SNPs-Analysis Variational Autoencoder (SAVE): A Semi-supervised Learning Approach for SNP-based Bleomycin-induced Toxicity Prediction

Authors: *Beh Qian Jun*[*], *Emmanuel Vikran*[*], *Than Hui Xin*[*] **Supervisors**: *Assistant Prof Folefac Aminkeng*[†] **Mentors**: *Lim Ting Wei*,[*] *Bharath Shankar*[*]

[*]*Special Programme in Science, National University of Singapore* [†]*Yong Loo Lin School of Medicine, National University of Singapore*

1. Introduction

- Drug-Induced Toxicity in Healthcare is caused by multiple factors.
- In our study, we try to build a tool that predicts Bleomycin-Induced toxicity (BIT) in patients administered
- Bleomycin is a common chemotherapy drug administered for Hodgekin's Lymphoma, a type of white blood cell cancer.[1]
- However, approximately 10% of patients experience BIT, with 3% resulting in fatality.[2]
- Multiple genetic factors have been stated to cause BIT.[2]
- Therefore, there is a need for an accurate method for **predicting BIT** for public health purposes based on multiple genetic and clinical factors.

2. Challenges

- Collected data of 127 patients administered with Bleomycin and their genetic markers.

Low sample size & High-dimensionality

Imbalanced Dataset

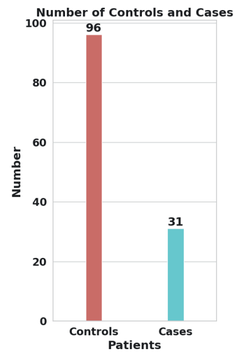
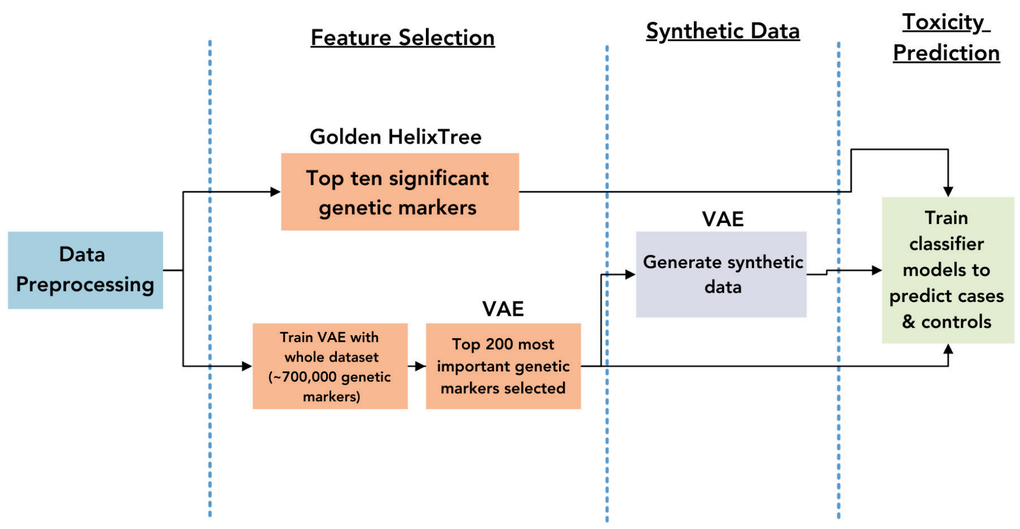


Figure 1: Bar plots of the number of patients with toxicity (cases) and the number of patients without toxicity (controls).

3. Materials & Methods

We propose a pipeline as such:



- **Classifier Model** is a ML algorithm that predicts if an observation belongs to a particular class.
- **Golden HelixTree** is a software for analysing population genetics.

Variational Autoencoder (VAE)

- A machine learning (ML) algorithm that can learn to **create new data** that is similar to the data it was trained on.
- **Reduce the dimensionality** of the input data, by extracting out the important features representative of the data.
- **Mitigate the problems of low sample size and imbalanced datasets** by generating synthetic data.

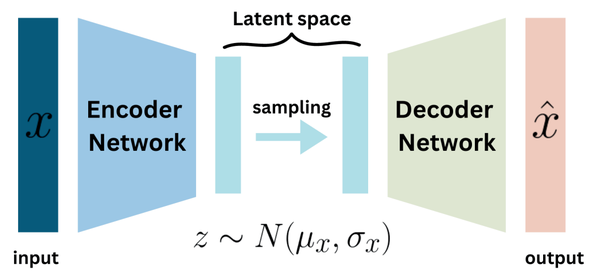


Figure 2: A simple visual representation of the conventional VAE, in which the blocks' sizes correspond to the dimensionality of the data at each stage. The dimension of the input data is being significantly reduced in the lower-dimensional space, also known as the latent space.

4. Results & Discussions

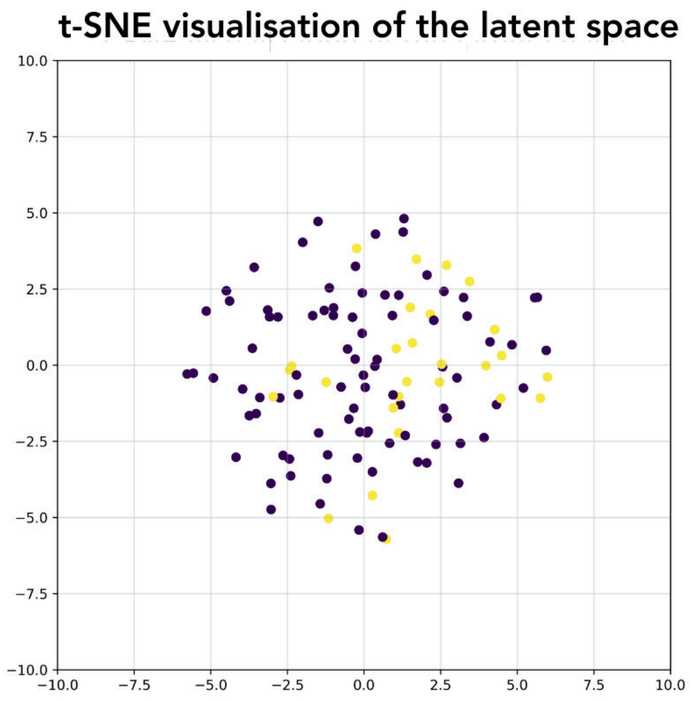
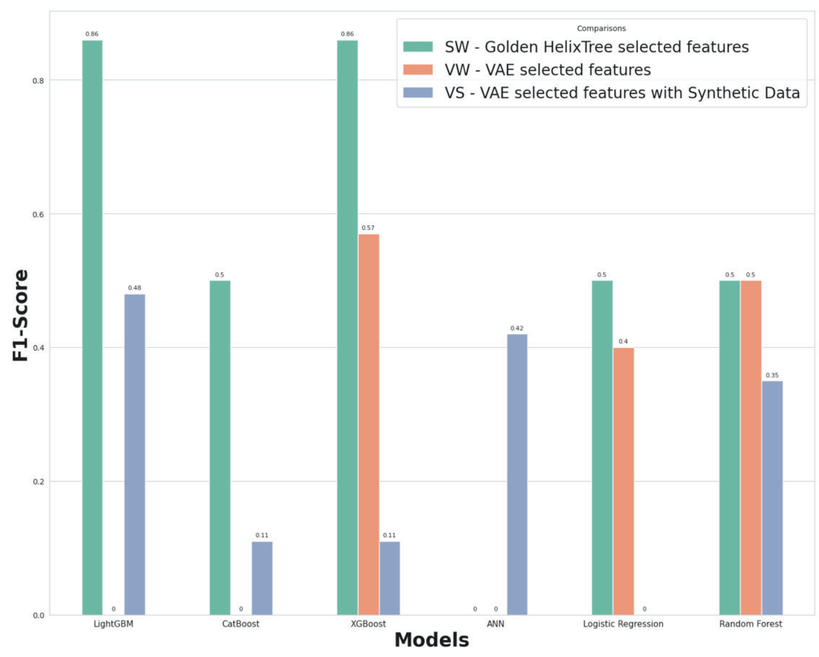


Figure 3: VAE's latent space trained with the hyperparameters = {Learning rate: 0.000612; Batch size: 78}. The yellow data points indicate the case, while the blue represents the control.

Figure 4: F1-Scores of the six classifier models to predict cases in the three variants of data: SW - Golden HelixTree selected features, VW - VAE selected features and VS - VAE selected features with synthetic data



- The data points are generally clustered, signifying that they are somewhat realistic.
- However, as the points are intermingled and are not cleanly distinguished as their two classes, F1-scores were used to evaluate model performance as it accounts for both precision and recall.
- High precision: Individuals are not wrongly classified as having BIT (cases) when they do not (controls).
- High recall: Individuals with BIT are being identified by the models.

- The classifier models performed poorly on the VS dataset, with the F1-scores for all models < 50%, likely due to the VAE model performing below expectations.
- The performance of the models trained on the VW dataset was worse than with the synthetic data: Despite generating synthetic data that did not fully capture the underlying structure of the dataset, the sheer increase in the number of data points improved the performance of the classifier models.
- However, the models trained on the SW dataset acquired the best performance, with F1-Scores of 86% for two models — LightGBM and XGBoost.

5. Conclusions

We are able to build a tool that is able to predict BIT. These results demonstrate that the proposed pipeline for generating synthetic data and feature selecting using the VAE to combat a small, imbalanced dataset with high dimensionality is limited. Using the Golden HelixTree software for feature selection can build more meaningful models to predict if a patient is likely to have BIT.

6. Future Work

Improvements on the Pipeline

- Use Supervised-VAE (SVAE)[3]
- Feed the input genetic markers to separate encoders, according to the respective chromosome number.
- Improved hyperparameter tuning.

Applications of the Pipeline

- Personalised medicine: Customised treatment based on an individual's susceptibility to BIT.
- Clinical trials: Screen participants in trials for their susceptibility to BIT.
- Drug development: Modification of the existing drugs to make them safer.
- Public health: Identification of populations with a higher risk of BIT.

7. References

(1) Hodgkin lymphoma - Latest research and news | Nature. https://www-nature-com.libproxy1.nus.edu.sg/subjects/hodgkin-lymphoma (accessed 2023-04-14).
(2) Ge, V.; Banakh, I.; Tiruvoipati, R.; Haji, K. Bleomycin-induced Pulmonary Toxicity and Treatment with Infliximab: A Case Report. Clin Case Rep 2018, 6 (10), 2011–2014. https://doi.org/10.1002/ccr3.1790.
(3) Ji, T.; Vuppala, S. T.; Chowdhary, G.; Driggs-Campbell, K. Multi-Modal Anomaly Detection for Unstructured and Uncertain Environments. arXiv December 15, 2020. https://doi.org/10.48550/arXiv.2012.08637.