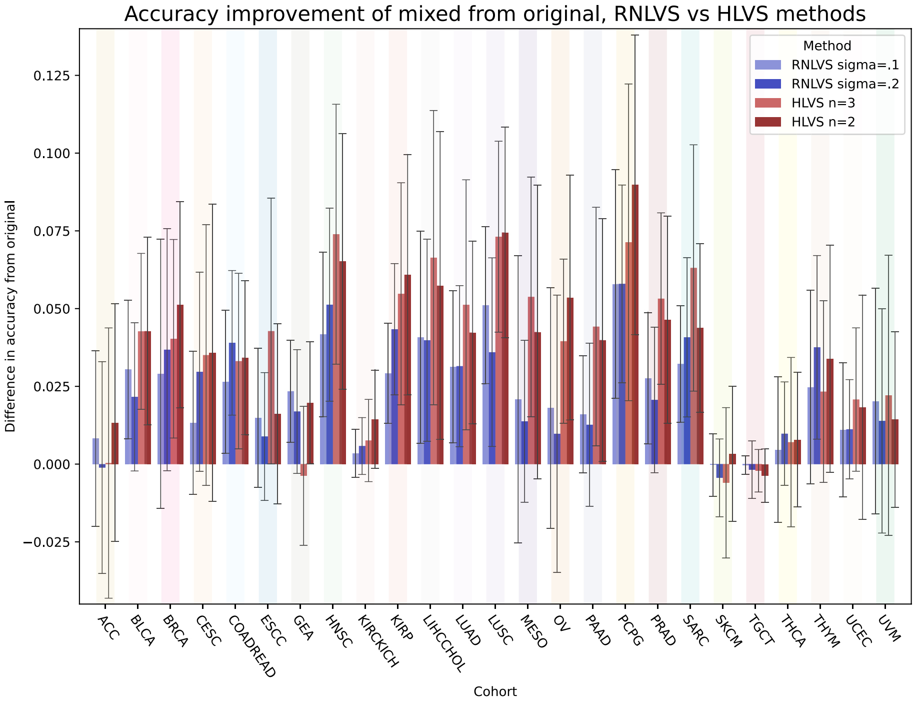

Fig. 2 Comparison of cancer subtype prediction accuracy improvement between the two RNLVS methods and two HLVS methods tested. With feature sets and model parameters fixed across primary cancer types, the HLVS methods return synthetic samples that result in greater accuracy improvement for 21 out of 25 cancer types

learning tasks, this resulted in improved classification accuracy for the majority of cancers. In our testing, 16 out of 25 cancers returned a statistically significant improved subtype prediction raw accuracy at a *p* value threshold of at least 0.05 as a result of mixing with the original 40 samples all of the 200 synthetic samples per subtype across the 25 experimental replicates.

Synthetic sample assessment

To quantify and compare the quality of the sample embeddings and generated synthetic samples, a Scikit-Learn RF classifier was selected based on its observed performance as a traditional ML method [27, 28]. The default hyperparameters of the RF classifier were used. Within each cohort and experimental replicate, the RF was first trained on the 40 original samples then used to predict on the validation set. This training of the RF was repeated on the VAE reconstruction of the same 40 samples once they had been encoded then re-coded back to gene expression space at the end of the fine-tuning epochs. The RF trained on these re-coded samples was then used to predict on the same validation as was used to evaluate the original 40 samples. Finally, this RF training and validation scheme was repeated on the pure synthetic and the mixture of the 40 original samples with the 200 synthetic samples per subtype. Raw prediction accuracy [Scikit-Learn metrics] was utilized for these comparisons. For each of these four data phases, the RF model was trained on the test set five times and used to predict on the validation each time to control for stochasticity in the RF model. The results of these five runs were averaged. A comparison of the performance results for two configurations within both the HLVS and RNLVS latent feature modification methods across the 25 TCGA cancers is shown in Fig. 2. The error shown is standard deviation and the magnitude relates to subsampling effects of low sample sizes. This illustrates heterogeneity within cohorts and number of subtypes within cohorts.

Once establishing this baseline configuration of the VAE training to attain predictive accuracy improvement for the majority of cohorts, learning curves were generated. The original and mixed datasets were subsampled in incremental steps with the random forest again repeated five times and averaged on each subsample set at each increment size. Learning curves for four selected cancers that returned increased raw accuracy from the addition of synthetic samples are shown in Fig. 3 with learning curves for the other 21 cohorts in Supplemental Fig. 1.

To characterize the similarity of the gene expression value distributions within the respective subtype label categories for the synthetic samples with the original samples from which they were generated, maximum mean discrepancy (MMD) was calculated for each pairwise combination of samples within three cancer types representing a range of subtype counts shown in Fig. 4A. A scatter plot of 2D UMAP dimensionality reduction was applied to visualize clustering of samples by subtype with mixing of original and synthetic data (Fig. 4B). If the distance between the expression value distributions of the original and synthetic samples is minimal, it would be expected that original and synthetic samples would cluster randomly within each subtype, with subtype status driving the clustering. Affirmingly, when applied to a mixed set of the original and synthetic samples, this clustering shows general separation of samples consistent by subtype as illustrated in Fig. 4C. Clustering of synthetic samples within a given subtype may be

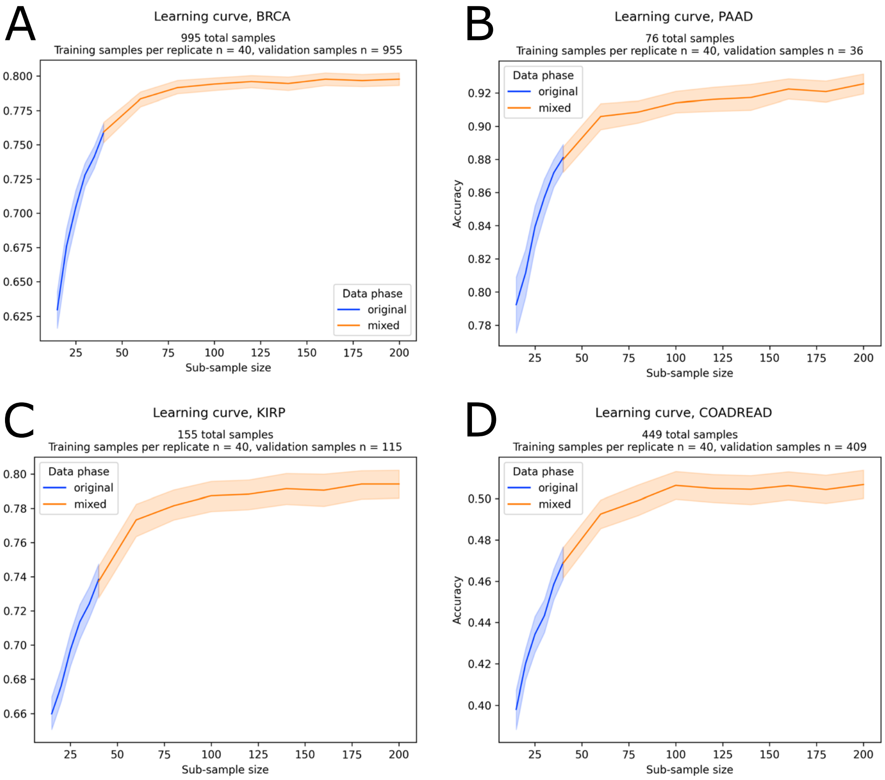

Fig. 3 Learning curve comparisons of individual cancers; predictive accuracy as a function of sample size aggregated across 25 experimental replicates. Original sample sets in blue showing subsampled accuracy growth up the 40 sample training threshold. Continuation of learning curves at larger sample counts with subsampling mixed original/synthetic sample sets in orange. **A** Breast invasive carcinoma learning curve, relatively smooth improvement in predictive accuracy with addition of synthetic samples up to a peak at approximately 150 samples. **B** Pancreatic adenocarcinoma, with 76 original samples shows a gradual improvement in predictive accuracy observed past 100 samples. **C** Performance improvement behavior of adding synthetic samples for kidney renal papillary cell carcinoma with 76 original samples, third smallest cohort. **D** Learning curve for colorectal adenocarcinoma, with more challenging to predict subtypes showing plateau in improved performance at around 50% accuracy

driven by the synthetic gene expression vectors being based on combinations of latent values from real samples resulting in synthetic samples being a non-linear interpolation of real samples. Although some degree of clustering by synthetic and original sample status is observed, despite this limitation, there is still an improvement in subtype predictive accuracy with either the pure synthetic or mixed data sets. A full survey covering another 22 TCGA cancer types can be found in Supplemental Fig. 2.

An additional quantitative inspection of the original and re-coded gene expression values was conducted with a root mean squared deviation (RMSD) comparison. $rmsd = \sqrt{mean((predictions - targets)^2)}$.

For each of the 40 samples in each experimental replicate, RMSD was calculated across the 5000 genes for the original and re-coded versions of the values. One thousand RMSD values, 40 samples times 25 replicates, for each cohort are shown in Fig. 5.

Recursive feature elimination, a statistical feature selection algorithm, was applied to identify specific gene features of importance within the original, re-coded, and synthetic samples. For three selected primary tumor types, BRCA, LUAD, and PRAD,

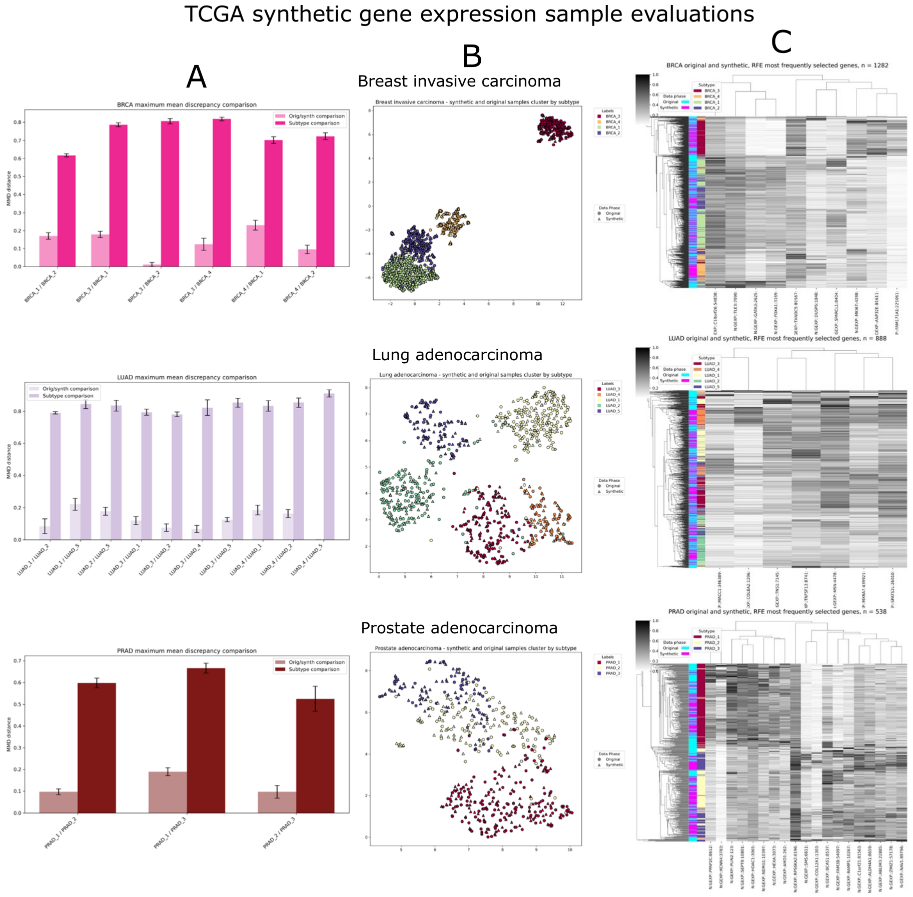

Fig. 4 **A** MMD statistics for each pair of cancer subtypes within each primary cancer type comparing the difference of gene distributions with samples split by subtype vs. samples split by original/synthetic. **B** Scatter plots of 2D UMAP projections showing interspersed clustering of original and synthetic samples separated by cancer type. **C** Cluster maps showing propensity of samples to cluster by subtype with interspersion of synthetic and original samples within each subtype. Color bars on left in pink and light blue show original or synthetic sample status and saturated color bars on right show subtype sample status

the intersections of features selected across the three data phases are presented in Fig. 6A. Consistency in the specific features selected from each phase of the data would be expected in the case of consistency in the gene expression values across the data phases. For these three cancers, this pattern of consistency was observed—in BRCA, 71 features were commonly selected across all three phases of the data compared with 39, 17, and 16 features commonly selected across the pairwise combinations of the data phases. Eighty-four and 66 features were commonly selected across all data phases for LUAD and PRAD, respectively, with lower numbers again observed for any pairwise combinations of data phases. This observation indicates biological consistency of the synthetic data with the original samples. Permutation-based feature importance scores were calculated within each of the three data phases for each of these three cancers for these selected features shown in Fig. 6B. The gene FOXA1

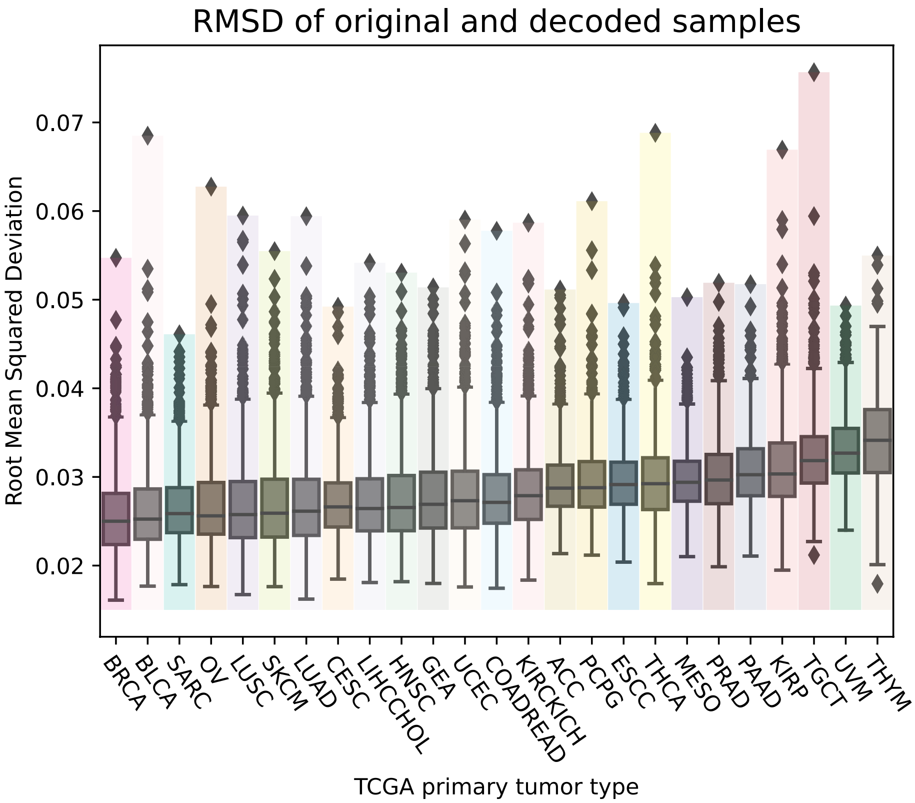

Fig. 5 Correlation of gene expression RMSD with the difference in prediction accuracy by primary cancer cohort. The gene expression RMSD is the average root mean squared deviation across each sample's 5000 gene expression values input to the VAE with the corresponding re-coded values of encoding and decoding these input values. The y-axis, delta accuracy is the change in average subtype predictive accuracy across the 25 replicates of 40 input samples vs. the average of the predictions at 140 and 160 sample size mixed sample sets of the original 40 samples and synthetic samples within each experimental replicate

scored in the top three of the most important features for BRCA across all data phases and SEPT9 scored in the top three across all phases for PRAD.

For further validation of the VAE-based genomic samples, we tested the algorithm on single-cell data, by using oligodendroglioma intra-tumor heterogeneity gene expression data obtained from the Broad Single Cell Portal [25]. To create two distinct cohorts, this data was filtered for malignant and Microglia/Macrophage cell labels which were the analog to the cancer subtype labels in the original experiments. The Microglia/Macrophage class was down-sampled to 250 samples to approximately match the 235 samples in the malignant class. Filtering samples with missing expression values from this set yielded a prepared set of 418 samples with 235 samples of the Microglia/Macrophage class and 183 samples of the malignant class. The 23,686 raw gene features were reduced to the 5000 gene features with the same greatest mean absolute deviation method utilized in the original experiments. The data was randomly split into a pre-training set of 268 samples and a fine-tuning set of 150 samples for input to the VAE sample generation tool in its same configuration from the original experiments. The generated data were evaluated against the original data

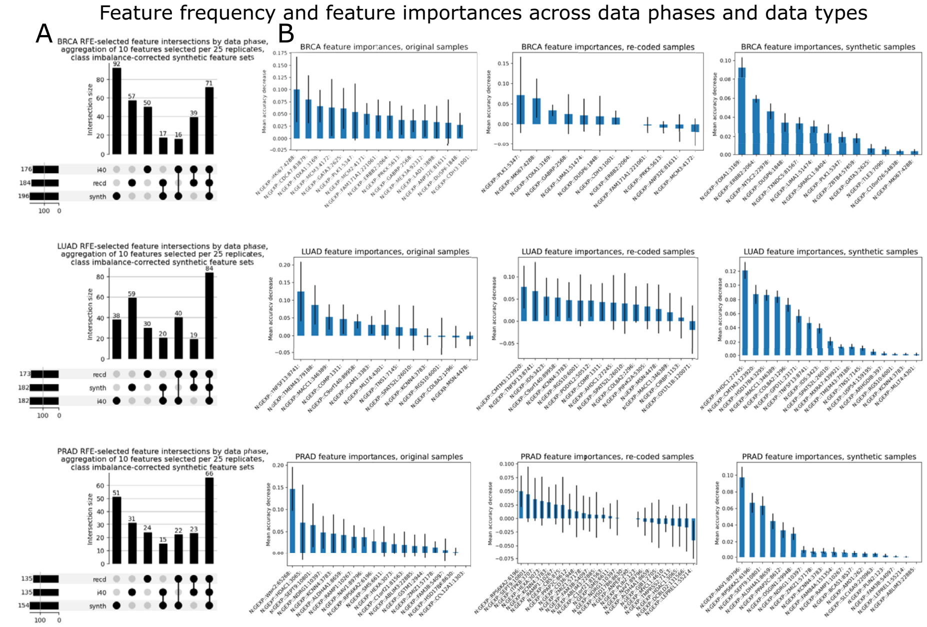

Fig. 6 **A** Intersections of features across original, re-coded, and synthetic samples. **B** Feature importance scores calculated with Scikit-Learn Permutation Importance algorithm for features selected three or more times across the 25 experimental replicates

with UMAP clustering (Supplemental Fig. 3), showing synthetic and original single-cell samples clustering by cell type and not clustering by real or synthetic status.

Discussion

In order to test the robustness of our method, we benchmarked the recognition of cancer subtypes as defined by the TCGA cohort. Because each tissue type has extremely different dynamics, and the subtypes within each of these cancers are defined by different rules, this allowed us to perform robust benchmarking in translation, by removing entire cancer types from the original training set. Additionally, the dataset has cohorts of extremely different sample sizes, with groups with 995, such as the case of breast invasive carcinoma (BRCA) and as few as 74 in the case of mesothelioma (MESO) and uveal melanoma (UVM). In our tests with the TCGA dataset, the sample size limitation is most pronounced in cancers with rare subtypes such as bladder urothelial carcinoma (BLCA) or kidney renal papillary cell carcinoma (KIRP), primary tumors with subtypes containing less than 10 samples. Using a leave-tissue out cross fold strategy, every cancer type was tested as if it was a rare cancer type. Our method to increase sample sizes of rare, molecularly defined subtypes to solve the class imbalance problem could be of particular utility for feature sets reduced to the number of samples required to train accurate models.

Augmenting datasets with synthetic samples created with the HLVS methods outperformed the RNLVS derived samples in 20 out of 25 of the specific machine learning tasks tested. The three-sample and two-sample variations of the HLVS method performed comparably well with average predictive improvement over the original samples of 3.64% ± 0.04% and 3.67 ± 0.04% percentage points, respectively. Although random noise methods combined with generative modeling improved performance for the majority of tested cancers, the performance gains were greater across most cancers with the combination of generative modeling and HLVS methods.

This study sought to leverage the representation learning capabilities of generative modeling with the interpretability of traditional ML to develop a method for transcriptomic sample generation. The software tool developed can be directly applied to supervised categorical prediction tasks with gene expression data sets and potentially adapted to other transcriptomic based ML tasks including regression. This was an improvement on previous methods in robustness for this type of genomics prediction task characterized by a large ratio of features to samples. By using transfer learning techniques to train a model on data related to the fine-tuning data and final prediction domain, the model is less prone to overfitting. The training method utilized in this study was to include all of the TCGA cohorts, except the cancer type designated for testing, to prepare the VAE for fine-tuning. The RMSD statistics characterizing the reconstruction values between the best-fit cancer, BRCA, and poorest-fit cancer, THYM, showed that the mean of every tested cancer was within the error of every tested cancer. This demonstrates generalizability of a transfer learning strategy where fewer epochs are used for training than fine-tuning and the batch size is reduced in the fine-tuning from the training.

In the benchmarking seen in Fig. 2, the cancer types that received no performance improvement, namely SKCM and TGCT, is likely due to issues beyond sample generation. The subtypes in skin cutaneous melanoma (SKCM) were originally defined using mutation markers. Training ML models on gene expression fails to capture that original information used for defining the subtyping, and instead relies on gene expression values that happen to be correlated with the subtype, rather than elements with direct biological implications. Similarly testicular germ cell cancer (TGCT) subtypes are largely defined by DNA methylation and miRNA [29]. In these cases, boosting the population of gene expression data will do very little to better illuminate the underlying biology.

To quantify the similarity of the synthetic and original data, maximum mean discrepancy (MMD), a nonparametric distance statistic that is robust in comparing sample groups comprising different distributions [30], was calculated for each subclass pair within three primary cancer types of differing numbers of subtypes. For all subclass pair comparisons, the distance between subclasses was significantly greater than the distance between the original and synthetic samples as shown in Fig. 4A. This observation is reinforced with UMAP clustering behavior shown in Fig. 4B, where original and synthetic samples cluster uniformly within each cancer subtype. The sample cluster map of gene expression value experiments, seen in Fig. 4C, also showed aggregation of samples within subtypes of mixed synthetic and original data.

The feature selection experiments reveal a greater intersection of features across the original, re-coded, and synthetic samples than within any pairwise combination of these three phases as shown in Fig. 6A. This observation is validating of both the model encoding and the synthetic data.

The feature importance scores indicate reduced error associated with the synthetic data compared with the original and re-coded feature importance scores as shown in Fig. 6B. This effect is driven by improved statistical power of synthetic data sets and the solving of the class imbalance problem with 200 synthetic samples per cancer subtype vs. 40 total original samples within each replicate. This demonstrates the potential utility of the method to improve confidence in biomarker target identification for rare cancer subtypes.

Conclusions

This work demonstrates that generative models based on neural networks can be combined with traditional ML as an effective means to generate synthetic gene expression samples. This allows for information from other tissue and cancer types to provide priors for learning patterns in a new cohort. Rare cancers, which traditionally see much lower rates of collection and sequencing, can benefit from augmenting their dataset. Additionally, non-DL machine learning methods, traditionally seen as more trustworthy or easier to interpret than DL models, can still benefit from these methods.

Methods

Data provenance and feature engineering

The data utilized for developing this sample generation method and software tools were derived from a TCGA-based curated dataset from the Tumor Molecular Pathology working group and can be downloaded from the NCI's Genomic Data Commons [31] [https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022]. These data files were tabular comprising 8009 samples across 25 primary tumor types and 99 subtypes. The gene expression features utilized in this study were down-selected via mean absolute deviation to the 5000 most differentially expressed features per the original Tybalt method [26]. The raw expression values were normalized with the Scikit-Learn MinMaxScaler function within each cohort and within each feature. Four of the cancers utilized in this study have only two subtypes making them a binary supervised classification problem whereas the remaining cancers are multiclass with three to seven subtypes per primary tumor type.

Generative modeling framework

The sample generation model (Fig. 1) was built around a variational autoencoder (VAE) adapted from [26]. A latent feature dimension of 250 was used for all experiments and all experiments used 150 epochs for model fine-tuning. One cohort at a time was designated for generating synthetic samples and removed from the combined TCGA set. The VAE was then trained on all of the remaining TCGA samples for 1, 2, 3, 4, 10, 20, or 30 epochs. The batch size was set at 50 for each of these initial TCGA trainings. From the cohort selected for sample generation, a training set of 40 samples was randomly selected without replacement. The remaining samples were used as a validation set of size $n_v = n - 40$. The various epoch-count and feature set versions of the TCGA-trained VAE were then each fine-tuned for 150 epochs at batch size of 10 on the 40 samples within each replicate. A learning rate of 0.0005 was used for both the TCGA training and fine-tuning steps. This framework is represented symbolically in Algorithm 1.

```
Given N samples within $\Gamma$ cohorts such that for each cohort $\Psi \subset \Gamma$,
contains ん number of classes $\psi_i$:
------------------------------
$\varrho$ = an experimental replicate
$\rho$ = 25, n replicates to repeat for each $\Psi$
$\lambda$ = a subset of $\Psi$ for training VAE
$n \leftarrow 40$, n samples for fine-tuning and synthesis within $\Psi$
$\gamma \leftarrow 50$, random sample repeats to attain minimum $\nu$
$\Theta$ = Latent sample vectors of $\Psi$
$\theta$ = Latent sample vectors of $\psi_i$
$\nu \leftarrow 3$, n samples within $\theta$ from which to generate a synthetic sample
$\vartheta$ = $\nu$ samples from within $\theta$
dim $\vartheta = (n, 250)$
$\underline{\omega}$ = a synthetic latent feature vector
$\omega$ = a synthetic latent feature value
$\Omega$ = Synthetic latent feature vector set from $\Theta$
$\varphi \leftarrow 200$, n samples to generate for each $\psi_i$
------------------------------
for $\varrho$ to $\rho$ do
 |  $\lambda$ = sample($\Psi$, $n$)
 |  while min($\psi_i$ of $\lambda$) $\geq \nu$ do
 |   |  $\lambda$ = sample($\Psi$, $n$)
 |   |  $\iota$ += 1
 |   |  if $\iota = \gamma$ then
 |   |   |  continue
 |   end
 |   train VAE on $\Gamma - \Psi$ and fine-tune on $n$
end
------------------------------
for VAE trained on $\Gamma - \Psi$ and fine-tuned on $n$ do
 |  D $\leftarrow$ decoder extracted from VAE
 |  for $\theta$ in $\Theta$ to ん do
 |   |  $\vartheta$ = sample($\theta$, $\nu$)
 |   |  for $\underline{\omega}$ in $\Theta$ to ん do
 |   |   |  $\omega$ = sample($\underline{\omega}$, 1)
 |   |   |  $\Omega$.append($\omega$)
 |   |  end
 |  end
 |  D($\Omega$)
end
```

Algorithm 1. Categorically labeled synthetic sample generation from the latent feature vectors of a variational autoencoder, VAE

The initial validation split of 40 fine-tuning samples within the cohort designated for sample generation defined each experimental replicate. Within each replicate, the samples not selected into the set of 40 for fine-tuning are designated as the validation set such that the number of validation samples varies by cohort because each cancer cohort contains a different number of total samples. Results for 25 replicates were produced for each cohort. Replicates returning less than three (or two in the alternate HLVS version) samples for any subtype within the random 40 cohort samples were rejected because this was the sampling threshold for the latent feature recombination algorithm, described below.

The training/validation split constituted an experimental replicate and was repeated 25 times for each cohort. If a training set contained less than three samples within a subtype, the sampling was repeated up to 50 times attempting to obtain at least three samples per subtype. The replicate was omitted if three (or two) samples were not obtained over these 50 repeats. The latent feature object was subset by subtype. Three samples at a time were chosen without replacement and sent to a function where the latent feature values from these three samples were randomly recombined into a novel latent feature vector. Two hundred synthetic samples were generated within each subtype for each primary tumor type. This 200 synthetic subtype sample by 150 synthetic latent feature object was returned to the original 5000 dimension feature space using the trained VAE decoder.

To evaluate the HLVS results, a set of experimental control results were generated with RNLVS derived from Gaussian noise injection. The effectiveness of Gaussian noise injection has been mathematically described for multi-layer perceptron neural networks in terms of the heat kernel and Taylor expansions [32]. This form of noise injection was implemented in the present study with sigma values of 0.1 and 0.2 for the Gaussian function applied to corresponding sets of latent feature values with a zero-floor or rectification operation to prevent negative expression values.

Within each experimental replicate, the 40 training samples were used to train a Scikit-Learn random forest model with default hyperparameters. This random forest was trained on the original training samples of the data then was used to predict on the validation set as to establish a baseline accuracy score with which to compare with the synthetic samples. The process of training the random forest and predicting on the validation set was repeated for the re-coded, synthetic, and mixed sample sets denoted by the green, red, and orange arrows, respectively, in Fig. 1. The mixed sample set was the generated synthetic sample set blended with the original 40 training samples.

The imbalanced class problem was eliminated by adding 200 synthetic samples to each class. The result was that subtypes with relatively few samples were augmented with proportionally more synthetic samples.

For the comparisons of the distributions of the original and synthetic samples within the cancer subtype class pairs shown in Fig. 4A, the MMD formula utilized is given in Algorithm 2.

```
Given two sets of samples, $gexp\_1$ and $gexp\_2$, and a kernel parameter $\gamma$:
$K\_XX \leftarrow \mathrm{rbf\_kernel}(gexp\_1, gexp\_1, \gamma)$
$K\_XY \leftarrow \mathrm{rbf\_kernel}(gexp\_1, gexp\_2, \gamma)$
$K\_YY \leftarrow \mathrm{rbf\_kernel}(gexp\_2, gexp\_2, \gamma)$
$m \leftarrow$ number of rows in $gexp\_1$
$n \leftarrow$ number of rows in $gexp\_2$
$mmd \leftarrow (\sum(K\_XX)\text{-}trace(K\_XX))/(m*(m\text{-}1))$
$mmd \mathrel{+}= (\sum(K\_YY)\text{-}trace(K\_YY))/(n*(n\text{-}1))$
$mmd \mathrel{-}= 2*\sum(K\_XY)/(m*n)$
$mmd \leftarrow \max(mmd, 0)$
$mmd \leftarrow \sqrt{mmd}$
return $mmd$
```

Algorithm 2. Compute MMD

The UMAP clusterings of original with synthetic samples within each intended cancer subclass shown in Fig. 4B were done by subsampling the pool of generated samples within each subtype the same number of synthetic samples as unique original samples in the aggregated input across the 25 experimental replicates. This unified set of balanced counts of original and synthetic samples within each subtype for each primary tumor type was input to the UMAP dimensionality reduction algorithm for subsequent scatter plotting. The clustering algorithm was the default "average" method implemented in the Scipy dependency of the Seaborn Clustermap function [33].

The feature importance algorithm utilized was Scikit-Learn Permutation Importance and was run on each of the 25 experimental replicates within the original gene expression data, the reconstructed expression data, and the synthetic sample expression data. Ten features were selected from each replicate within each data phase. The intersections of every combination of selected features were identified and binned for plotting in the UpSet plot.

Software tool requirements:

- TensorFlow 2.10
- Python 3.9
- Scikit-Learn 1.1.3

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03431-3.

Additional file 1: Table S1. Summary of curated TCGA genomic input data. The data source was the NCI's Genomic Data Analysis Network Tumor Molecular Pathology Analysis Working Group.

Additional file 2: Fig. S1. Learning curves showing predictive accuracy as a function of sample size for the 21 cancer cohort test sets in addition to the four cancers presented in Fig. 3. Subsampling of original samples in blue showing growth in accuracy up to the 40 total test samples. Continuation of accuracy response to increased sample sizes by means of adding synthetic samples shown in orange.

Additional file 3: Fig. S2. MMD, UMAP, and cluster maps showing similarities and differences of samples by original or synthetic status with cancer subtype status; 22 of 25 cancers studied not presented in main Fig. 4

Additional file 4: Fig. S3. Application of VAE-based genomic sample generation method to scRNA-seq data obtained from the Broad's Single Cell Portal. Inset shows single cell sample scores of stemness and differentiation; yellow box denotes region of samples with class (cell type) separation selected for sample generation.

Additional file 5. Review history.

Peer review information

Kevin Pang and Andrew Cosgrove were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 5.

Authors' contributions

KE and BK developed the hypotheses, designed the experiments, and wrote the manuscript; KE, RK, and BK developed the generative VAE model; BK wrote the code, executed the experiments, and created the figures; KE, JL, MP, LB, and JG consulted on the scientific and technical aspects through the development iterations.

Funding

Support from the National Cancer Institute and Oregon Health and Science University's Knight Cancer Institute is gratefully acknowledged, including NIH Bridge2AI grant 1U54HG012517, NCI GDAN 5U24CA264007, and NIH 1S10OD034224.

Data availability

The software tool, SyntheVAEiser, is available at https://github.com/ohsu-comp-bio/syntheVAEiser [34] and https://doi.org/10.5281/zenodo.13948571 [35] under the Apache 2.0 license.

Declarations

Ethics approval and consent to participate

N/A.

Competing interests

N/A.

Received: 11 October 2023 Accepted: 30 October 2024
Published online: 18 December 2024

References

1. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16:321–32.
2. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23:40–55.
3. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol. 2015;19:A68–77.
4. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell. 2015;163:506–19.
5. Network CGA. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517:576–82.
6. Roh W, Geffen Y, Cha H, Miller M, Anand S, Kim J, et al. High-resolution profiling of lung adenocarcinoma identifies expression subtypes with specific biomarkers and clinically relevant vulnerabilities. Cancer Res. 2022;82:3917–31.
7. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474:609–15.
8. Fishbein L, Leshchiner I, Walter V, Danilova L, Robertson AG, Johnson AR, et al. Comprehensive molecular characterization of pheochromocytoma and paraganglioma. Cancer Cell. 2017;31:181–93.
9. Picornell AC, Echavarria I, Alvarez E, López-Tarruella S, Jerez Y, Hoadley K, et al. Breast cancer PAM50 signature: correlation and concordance between RNA-Seq and digital multiplexed gene expression technologies in a triple negative breast cancer series. BMC Genomics. 2019;20:452.
10. Jensen M-B, Lænkholm A-V, Balslev E, Buckingham W, Ferree S, Glavicic V, et al. The Prosigna 50-gene profile and responsiveness to adjuvant anthracycline-based chemotherapy in high-risk breast cancer patients. NPJ Breast Cancer. 2020;6:7.

11. de Melo CM, Torralba A, Guibas L, DiCarlo J, Chellappa R, Hodgins J. Next-generation deep learning based on simulators and synthetic data. Trends Cogn Sci. 2022;26:174–87.
12. Hosna A, Merry E, Gyalmo J, Alom Z, Aung Z, Azim MA. Transfer learning: a friendly introduction. J Big Data. 2022;9:102.
13. Ruthotto L, Haber E. An introduction to deep generative modeling. GAMM-Mitt. 2021;44. Available from: https://onlinelibrary.wiley.com/doi/10.1002/gamm.202100008.
14. Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. arXiv [stat.ML]. 2014. Available from: http://arxiv.org/abs/1401.4082.
15. Kingma DP, Salimans T, Welling M. Variational dropout and the local reparameterization trick. arXiv [stat.ML]. 2015. Available from: http://arxiv.org/abs/1506.02557.
16. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv [stat.ML]. 2013. Available from: http://arxiv.org/abs/1312.6114v11.
17. Bilodeau C, Jin W, Jaakkola T, Barzilay R, Jensen KF. Generative models for molecular discovery: recent advances and challenges. Wiley Interdiscip Rev Comput Mol Sci. 2022;12. Available from: https://onlinelibrary.wiley.com/doi/10.1002/wcms.1608.
18. Kokol P, Kokol M, Zagoranski S. Machine learning on small size samples: a synthetic knowledge synthesis. Sci Prog. 2022;105: 368504211029777.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. JAIR. 2002;16:321–57.
20. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14: 106.
21. Osuala R, Kushibar K, Garrucho L, Linardos A, Szafranowska Z, Klein S, et al. Data synthesis and adversarial networks: a review and meta-analysis in cancer imaging. Med Image Anal. 2023;84: 102704.
22. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. Journal of Big Data. 2019;6:1–48.
23. Baur C, Albarqouni S, Navab N. MelanoGANs: high resolution skin lesion synthesis with GANs. arXiv [cs.CV]. 2018. Available from: http://arxiv.org/abs/1804.04338.
24. Ahmed KT, Sun J, Cheng S, Yong J, Zhang W. Multi-omics data integration by generative adversarial network. Bioinformatics. 2021;38:179–86.
25. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–7. https://doi.org/10.1200/JCO.2008.18.1370.
26. Way GP, Greene CS. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pac Symp Biocomput. 2018;23:80–91.
27. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on machine learning. New York: Association for Computing Machinery; 2006. p. 161–8.
28. Kim AA, Rachid Zaim S, Subbian V. Assessing reproducibility and veracity across machine learning techniques in biomedicine: a case study using TCGA data. Int J Med Inform. 2020;141: 104148.
29. Shen H, Shih J, Hollern DP, Wang L, Bowlby R, Tickoo SK, et al. Integrated molecular characterization of testicular germ cell tumors. Cell Rep. 2018;23:3392–406.
30. Gretton A, Borgwardt KM, Rasch MJ. A kernel two-sample test. J Mach. 2012. Available from: https://www.jmlr.org/papers/volume13/gretton12a/gretton12a.pdf?ref=https://githubhelp.com.
31. Kyle Ellrott, Christopher K. Wong, Christina Yau, Mauro A. A. Castro, Jordan A. Lee, Brian J. Karlberg, Jasleen K. Grewal, Vincenzo Lagani, Bahar Tercan, Verena Friedl, Toshinori Hinoue, Vladislav Uzunangelov, Lindsay Westlake, Xavier Loinaz, Ina Felau, Peggy I. Wang, Anab Kemal, Samantha J. Caesar-Johnson, Ilya Shmulevich, Alexander J. Lazar, Ioannis Tsamardinos, Katherine A. Hoadley, The Cancer Genome Atlas Analysis Network, A. Gordon Robertson, Theo A. Knijnenburg, Christopher C. Benz, Joshua M. Stuart, Jean C. Zenklusen, Andrew D. Cherniack, Peter W. Laird. TCGA cancer subtype assignment of patient samples using compact feature sets. Available from: https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022. Cited 2024 Oct 24.
32. Grandvalet Y, Canu S, Boucheron S. Noise injection: theoretical prospects. Neural Comput. 1997;9:1093–108.
33. Müllner D. Modern hierarchical, agglomerative clustering algorithms. arXiv [stat.ML]. 2011. Available from: http://arxiv.org/abs/1109.2378.
34. Karlberg B, Kirchgässner R, Lee J, Peterkort M, Beckman L, Goecks J, Ellrott K. SyntheVAEiser Github; 2024. Available from: https://github.com/ohsu-comp-bio/syntheVAEiser.
35. Karlberg B, Kirchgässner R, Lee J, Peterkort M, Beckman L, Goecks J, Ellrott K. SyntheVAEiser Zenodo; 2024. Available from: https://zenodo.org/doi/10.5281/zenodo.13948571.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.