EVENFLOW

Robust Learning and Reasoning for Complex Event Forecasting

EVENFLOW White Paper

The Personalised Medicine Use Case of EVENFLOW

www.evenflow-project.eu

WHITE PAPER

The Personalised Medicine Use Case of EVENFLOW

Date:	December 9th, 2025
Authors:	Guillermo Prol-Castelo, Davide Cirillo, Alfonso Valencia

Executive Summary

Cancer is a global health problem, affecting millions of people worldwide. However, many types of cancer and cancer subtypes exist, so when stratifying different cancer states and conditions, the number of available samples in biomedical datasets for each case tends to dwindle. In this work, we summarise the Personalised Medicine Use Case contributions to the EVENFLOW project, which concentrates on the generation of synthetic data in highly specific scenarios. We developed a methodology for Synthetic Data Generation based on the Variational Autoencoder (VAE), a Deep Learning technique for variational inference.

By leveraging the generative capabilities of the VAE, we have applied it in two different scenarios. The first one is a static scenario, where we study the boundaries of subgroups in medulloblastoma, a cancer that affects children. By doing so, we were able to characterise an underrepresented subgroup, which requires differentiated treatment from other similar subgroups.

The second scenario is dynamic, considering cancer stages. Cancer is not a static disease, as it is able to grow and invade near or even distant tissues. Here, we generate trajectories of synthetic patients as they progress from early to late stages of the disease.

Both cases are helpful to more accurately determine cancer attributes where data availability is low. These findings may help pave the way toward more personalised treatments in the considered cancers, as well as contribute to the generation of robust and reliable synthetic data in a critical yet general scenario such as biomedicine.

Table of Contents

List of Figures

Definitions, Acronyms and Abbreviations

Acronym/Abbreviation	Title
TNM	Tumour Node Metastasis
SDG	Synthetic Data Generation
GSEA	Gene Set Enrichment Analysis
TCGA	The Cancer Genome Atlas
VAE	Variational Autoencoder
AI	Artificial Intelligence

Term	Definition
Variational Autoencoder (VAE)	Deep Learning technique that creates embeddings from unlabeled data by considering a variational Bayesian inference to learn a probabilistic latent space.
Medulloblastoma	Childhood tumour of the cerebellum

1 Introduction

1.1 Biology of Cancer

Cancer is a disease where body cells start to divide uncontrollably and may spread beyond their point of onset. This definition is very wide. Moreover, cancer can manifest in a variety of ways. Some types, like solid tumours, affect tissues such as skin, while others, like leukemia, are liquid and impact the blood's white cells. Cancers can target different tissues, vary in their invasiveness, and affect distinct populations, for example, children versus adults. They may arise at their original site (primary cancer) or spread to another location (secondary cancer), among other variations. Hence, cancer is not one disease, but many. For reference, the National Cancer Institute (NCI) lists more than 100 types and explicitly states there are over 100 distinct cancers (https://www.cancer.gov/types).

Each cancer type has its own subtypes. These may depend on multiple factors such as the onset site, patient demographics, prognosis, etc. In EVENFLOW, we are particularly interested in defining cancer subtypes using molecular signatures derived from gene-level data. Different genes may affect different patients, resulting in a variety of symptoms. In any disease, it is important to know which genes are more or less active compared to normal conditions, in order to develop specific treatments.

Adding to the complexity, cancers evolve over time. They generally start in a non-invasive state, known as benign tumours. However, they may begin invading tissues beyond their starting point. To characterise this, different systems have been developed. One of the most common ones is the TNM (Tumour Node Metastasis), which defines a series of stages tumours go through based on their size and the involvement of lymph nodes (Figure 1). This way, a non-invasive tumour is at stage 0, and when it becomes invasive, it transitions to stage I. At that point, it may continue growing and invading nearby tissues. It may even be able to go on and invade other organs or body parts, known as metastasis, and the final stage IV. At the point of metastasis, the chances of survival become slim.

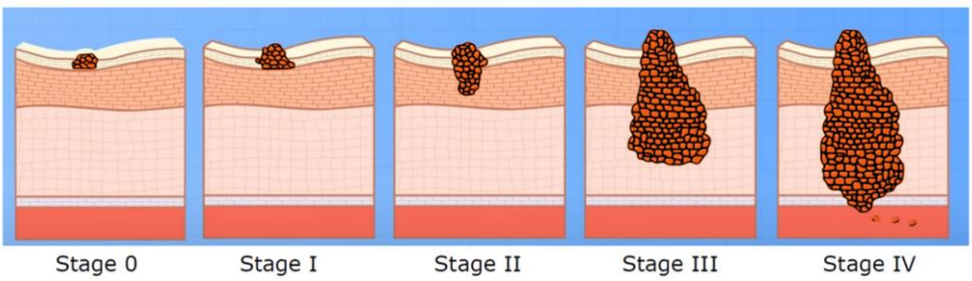

Figure 1: Cancer Stages. Cancer may start as a benign, non-invasive cluster of cells, at stage 0. Once cancer becomes invasive, it is known as a malignant tumour, with an associated larger stage (I-III) as more nearby tissues are invaded. At Stage IV, or metastasis, cancer has made its way into the blood stream and can spread throughout the body. Source: Cleo Abram, YouTube, youtu.be/USmu0scNcSs

1.2 Problem Statement

Despite advances in cancer diagnosis and prevention, there is limited understanding of tumour heterogeneity across subtypes and stages. Effective treatment requires precise tumour characterisation, but the availability of patient samples is low when stratified by tumour type, subtype, stage, or demographics. This scarcity of data hinders research and limits the ability to develop personalised treatment strategies. With EVENFLOW technology, we can address this challenge by expanding and analysing samples across cancer subtypes and stages, enabling more comprehensive insights and tailored treatment approaches.

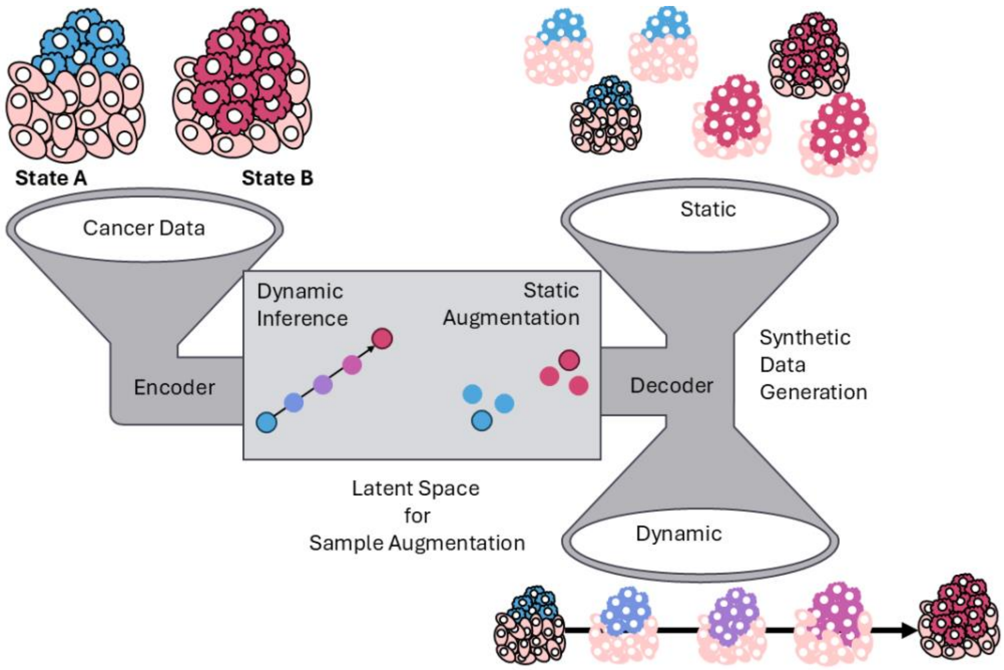

Figure 2: Summary of the Synthetic Data Generation (SDG) pipeline applied to cancer data. The VAE, which consists of two parts, an encoder and a decoder, is the centerpiece of our SDG methodology. The encoder takes in biological cancer data from different states or conditions (colored in blue and dark pink) and is able to compress it into a space of lower dimensionality, the latent space. This space may be sampled by slightly altering the original representations (dots with a dark border) either in a directed way between dots, i.e., dynamic inference, or by simply augmenting the number of dots, to produce new representations (dots without a border). In either case, the decoder is then able to produce new, synthetic samples in the original space.

2 Synthetic Data Generation (SDG)

2.1 Cancer Data

The completion of the Human Genome Project in 2003 enabled the study of biological processes and diseases at the molecular level through informatics tools. There are many kinds of large-scale data available that contain biological information, known as omics. When the omics focuses on RNA is called transcriptomics data, which may contain information of individual cells, or a collection of them. We focus on the latter, called bulk-RNASeq, from cancer tissue. This data helps researchers understand the regulation of genes in cancer, i.e., which processes are fundamental in this disease.

The production of omics data, however, requires specialised equipment and training. Therefore, there have been significant efforts in making these datasets available to the wider scientific community in an open way. Here, we highlight two open-access data repositories, which are the ones used during the project. The Gene Expression Omnibus (GEO) is a large database where any researcher can make their omics data available and hence includes a large collection of experiments on many different diseases. The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/ccg/research/genome-sequencing/tcga), on the other hand, was a project that specifically focused on cancer and was finalised in 2018, producing a large dataset on many different types of cancer across thousands of patients.

However, even among these large datasets, when considering very specific cancer types, subgroups, or stages, the number of available samples rapidly decreases. It is for this reason that we turn to generative AI to engineer a pipeline to augment the number of available samples in these highly specific scenarios.

2.2 Generating Synthetic Data

Synthetic data is any artificial data that has been generated artificially instead of originating from real-world observations. In silico, synthetic data can be generated from statistical methods or from Artificial Intelligence (AI) methods. In our case, we have resorted to the Variational Autoencoder (VAE) [REF-01], an AI model with generative capabilities. The VAE is based on Bayesian statistics, inferring from the input data a low-dimensional embedding, called the latent space. It does so in two steps. First, a neural network known as the encoder takes the input data and learns the latent space. A second neural network, the decoder, then reconstructs the latent space into the original input data space. Through many iterations, the model can progressively learn and adapt the latent space to make it more suitable for encoding and decoding the original data.

The latent space itself is useful due to its reduced dimensionality. For this reason, it has been extensively used in biomedical research, but its generative abilities have remained underexplored in this field [REF-01]. The VAE is nonetheless adequate for this task. It is able to learn non-linear relationships within complex datasets, such as those used in transcriptomics data [REF-02]. For this reason, we have created a Synthetic Data Generation (SDG) process that is based on the VAE. This process is described in Figure 2. The VAE is fed data from different cancer patients that may display different cancer states, such as a subtype or stage. The VAE is unsupervised, meaning it does not directly learn from labels describing the state of cancer, only the gene expression table data. This way, the inferred latent space retains the characteristics of the original data, reflecting differences among samples. At the same time, the decoder has learned how to reconstruct the embeddings in the latent space back to the original space. The latent space is fully accessible, meaning we can not only place in it each of the original samples, but we can also sample new points and use the decoder to project those new points into the original space. This way, it is possible to generate new, synthetic samples that resemble the original ones.

During the project, we have implemented this pipeline in two different scenarios. The first one is static data augmentation. Knowing the position in the latent space of the real samples, we can sample near those samples in the latent space and reconstruct them with the decoder to augment the number of available samples of a given cancer subgroup. We applied this to the case of medulloblastoma subgroups (see section 3). The second scenario concerns dynamic data generation. Instead of sampling around a point, we can define a source and a target point in the latent space, both from real samples. We can then trace a line between both points and use the decoder to reconstruct the intermediate cases. We have applied this dynamic ability to replicate the cancer stage transitions in kidney cancer (see section 4).

3 Medulloblastoma

3.1 Biology of Medulloblastoma

Medulloblastoma is a childhood tumour that affects the cerebellum and has an embryonic origin [REF-03]. The first diagnosis most commonly occurs between ages 6 and 8 but may happen anywhere from birth to adulthood. Even though it is one of the most common central nervous system tumours, it is a rare disease, affecting 5 in every million individuals.

Medulloblastoma has historically been divided into four subgroups: SHH, WNT, G3, and G4 [REF-04]. WNT and SHH are well differentiated from each other and from G3 and G4. However, G3 and G4 share more similarities, to the point where there is a debate on whether they may contain different, intermediate subgroups or whether they may even be part of the same continuum. However, patients showing intermediate characteristics are rare cases.

It is crucial to correctly diagnose patient subgroups given the differences in prognosis. G3 patients show the worst prognosis among all subgroups: less than 60% survival 5 years after diagnosis [REF-03]. G4, on the other hand, is milder. Consequently, different groups are assigned treatments with varying aggressiveness. If a patient is assigned an incorrect subgroup, the given treatment could also be incorrect. By detecting an intermediate patient, we could reduce the risk of incorrect treatment assignment and help define a more suitable treatment instead.

3.2 Static Data Augmentation in Medulloblastoma

Our contribution to medulloblastoma centers around exploring the boundaries of the G3 and G4 subgroups [REF-05]. We determined the dubious cases: those patients that lay right in between both groups, which we labeled as G3-G4. Even though we have used the largest transcriptomic dataset available from medulloblastoma to date [REF-04], G3-G4 patients were very few compared to the other subgroups, we used our SDG pipeline to augment the number of G3-G4 samples, balancing them out with the samples in G3 and G4.

We compared the classification performances of the synthetically balanced data, comparing two subgroups (G3 and G4) and three (G3, G4, and G3-G4). The results show very similar performances across different metrics. This experiment suggests that the generated G3-G4 samples are significantly different from G3 and G4. Given that the fine distinction between subgroups has been kept, we argue that the three-subgroup scenario is fairer. Besides, G3-G4 shows important differences in the expression of certain key genes such as KMT2C, MYC, SNCAIP, SYNCRIP, and TP53. Together, these facts contribute to the hypothesis that G3-G4 should be considered as a subgroup with differentiated characteristics from G3 and G4.

The developed methodology is expandable to research scenarios prone to data scarcity, even if the data is highly specific. Our method is especially crusial when studying rare diseases and when dealing with underrepresented patients. Effectively, this work hence extends the principles of responsible AI by addressing a critical, clinical concern that is treatment discrimination.

4 Kidney Cancer

4.1 Biology of Kidney Cancer

Renal tumours invade large vessels by forming tumour thrombi [REF-06]. This is a significant difference compared to other tumours, and it is reflected in transcriptomics data, allowing the study of differences between tumour stages. Commonly, stages I and II may be grouped together as an early stage, and III and IV as a late stage. However, the characterisation of the different stages of cancer remains a challenge [REF-07]. The Cancer Genome Atlas (TCGA) is a large dataset that contains the molecular information of hundreds of patients, as well as the stage at diagnosis. However, patients are not followed up through the different stages. Our objective is to generate a synthetic dataset that replicates a longitudinal stage progression of kidney cancer patients.

4.2 Dynamic Trajectory Inference in Kidney Cancer

We simulate gene-expression trajectories by leveraging our SDG pipeline, trained on real TCGA data. This model can generate synthetic intermediate points between an origin (here, an early-stage patient) and a destination (a late-stage patient). In this way, we generated 50 pseudotime-points for each gene expression trajectory. In parallel, we trained a classifier on the real TCGA data and proceeded to apply it to the synthetic trajectories. This reveals a correct gradual classification, starting from assigning early trajectory points to early stages. As the pseudotime advances, the classification gradually reverses, with a mix of results in the middle points, until it ends up switching to the late-stage assignment in the final pseudotime points, as expected. This phenomenon of pseudo-time-dependent stage classification indicates that the synthetic data reflect meaningful stage transitions.

Our collaboration with NCSR Demokritos in the framework of this use case enabled probabilistic risk estimates of stage advancement from incomplete temporal information of the synthetic trajectories through symbolic finite-state automata (SFA) and Complex Event Forecasting for temporal pattern recognition. These methods allowed for the early prediction of worsening cancer conditions.

Our synthetic data was further assessed through Gene Set Enrichment Analysis (GSEA). This is a methodology that establishes the activity of a known biological function, composed of a set of known genes. When applied to our data, GSEA returned a pseudo-time dependent activity of many biological functions, which are also in agreement with results found in studies of cancer in general [REF-08] and kidney cancer in particular [REF-09].

All in all, our methodology can turn a static dataset into a dynamic database, while retaining the key biological characteristics of cancer stages, and is validated through standard methodology, assessed on independent studies.

5 Conclusion

In the Medicine Use Case of the EVENFLOW project, we have applied an SDG pipeline based on the VAE to extend its usability in biomedical data by reliably producing synthetic data that mimics the original data. We have applied this methodology in two cases. In a static case, we have augmented the number of samples of a rare case within medulloblastoma to determine an intermediate subgroup between the recognised G3 and G4 subgroups. In a dynamic scenario, we generated intermediate patients between the early and late stages of kidney cancer. These are important steps towards the development of more personalised treatments for specific subgroups and stages of cancer.

Our approach presents, however, certain limitations. The most obvious is that there is no straightforward way to validate the generated data. Future work should consider closer collaboration with applied biological researchers to determine the viability of the generated samples. Moreover, in the dynamic case, we have been able to assess the synthetic trajectories between early and late stages because kidney cancer stages are particularly separable with transcriptomics data, a rare case among tumours. We recommend that future studies focus on following up patients as they progress through several cancer stages.

We envision that our work will set the foundations for further application of synthetic data in biomedicine. Other generative models, such as optimal transport, have already been implemented for data augmentation and trajectory inference in single-cell data [REF-10]. However, we should establish patient-centric methodologies that are able to closely mimic general cancer behaviour.

6 References

[REF-01]	Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
[REF-02]	Prol-Castelo, Guillermo, Davide Cirillo, and Alfonso Valencia. "10 Years of Variational Autoencoder: Insights from Cancer Temporal Progression Studies, a Systematic Literature Review." bioRxiv (2025): 2025-05.
[REF-03]	Way, Gregory P., and Casey S. Greene. "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders." In PACIFIC SYMPOSIUM on BIOCOMPUTING 2018: Proceedings of the Pacific Symposium, pp. 80-91. 2018.
[REF-04]	Northcott, P.A., Robinson, G.W., Kratz, C.P. et al. Medulloblastoma. Nat Rev Dis Primers 5, 11 (2019). https://doi.org/10.1038/s41572-019-0063-6
[REF-05]	Cavalli, Florence MG, Marc Remke, Ladislav Rampasek, John Peacock, David JH Shih, Betty Luu, Livia Garzia et al. "Intertumoral heterogeneity within medulloblastoma subgroups." Cancer cell 31, no. 6 (2017): 737-754.
[REF-06]	Prol-Castelo, Guillermo, Alejandro Tejada-Lapuerta, Beatriz Urda-García, Iker Núñez-Carpintero, Alfonso Valencia, and Davide Cirillo. "Exploring the Boundaries of Medulloblastoma Subgroups with Synthetic Data Generation." bioRxiv (2024): 2024-12.
[REF-07]	Kim, K., Zhou, Q., Christie, A. et al. Determinants of renal cell carcinoma invasion and metastatic competence. Nat Commun 12, 5760 (2021). https://doi.org/10.1038/s41467-021-25918-4
[REF-08]	Benitez Fuentes JD, Morgan E, de Luna Aguilar A, et al. Global Stage Distribution of Breast Cancer at Diagnosis: A Systematic Review and Meta-Analysis. JAMA Oncol. 2024; 10(1):71–78. doi:10.1001/jamaoncol.2023.4837
[REF-09]	Gaude, E., Frezza, C. Tissue-specific and convergent metabolic transformation of cancer correlates with metastatic potential and patient survival. Nat Commun 7, 13041 (2016). https://doi.org/10.1038/ncomms13041
[REF-10]	Morris, M., Latif, F. The epigenetic landscape of renal cancer. Nat Rev Nephrol 13, 47–60 (2017). https://doi.org/10.1038/nrneph.2016.168
[REF-11]	Bunne, C., Schiebinger, G., Krause, A. et al. Optimal transport for single-cell and spatial omics. Nat Rev Methods Primers 4, 58 (2024). https://doi.org/10.1038/s43586-024-00334-2

This project is funded by the EU under Horizon Europe Grant Agreement No 101070430.
Views and opinions expressed are those of the EVENFLOW consortium only and do not necessarily reflect those of the EU or the European Commission, neither the EU nor the granting authority can be held responsible for them.