Proceedings of the 7th International Conference on Statistics: Theory and Applications (ICSTA'25)
Paris, France - August, 2025
Paper No. 174
DOI: 10.11159/ICSTA25.174

Investigating Tabular Generative Models for Synthetic Data Generation in PDAC Bulk Gene Expression Data

Sultan Sevgi TURGUT ÖGME, Zeyneb KURT[2], Nizamettin AYDIN[3]
[1]Department of Computer Engineering, Yildiz Technical University, Istanbul, Türkiye
sturgut@yildiz.edu.tr
[2]Information School, The University of Sheffield, The Wave, 2 Whitham Rd, Sheffield S10 2SJ, United Kingdom
z.kurt@sheffield.ac.uk
[3]Department of Computer Engineering, Istanbul Technical University, Istanbul, Türkiye
naydin@itu.edu.tr

Abstract - Pancreatic Ductal Adenocarcinoma (PDAC) is among the deadliest cancer types, with early detection being critical to improving survival rates. However, developing effective detection models is challenging due to the need for high-quality, class-balanced datasets. Generative models have recently gained attention for addressing this issue. In this study, we compare three tabular data-based generative models: Conditional Tabular Generative Adversarial Networks (CTGAN), Tabular Variational Autoencoder (TVAE), and Gaussian Copula (GC) using PDAC gene expression data. We first constructed an integrated dataset by curating six PDAC studies and applied an ensemble-based feature selection approach combining Differential Expression (DEG) analysis, ANOVA, Lasso, and Mutual Information. The synthetic data were evaluated both statistically (using Correlation Discrepancy (CD), Kolmogorov-Smirnov(KS), and Statistical Similarity(SS) metrics) and biologically (via PDAC marker genes), as well as visually in 2D-PCA space. The GC model produced the most realistic synthetic data with 0.1482 CD, 0.8120 KS, and 0.9529 SS metric values, similar expression level with PDAC markers, and uniform distribution with real data. TVAE followed GC. Based on these findings, we proposed an ensemble model combining GC and TVAE-generated samples. Classification experiments using Random Forest (RF) and Support Vector Machine (SVM) demonstrated that, while the ensemble generative model did not achieve the highest performance (0.8541 precision, 0.8570 recall, 0.8533 F1-measure and 0.9236 AUC) for SVM but achieved (0.8549 precision, 0.8623 recall, 0.8568 F1-measure and 0.9246 AUC) for RF, so it is a promising model for future applications.

Keywords**:** generative models, gene expression, pancreatic cancer, ensemble

1. Introduction

RNA-sequencing (RNA-seq) is a technique that utilizes next-generation sequencing (NGS) to process RNA molecules of a biological sample. It is widely used in cancer studies and helps to understand the formation and development of cancer, identify cancerous tissues and types, and develop cancer prevention and treatment solutions. Pancreatic Ductal Adenocarcinoma (PDAC) is one of the most lethal cancer types that progresses rapidly, highly metastatic, and difficult to treat. While the five-year survival rate is stated to be 9% due to late diagnosis [1], early diagnosis increases the survival rate by almost 20% [2]. Bulk gene expression profiling has become a crucial tool in understanding PDAC and guiding diagnosis and prognosis. However, the scarcity of high-quality labeled datasets induces the imbalanced distribution across the two classes(tumor vs normal), therefore an important challenge emerges to the development of robust, unbiased machine-learning models. Generative models are a promising solution to this problem and can create realistic artificial samples, thereby facilitate data augmentation and improved downstream analysis.

Kumar and Kiran conducted a comparative study with TVAE and CTGAN using two highly imbalanced Malware Detection and Wafer Anomaly datasets. Their results show that TVAE fails to produce data from a small number of samples [3]. Another study used the TCGA Pan-Cancer gene expression RNA-seq dataset to evaluate the performance of GANs and Diffusion Models (DMs). They applied classification methods and pathway analyses to assess statistical and biological similarity. It was concluded that GAN models produced higher-quality data [4]. Eltager et al., stating that generative models are increasingly gaining interest in computational biology, compared different versions of the VAE model on the cancer genome atlas (TCGA) Pan-Cancer data. Beta Total Correlation Variational Autoencoder (β-TCVAE) and Disentangled Inferred Prior VAE (DIP-VAE) were found to be the best versions and it was emphasized that the selection of hyperparameters for these models significantly affects the performance [5].

In this study, we investigate several state-of-the-art generative models to augment bulk gene expression data from human PDAC samples. Specifically, we compare tabular data-based models like Conditional Tabular Generative Adversarial Network (CTGAN) [6], Tabular Variational Autoencoder (TVAE) [6], and Gaussian Copula (GC). Each of these methods represents a distinct type of synthesis strategy: adversarial learning, variational inference, statistical modeling, respectively. While generative models have been mostly applied to images and text, their use in biomedical data particularly in high-dimensional gene expression datasets remains a field that requires further development. Our goal is to evaluate the synthetic data produced by these generative models in terms of statistical properties, biological relevance and similarity to the real data, and to assess their performances in downstream analyses such as classification. The results of this study aim to inform the development of more reliable synthetic data generators for cancer genomics and highlight the potential of generative modeling in supporting data-driven cancer research.

2. Methodology

Figure 1 shows the workflow of this study.

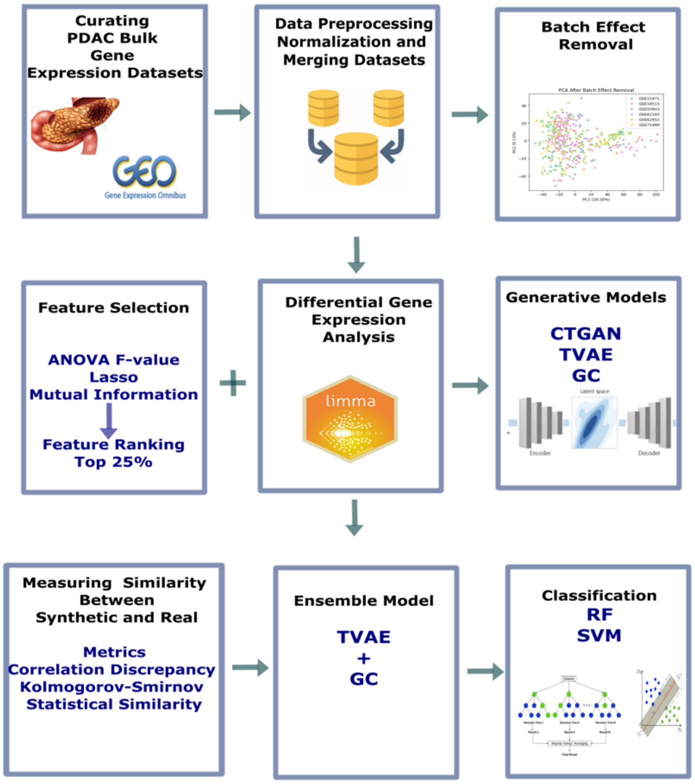

Fig. 1: Workflow

2.1. Dataset

We utilized six publicly available bulk gene expression datasets derived from human PDAC tissue samples. Datasets are available at the GEO repository [maybe a link refernce to GEO main page] with GSE15471[7], GSE16515[8], GSE62165[9], GSE71989[10], GSE62452[11], GSE55643[12] Geo accession numbers. We formed an integrated dataset by combining them to increase the number of samples. Table 1 summarizes details of the individual datasets including GEO ID, PubMed ID, number of tumour/normal samples.

Table 1: Datasets

GEO ID	Type	Sample Size	# of tumor	# of normal
GSE15471	PDAC-Human	78	39	39
GSE16515	PDAC-Human	52	36	16
GSE62165	PDAC-Human	131	118	13
GSE71989	PDAC-Human	22	14	8
GSE62452	PDAC-Human	130	69	61
GSE55643	PDAC-Human	53	45	8

We first removed samples and genes which have missing or incomplete expression values for each dataset. We applied the Robust Multi-array Average (RMA) normalization method to each dataset individually. Then we merged datasets using common gene symbols and obtained 466 samples and 10106 genes in the end. Since the datasets were originally generated using different experimental protocols and platforms across multiple laboratories, their integration introduces batch effects. To mitigate these inter-dataset discrepancies, we employed the removeBatchEffect function from the Limma R package. We splitted original data using five fold cross validation as 80% train(372 samples, 116-normal, 256-tumor) and 20% test(94 samples, 29-normal, 65-tumor).

2.2. Feature Selection

We first performed the Differential Gene Expression (DGE) analysis to identify genes with significant changes in expression between tumor and normal samples. Genes with an adjusted p-value < 0.05 and |log2 fold change| > 2 were assumed to be differentially expressed. This analysis resulted in the encompassing both upregulated and downregulated genes associated with PDAC. To further refine the feature space and enhance the relevance of the input for generative modeling, we employed an ensemble-based feature selection strategy. This approach integrates multiple feature selection techniques to capture robust and complementary information. We implemented the following methods:

- ANOVA F-value using SelectKBest with the f_classif scoring function to measure the variance between class labels.
- Mutual Information (MI): Captures non-linear dependencies between features and class labels.
- Lasso Regression: Performs regularization-based feature selection by shrinking less relevant coefficients toward zero.

For each method, features were ranked based on their importance scores, and the ranks were aggregated. Final feature selection was based on a percentile threshold (top 25%) of the combined scores, resulting in a robust set of predictive features. Finally, the DEGs and ensemble-selected features were combined for generative modeling and downstream analysis.

2.3. Generative Models

GAN [13] consists of two neural networks, one is generator, and the other one is discriminator. The generator aims to produce realistic synthetic data from random noise. The discriminator evaluates the synthetic data as real or fake and calculates loss to give feedback to the generator. Networks learn together. CTGAN is a GAN-based model specifically designed for synthesizing tabular data by addressing handling discrete and continuous variables through a conditional generator.

Variational Autoencoder [14] also consists of two neural networks, encoder and decoder. A variational inference-based encoder learns from input data and encodes it to the latent space while keeping the most important variables. A decoder aims to reconstruct latent variables by learning a continuous probabilistic representation of latent space. TVAE was constructed by modifying the loss function.

Gaussian Copula, a statistical approach that models dependencies between variables using copulas, generating synthetic data by transforming marginal distributions into a Gaussian space and sampling from a multivariate normal distribution.

We used Synthetic Data Vault (SDV) library to use ctGAN, TVAE and GC. To address the class imbalance in the dataset, we generated 140 normal samples, as this was the underrepresented class, using generative models.

2.4. Classifiers

We employed two widely used machine learning classifiers Random Forest (RF) and Support Vector Machine (SVM) to evaluate the impact of synthetic data on classification performance.

Random Forest is an ensemble learning method that consists of multiple decision trees and output is obtained with majority voting of the individual trees. It handles high-dimensional data effectively like gene expression, is robust to overfitting due to its averaging nature, and can model complex interactions between features.

Support Vector Machine, on the other hand, is a supervised learning algorithm that seeks to find the optimal hyperplane that maximizes the margin between classes. SVM is particularly effective in high-dimensional spaces and has high generalization ability. We used SVM with a radial basis function (RBF) kernel.

2.5. Evaluation Metrics

Comparison between real and synthetic samples were made using statistical metrics described below:

Correlation Discrepancy (CD) calculates the difference of Pearson correlation between synthetic and real data and gets the mean discrepancy as a summary metric. Lower values are better in terms of realistic synthetic data generation.

$$CD = mean(\left| R(real) - R(synthetic) \right|) \tag{1}$$

Average Kolmogorov-Smirnov (KS) statistic measures the maximum difference between the cumulative distribution functions (CDFs) of real and synthetic data. *KS(real[i], synthetic[i])* represents the maximum difference between the CDFs of feature *i* in real and synthetic data. KS produces values in the interval [0,1], where 1 is better, 0 is worst.

$$Average\ KS = \frac{1}{n}\sum_{i=1}^{n}(1 - KS(real[i], sytnhetic[i]) \tag{2}$$

Average Statistical Similarity (SS) function from sdmetrics python library calculates mean, median and standard deviation of real and synthetic data. SS produces values between [0,1], where 1 is better, 0 is worst.

$$Average_{SS} = \frac{SS_{mean(real,\ synthetic)} + SS_{std(real,\ synthetic)} + SS_{median(real,\ synthetic)}}{3} \tag{3}$$

To assess the performance of classifiers, we employed standard classification metrics:

Accuracy measures the proportion of correctly classified instances among all instances.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + TN} \tag{4}$$

Precision quantifies the proportion of correctly predicted positive instances among all instances predicted as positive.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall measures the proportion of positive instances from all the actual positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

The F1-score is the harmonic mean of precision and recall.

$$F1\ Measure = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \tag{7}$$

3. Results and Discussion

Generative models aim to produce samples that resemble real data. Therefore we employed several statistical metrics to compare synthetic data generated by CTGAN, GC and TVAE. Table 2 summarizes the results of CD, KS and SS metrics.

Table 2: Comparison synthetic and independent test data

	CD↓	KS↑	SS↑
Original Train - Test	0.1399	0.8361	0.9603
CTGAN - Test	0.4699	0.6971	0.8746
TVAE - Test	0.3736	0.7522	0.9330
GC - Test	**0.1482**	**0.8120**	**0.9529**

According to Table 2, the GC model generated synthetic data that is most similar to the real data. Its results for the CD, KS, and SS metrics closely match with the original train:test similarity comparison, indicating high fidelity of the synthetic cells. Therefore, the GC model performed best in terms of generating more realistic samples, followed by TVAE, while CTGAN yielded the poorest results.

Figure 2 shows the distribution of synthetic samples alongside the combined original train and test samples. The distribution of synthetic samples generated by the GC model is uniform and closely matches to that of the real data. In Figure 2-B, TVAE shows a similar distribution, although not as uniform as GC. In contrast, CTGAN produces a noticeably different distribution that does not resemble or mingle with the real data.

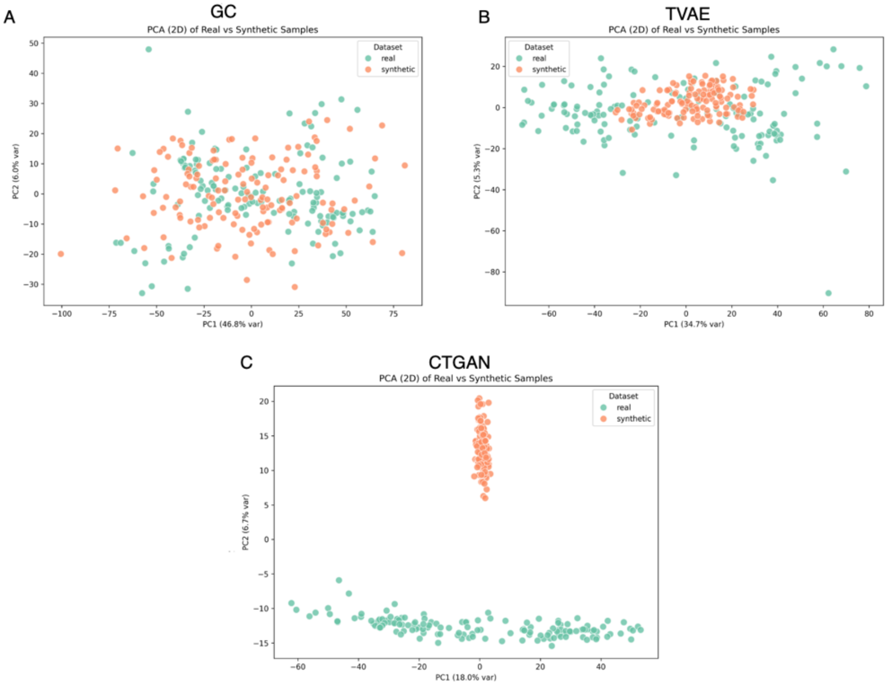

Fig. 2: Distribution of synthetic and real(train+test) data (A) GC (B) TVAE (C) CTGAN

Marker genes are crucial in cancer detection studies; Therefore, synthetic samples should preserve similar expression levels for such marker genes. Several PDAC marker genes have been reported in the literature. We explored GeneCards [15] to identify marker genes that are also present in our integrated dataset. Violin plots were then generated for the selected marker genes' expression profiles to assess the similarity between the real (test data) and the synthetic samples. Figure 3 presents violin plots for each marker gene, and the Mann–Whitney statistical test was applied to detect significant differences. A significant result (p-value < 0.05) indicates that the expression levels between the compared expressions differ significantly. While non-significant results do not confirm similarity, they allow us to visually compare the overall shape of the violin plots. In cases of observing significant differences, we can conclude that the expression of the corresponding marker gene differs meaningfully between the real and synthetic samples (so the corresponding generative model fails at generating realistic samples). CTGAN shows significant differences in the expression of the marker genes *CPA1, ACTA2, COL1A1,* and *VIM*. TVAE exhibits significant differences with *KRT19* and *VIM*, while GC differs significantly only for one marker gene, *VIM*. These results indicate that the expression of certain marker genes may not be well preserved in the synthetic samples. However, visual inspection suggests that GC and TVAE generally maintain expression levels similar to the real (test) data.

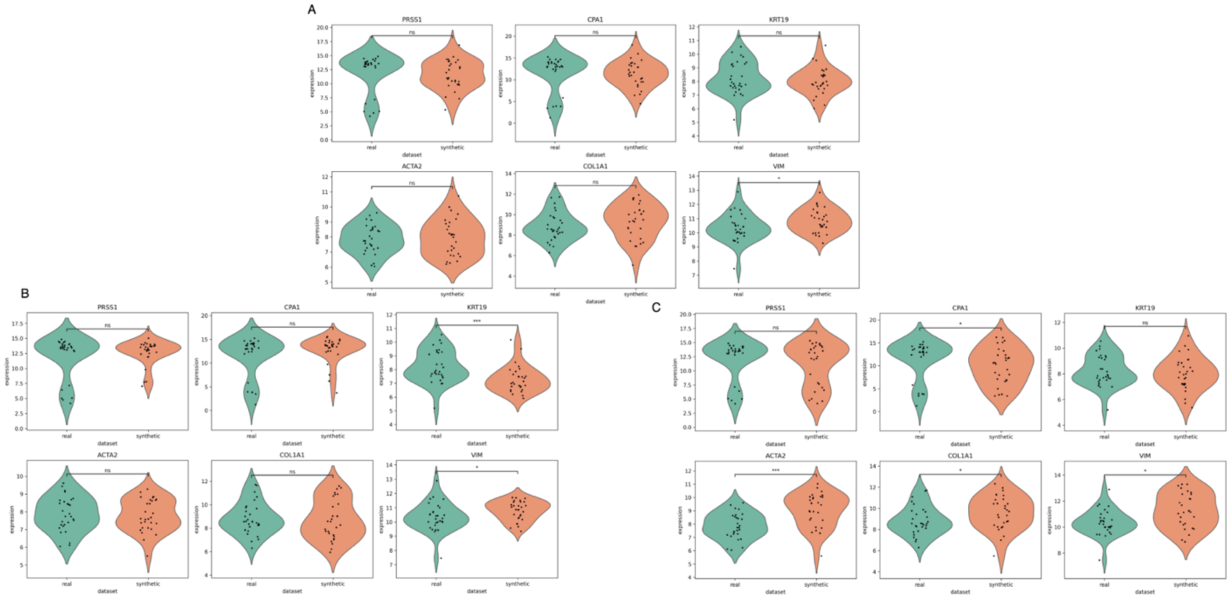

Fig. 2: Violin plots of PDAC Marker Genes (A) GC (B) TVAE (C) CTGAN

We conducted further classification experiments to evaluate the impact of synthetic samples on predictive model performances. The training set was constructed by combining original training samples with synthetic samples, and an independent test set was used for evaluation and test set was not used for data generation to avoid data leakage. We applied 5-fold cross-validation to assess model robustness. Additionally, we proposed an ensemble generative model that combines synthetic samples generated by the TVAE and GC models. Due to the results of previous analyses, the CTGAN model was not included in the ensemble model. 'Original' refers to the case where only the original samples used and no synthetic samples considered. Acknowledging the success of GC model, we sampled 70% of synthetic samples using GC and the remaining 30% using TVAE for the ensemble approach. Table 3 presents the classification results. TVAE performed well with the tree-based classifier Random Forest (RF), while GC achieved better results with the Support Vector Machine (SVM). Therefore, we combined their strengths in an ensemble model. Although the ensemble model did not outperform GC for SVM, it showed highly promising results overall. Increasing the number of synthetic samples may further enhance performance and better reveal differences between the models.

Table 3: 5-Fold CV Results

	RF			
Model	Precision	Recall	F1 Measure	AUC
Original	0.8380	0.8165	0.8251	0.9204
ctGAN	0.8431	0.8344	0.8371	0.9150
TVAE	0.8450	0.8416	0.8407	0.9188
GC	0.8326	0.8410	0.8337	0.9180
Ensemble	**0.8549**	**0.8623**	**0.8568**	**0.9246**
	SVM			
Original	0.8432	0.7929	0.8098	0.9122
ctGAN	0.8418	0.8002	0.8143	0.9087
TVAE	0.8394	0.8237	0.8291	0.9169
GC	0.8538	**0.8627**	**0.8553**	**0.9238**
Ensemble	**0.8541**	0.8570	0.8533	0.9236

4. Conclusion

We conducted a comparative study on creating synthetic gene expression data using tabular data-based generative models: CTGAN, TVAE, and Gaussian Copula (GC). The performance of these models was evaluated both statistically and biologically. First, we curated six PDAC gene expression datasets to create integrated dataset. An ensemble-based feature selection approach was applied, combining Differential Expression Gene (DEG) analysis, ANOVA, Lasso, and Mutual Information (MI). To address the class imbalance, we generated synthetic normal samples to match the number of tumor samples in the integrated dataset.

The quality of synthetic data was initially assessed using common statistical metrics including Correlation Discrepancy (CD), Kolmogorov-Smirnov (KS), and Statistical Similarity (mean, median, std), where the GC model demonstrated the highest similarity to the real data. We also performed a visual evaluation by illustrating the distribution of real and synthetic samples in a 2D principal components (PC) space, in which GC showed the most uniform and realistic distribution where original and synthetic samples mixed together.

For biological relevance, we examined the expression levels of PDAC marker genes. The GC model, followed by TVAE, exhibited expression patterns most similar to the real data.

Finally, we assessed the impact of synthetic data on classification performance. Training sets were created by combining synthetic and real training samples, while an independent test set was retained for final model evaluation. We also proposed an ensemble model that integrates synthetic samples from both GC and TVAE. Using SVM and Random Forest (RF) classifiers, the ensemble model achieved the best performance with RF and also showed promising precision with SVM. Although it was not the top-performing model for SVM, the ensemble approach demonstrates strong potential overall.

References

[1] İr. Acer, F. O. Bulucu, S. İÇer, and F. LatiFoğlu, ‘Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations’, *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 31, no. 1, pp. 112–125, Jan. 2023, doi: 10.55730/1300-0632.3974.

[2] H. Hayashi, N. Uemura, K. Matsumura, L. Zhao, H. Sato, Y. Shiraishi, YI. Yamashita, H. Baba, ‘Recent advances in artificial intelligence for pancreatic ductal adenocarcinoma’, *World J Gastroenterol*, vol. 27, no. 43, pp. 7480–7496, Nov. 2021, doi: 10.3748/wjg.v27.i43.7480.

[3] A. Kiran and S. S. Kumar, ‘A Comparative Analysis of GAN and VAE based Synthetic Data Generators for High Dimensional, Imbalanced Tabular data’, in *2023 2nd International Conference for Innovation in Technology (INOCON)*, Mar. 2023, pp. 1–6. doi: 10.1109/INOCON57975.2023.10101315.

[4] Y. Wang, Q. Chen, H. Shao, R. Zhang, and H. Shen, ‘Generating bulk RNA-Seq gene expression data based on generative deep learning models and utilizing it for data augmentation’, *Computers in Biology and Medicine*, vol. 169, p. 107828, Feb. 2024, doi: 10.1016/j.compbiomed.2023.107828.

[5] M. Eltager, T. Abdelaal, M. Charrout, A. Mahfouz, M. J. T. Reinders, and S. Makrodimitris, ‘Benchmarking variational AutoEncoders on cancer transcriptomics data’, *PLOS ONE*, vol. 18, no. 10, p. e0292126, Oct. 2023, doi: 10.1371/journal.pone.0292126.

[6] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, ‘Modeling Tabular data using Conditional GAN’, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019. Accessed: May 15, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

[7] T. Idichi, N. Seki, H. Kurahara, K. Yonemori, Y. Osako, T. Arai, A. Okato, Y. Kita, T. Arigami, Y. Mataki, Y. Kijima, K. Maemura, S. Natsugoe, ‘Regulation of actin-binding protein ANLN by antitumor miR-217 inhibits cancer cell aggressiveness in pancreatic ductal adenocarcinoma’, *Oncotarget*, vol. 8, no. 32, pp. 53180–53193, Aug. 2017, doi: 10.18632/oncotarget.18261.

[8] L. Li, JW. Zhang, G. Jenkins, F. Xie, EE. Carlson, BL. Fridley, WR. Bamlet, GM. Petersen, RR. McWilliams, L. Wang, ‘Genetic variations associated with gemcitabine treatment outcome in pancreatic cancer’, *Pharmacogenet Genomics*, vol. 26, no. 12, pp. 527–537, Dec. 2016, doi: 10.1097/FPC.0000000000000241.

[9] R. Janky, MM. Binda, J. Allemeersch, A. Van den Broeck, O. Govaere, JV. Swinnen, T. Roskams, S. Aerts, B. Topal, ‘Prognostic relevance of molecular subtypes and master regulators in pancreatic ductal adenocarcinoma’, *BMC Cancer*, vol. 16, p. 632, Aug. 2016, doi: 10.1186/s12885-016-2540-6.

[10] J. Jiang, AC. Azevedo-Pouly, RS. Redis, EJ. Lee, Y. Gusev, D. Allard, DS. Sutaria, M. Badawi, OA. Elgamal, MR. Lerner, DJ. Brackett, GA. Calin, TD. Schmittgen., ‘Globally increased ultraconserved noncoding RNA expression in pancreatic adenocarcinoma’, *Oncotarget*, vol. 7, no. 33, pp. 53165–53177, Aug. 2016, doi: 10.18632/oncotarget.10242.

[11] S. Yang, P. He, J. Wang, A. Schetter, W. Tang, N. Funamizu, K. Yanaga, T. Uwagawa, AR. Satoskar, J. Gaedcke, M. Bernhardt, BM. Ghadimi, MM. Gaida, F. Bergmann, J. Werner, T. Ried, N. Hanna, HR. Alexander, SP. Hussain, ‘A Novel MIF Signaling Pathway Drives the Malignant Character of Pancreatic Cancer by Targeting NR3C2’, *Cancer Res*, vol. 76, no. 13, pp. 3838–3850, Jul. 2016, doi: 10.1158/0008-5472.CAN-15-2841.

[12] S. Lunardi, NB. Jamieson, SY. Lim, KL. Griffiths, M. Carvalho-Gaspar, O. Al-Assar, S. Yameen, RC. Carter, CJ. McKay, G. Spoletini, S. D'Ugo, MA. Silva, OJ. Sansom, KP. Janssen, RJ. Muschel, TB. Brunner, ‘IP-10/CXCL10 induction in human pancreatic cancer stroma influences lymphocytes recruitment and correlates with poor survival’, *Oncotarget*, vol. 5, no. 22, pp. 11064–11080, Nov. 2014, doi: 10.18632/oncotarget.2519.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, ‘Generative Adversarial Nets’, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2014. Accessed: May 15, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

[14] D. P. Kingma and M. Welling, ‘Auto-Encoding Variational Bayes’, Dec. 2013, Accessed: May 15, 2025. [Online]. Available: https://openreview.net/forum?id=33X9fd2-9FyZd

[15] M. Safran, N. Rosen, M. Twik, R. BarShir, T. Iny Stein, D. Dahary, S. Fishilevich, and D. Lancet, ‘The GeneCards Suite’, in *Practical Guide to Life Science Databases*, I. Abugessaisa and T. Kasukawa, Eds., Singapore: Springer Nature, 2021, pp. 27–56. doi: 10.1007/978-981-16-5812-9_2.