Liverpool John Moores University

Using Generative Topography Mapping to risk stratify breast cancer patients post successful treatment.

by

Christopher Ravenscroft

A thesis submitted in partial fulfillment for the
degree of Masters of Data Science

in the
Faculty of Science
Astrophysics Research Institute

September 2024
Word Count 9774 excluding Lists of Tables, List of Figures,
Contents, References and Tables

Liverpool John Moores University

Abstract

Faculty of Science
Astrophysics Research Institute

by Christopher Ravenscroft

There exists a range of negative outcomes post treatment of patients for breast cancer. This study aimed to develop a novel risk stratification model, built on Generative Topographic Mapping and Random Forest architecture, calculating risks of recurrence and death post treatment. The model was built using current breast cancer data from Clatterbridge Cancer Centre. The GTM performed well on the breast cancer dataset, with analysis identifying six distinct phenotypes, split mainly on TNM stages, ages and treatment drugs. These were effectively categorised in low, medium and high risk groups, and consequently random forest model were trained. Predictive accuracy varied, with high risk groups performing well with high negative outcome recall rates (0.52-0.70) and F1 Scores(0.49 - 0.61). Lower risk groups struggled with predicting the negative class. Finally, the approach suggested a novel combination using the probability of belonging to negative outcome classes to identify at risk individuals in a test dataset. This presents a promising idea for a model for personalised post breast cancer treatment management, although further model refinement and clinical consultation is needed prior to any clinical application.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In 2024, breast cancer remains as the most common cancer worldwide, and remains as the leading cause of mortality amongst female cancer patients. In 2020 alone, there was over 2.3 million diagnosed cases, making up 11.7% of total cancer diagnosis. This corresponds to 6.9% of all cancer mortality, despite majority affecting women[51]. Furthermore, recurrence in breast cancer is extremely high. Research indicates that even after achieving disease-free status for 10 to 32 years, patients still face a 16.6% chance of recurrence [41]. This combined problem of high initial diagnosis rates highlights the critical importance of continued post-clearance care. The monitoring of patients, and ability to find susceptible patients is important in preventing unnecessary negative outcomes. One method for this would be through the application of machine learning is for the risk stratification of patients [11].

This thesis has three main objectives in the hope of developing a novel risk stratification method for breast cancer patients post successful treatment of the cancer. The first is to implement Generative Topographic Mapping (GTM) [7] onto the breast cancer dataset. By using the GTM, we wish to accurately model and visualise the data, highlighting the potential utility of the framework, and discussing its success. Using the clusters found, the second objective is to perform detailed analysis of each cluster, identifying the phenotypes which exist within the data. This will be done based on the compositions of different variables within each cluster. Using this, we will then use random forest to build predictive models for recurrence and death post treatment. Using a combination of the random forest models and GTM, we will develop of novel risk stratification method, which assigns a score to the individual.

Finally, the thesis will discuss the implications of the findings in a clinical setting. We will suggest future research directions, and explain the utility of our model. By achieving this objectives, and explaining our implications, we aim to demonstrate the effectiveness

of GTM and ensemble methods in breast cancer research, and highlight the utility of machine learning tools in oncology risk assessment models.

Chapter 2

Literative Review

2.1 Important Features

There exists numerous shared risk factors for the development and progression of breast cancer [36] [36]. Physical activity, tobacco usage, obesity levels and diet are all key risk factors of not only breast cancer [30], but other dangerous diseases such as cardiovascular diseases (CVD) [43]. Whilst other factors play key roles, these relations show potential scenario for individuals, due to their susceptibility to these shared risk factors, to face the dual threat of developing both diseases concurrently or sequentially. In fact, in older women, cardiovascular disease emerges as a more significant cause of mortality in breast cancer patients than the cancer itself. [40]. This shows the importance of considering other co-morbidities that the patient struggles with, such as CVD. Other co-morbidites can impact treatment outcomes, for example mental health issues are known to be furthered by cancer treatment, and also decreases likelihood of attending following up consultations [9].

Breast cancer treatments, while crucial for combating the disease, can inadvertently lead to cardio toxicity. Notably, anthracycline-based chemotherapy regimens, such as doxorubicin have been associated with an increased risk of cardiac events, including cardiomyopathy and heart failure [45]. On the other hand, there also exists drugs which have shown preventative measures. Clinical trials involving women with advanced breast cancer have demonstrated that the administration of dexrazoxane, a cardioprotective agent, significantly reduces the overall incidence of cardiac events when administered prior to doxorubicin infusion [14]. However, whilst methods such as selective treatment, screening for heart issues, and checking family history are now common place, these do not offer complete safety from cardiovascular complications.

Furthermore, Vinca alkaloids such as vincristine are known to induce peripheral neuropathy, a condition which causes damage to the peripheral nerves, leading to symptoms such as numbness and pain in the hands and feet. This nephropathy can effect between 35% to 45% of patients undergoing treatment with these drugs[27]. The condition can result in significant functional disabilities, impacting fine motor skills and walking, and may persist long after treatment ceases, potentially worsening. This neurotoxicity is dose-dependent, with symptoms typically appearing within the first three months of treatment [60]. These highlight how specific treatments can lead to negative side effects, which may increase mortality rates even after successfully treating the cancer.

Breast cancer incidence is notably higher in women aged 50 and older, with a rate of 375.0 per 100,000 compared to 42.5 per 100,000 in women under 50. However, due to the larger proportion of younger women in the population, 23% of breast cancer cases occur in women under 50. Among women aged 35 and younger, African American women have slightly higher breast cancer incidence rates than White women, but this trend reverses in older age groups. Furthermore, despite lower total incident rates, African American women are shown to face a 37% higher death rate than White women, even though their incidence rates are lower[49]. Similarly, obesity not only increases the risk of developing breast cancer, particularly in post-menopausal women, but it also significantly impacts the risk of recurrence and mortality, making weight management a crucial component of breast cancer prevention and survivorship strategies [26]. Along with this, obesity disproportionately affects ethnic minorities, with African Americans and Mexican American women experiencing higher rates than their white counterparts [13]. This shows the differences between races when considering impacts of breast cancer treatment.

In identifying key features, reducing the dataset is crucial. In not relevant variables, there exists bias which may skew the model. Another study highlighted the importance of feature selection in breast cancer risk prediction, finding that Random Forest and Support Vector Machine Recursive Feature Elimination (SVM-RFE) were effective. Random Forest, in particular, was emphasized for its performance in improving model accuracy, making it valuable for developing accurate predictive models [31].

2.2 Prediction Models

A 2023 study by Wu et al [58] aimed to evaluate the effectiveness of various machine learning model in predicting the prognosis of breast cancer patients. Using the Surveillance, Epidemiology, and End Results (SEER) database, 10 different non deep learning models were used on an large group of 63145 patients, aiming to establish the best

model. They found the Multivariate Adaptive Regression Spline (MARS) model to be most effective for predicting five year prognosis of the disease, with a AUC of (0.831), and F1-Score 0.608, along with high sensitivity and specificity. This was a useful study which highlighted the predictive ability of a range of models. A further study by Parikh et al [39] explored the usage of machine learning algorithms in predicting short term mortality of all cancer patients. They effectively utilised electronic health records to identify patients at high risk of mortality within six months. Achieving observed mortality rates of 51.3% in the high risk group, and 3.4% in the low risk group, it shown best results with random forest modelling, and highlighted the potential of using machine learning models in risk stratifying patients,

Recurrence prediction models in breast cancer are more difficult to predict. [1]. There are various examples able to predict high accuracy such as a 2012 study by Ahmad et al [2], which compares various methods such as decision trees and support vector machines. However whilst reporting high accuracies, fail to mention the imbalanced classes, and disclose any other metric of evaluation, highlighting issues in taking accuracy as only metric. Further studies have shown attempts of applying deep learning methods to predict early recurrence in breast cancer by analyzing histopathological images. These models have shown promise in identifying features that correlate with recurrence risk, achieving competitive prediction accuracy's comparable to established clinical markers like estrogen receptor status and tumor grade [46].

Along with this, while not the primary aim of this thesis, our research has observed strong diagnostic accuracy in breast cancer prediction using machine learning models applied to both health records and image analysis [4, 32]. This further shows the utiity of machine learning within oncology.

2.3 Clustering

A further study by Shukla et al [47] in 2018 aimed to highlight the usability of clustering technique for breast cancer patients. Again using the SEER dataset, initially the data was mapped onto a lower dimensional space, using Self Organizing Map (SOM) [33]. Following this, DBSCAN, a automatic clustering algorithm which can detect both non-normal shaped clusters and determine the number of clusters itself, was applied [18]. This was able to find 9 unique clusters, each representing different combinations of the variables used. Furthermore, it showed the ability to then train different MLPs on each cluster, with each cluster showing higher predictive accuracy than without clustering methods. This shown the usage of individual models for each cluster.

Another study by Belcuig et al [5] showed the effectiveness of SOM on predicting recurrence with breast cancer. Using tumour details such as perimeter and area, they were able to achieve a 72% accuracy on test set. Moreover, a study conducted on gene expression data utilized k-means clustering in combination with neural networks and random forests. This approach improved the classification of high-risk cases by identifying differentiating genes, demonstrating the potential of clustering to enhance predictive models in breast cancer research[44].

2.4 GTM

GTM has been effectively utilized in various clustering tasks, demonstrating its utility in the oncology domain. A notable application involved collaborative clustering using the Wisconsin Diagnostic Breast Cancer dataset. In this study, GTM was employed to create topographic maps, which were then clustered using the Expectation-Maximization (EM) algorithm for Gaussian Mixture Models (GMMs). This method allowed for effective clustering, highlighting GTM's ability to manage complex datasets with multiple features, such as those found in breast cancer research [50]. However, it is quite under utilised in cancer generally. GTM has made a recurrence in research recently with various studies highlighting it's usage in healthcare and bio informatics. One study by Andrade et al [3] used GTM to visualise and map motor unit action potentials and compared its performance with other clustering methods like Self-Organizing Maps and GMMs Other studies experiment with it's utility in biological pathways [19].

2.5 Risk Stratification and Machine Learning

Risk stratification plays a crucial role in the care of oncology patients by identifying those at multiple levels of risks for various complications or negative outcomes, enabling further personalised treatment and additional care. Various risk stratification models have shown utility within the oncology . One example is in the use of managing cardiovascular toxicity in cancer patients, assessing related side effects and guiding preventative treatment[6, 53]. Other examples include in the case of febrile neutropenia, risk stratification models like the Multinational Association for Supportive Care in Cancer (MASCC) score and the Clinical Index of Stable Febrile Neutropenia (CISNE) are employed to identify patients at low risk for serious complications, allowing for safe outpatient treatment [28].

Further advancements in machine learning have shown improvements in risk stratification, offering more precise predictions and personalized treatment plans. One example

of this is in high-grade serous ovarian cancer. A multi model ML approach was developed to integrate clinical imaging and genomic data for risk stratification. This model combines radio-logical, histopathology, and clinicogenomic data to improve predictive ability of the model. The use of various data types helped make various decisions about maintenance therapy [8]. One example in prostate cancer, a radiomic-based ML model was developed using magnetic resonance imaging (MRI) features to stratify patients into different risk categories. This model employed several ML approaches, including support vector machines, logistic regression, and random forest classifiers, with the random forest model demonstrating the best predictive performance (AUC = 0.87) for high-risk patients [38]. Similarly, in the context of immunotherapy, an ML-based risk stratification model was created to predict mortality in cancer patients treated with atezolizumab. This model, which utilized algorithms such as extreme gradient boosting and random forest, effectively stratified patients into high and low-risk groups, providing valuable insights for treatment management [23]. Additionally, a pan-cancer risk prediction model was developed using routine health check-up data, employing ML techniques like Lasso-Cox, Random Survival Forests (RSF), and eXtreme Gradient Boosting (XGBoost). This model demonstrated the ability to predict the risk of developing various cancers by analyzing demographic and lifestyle variables, highlighting the potential of ML in broad cancer risk assessment [52]. These examples underscore the diverse applications of ML in cancer risk stratification, showcasing its potential to revolutionize personalized oncology care.

2.6 Limits of Previous Research

From prior research, we see that the SEER dataset is commonly used in various cancer prediction models. There is various reasons for this. The SEER database is incredible robust for populations aged 65 and older, is publically available and has a comprehensive, complete dataset on various cancer details. Furthermore, it recently expanded between 2011 to 2015, now including 857,056 cancer patients and 601,470 non cancer controls, proving a large resource for cancer research[17]. Despite this, there are various limitations to the database. One is the demographic make up of the dataset. It is solely U.S popuilation, and it has an over-representation of foreign born and urban inhabitants, as well as particular ethnic groups [59]. Along with this, it has been extensively used in various breast cancer prediction models, which highlights the over-reliance on the database and the need to explore different available datasets, to enhance robustness of predictive models and a more diverse understand of breast cancer prognosis. Along with this, we see commonly issues with metrics chosen to judge models on. Often accuracy is taken as the key metric, as it's high scores in imbalance classes provides a false sense

of effective modelling. We will aim to fix upon this, by choosing alternative metrics to judge our model on.

On the other hand, commonly SOM is selected as the choice of clustering. Whilst it is a powerful method for clustering, it does present several limitations. One being the lack of a well-defined objective function. On the other hand, GTM is a similar model to SOM built with intentions of improving on some of the issues of it, and since SOM has proven effective, we could expect to see similar results. GTM has a clear defined objective function and is probabilistic in nature. This may give more easily interpreted results than SOM does, and enable for easier comparison of the effectiveness of the clustering method.

Chapter 3

Data

3.1 Overview

The data for this study was sourced from the NHS Clatterbridge Cancer Centre, an NHS trust based in Merseyside, England, specialized on the treatment of referred patients with all types of cancer [12]. As a leading cancer treatment center, NHS Clatterbridge offers comprehensive care and advanced therapies to improve patient outcomes. In 2016, the center transitioned to a new, sophisticated data storage system designed to enhance the management and accessibility of patient information. Consequently, the available data predominantly spans from 2016 to the present, except for specific cases of migration in which prior patient information is needed, for example in recurrence cases. Whilst this limits available data, this ensures the dataset reflects current treatment practices and outcome information.

3.1.1 Inclusion Criteria

To ensure data completeness, various inclusion criteria were specified as to ensure this.

1. Primary Diagnosis of cancer is breast cancer.
2. Patient has successfully completed treatment, and acknowledged as cleared.
3. Patient has complete entries for each variable.
4. Minimum 3 years from 1st August 2021 since treatment completed treatment for adequate follow-up.

In this study, one particular limiter was the choice to ensure that each variable had an entry, and all patients with incomplete variables were removed. Whilst this limited the sample size, we implemented this decision to avoid imputation which in turn helped prevent make assumptions with patient data. By making sure the results were all observed data, it kept the integrity of the model, and reduced bias.

Data which did not meet the minimum 3 years of follow up was kept to use as example data for the application of unseen data in the risk stratification model. Figure 3.1 shows the process of the inclusion criteria, and how many patients exist at each stage.

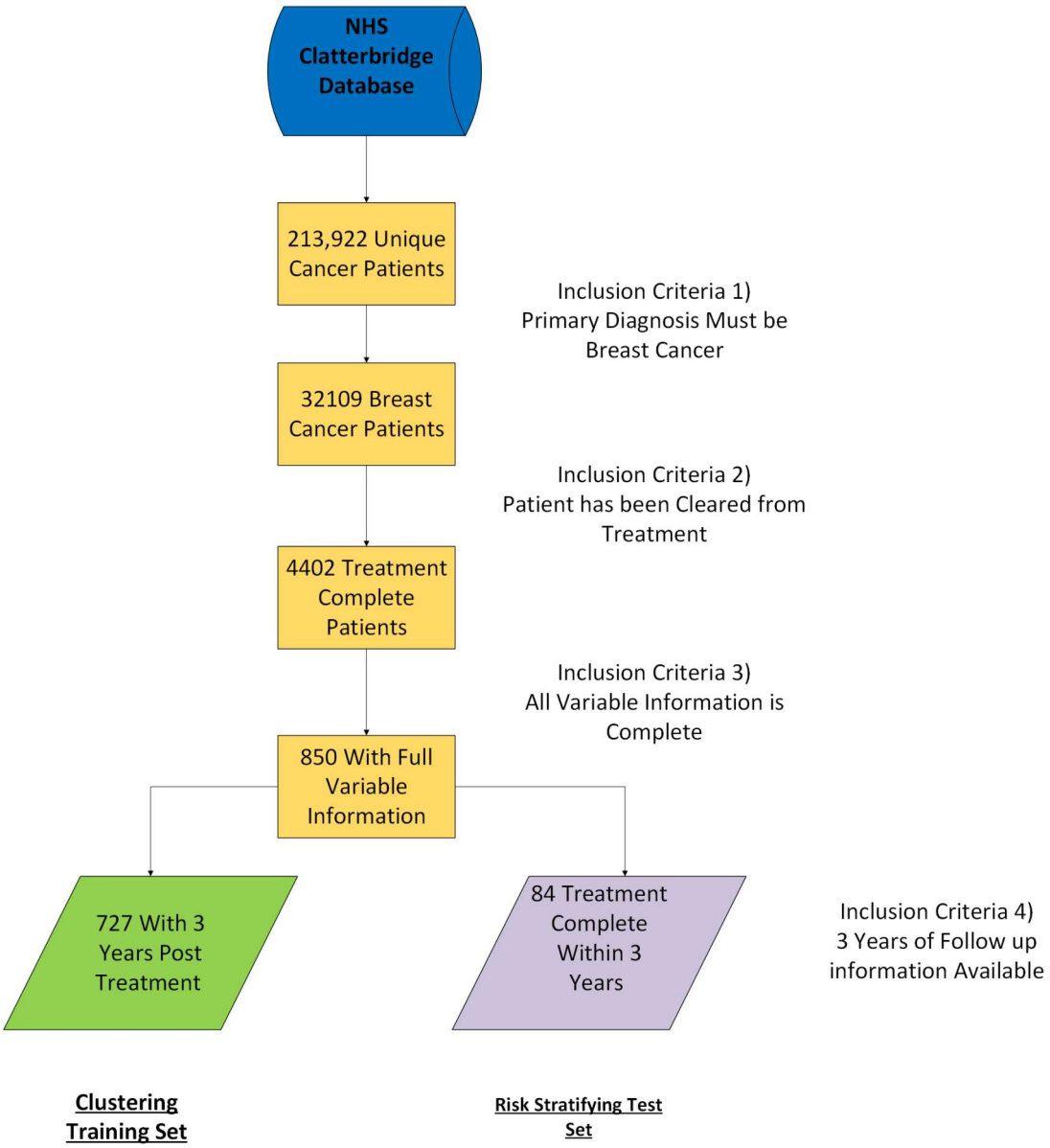

FIGURE 3.1: Inclusion Criteria Flow Diagram

3.2 Dataset Creation

The dataset for this study was divided into several key sections: Patient Demographics, Treatment Drugs, Tumor Details, Blood Tests, Treatment Side Effects, and Outcomes. These broad categories were chosen to investigate based on frequent usage in prior literature, which ensures comparability with current research. Additionally, they were chosen for the availability and completeness they presented within the database.

3.2.1 Outcomes

In our investigation, we focused on two key outcomes: death and recurrence following treatment. For the purpose of this study, recurrence encompasses all forms of cancer reappearance, including both localized and distant recurrence. All types of recurrence were studied collectively, and referred to here on just as recurrence. These were recorded as 0 for not occurred and 1 for occurred, differentiating between our two groups. Furthermore, death included all form mortality. With 65 cases of mortality, and 105 cases of recurrence, this makes up a small subset of our data, and hence we will expect to see imbalances classes.

3.2.2 Patient Demographics

In the demographic data, we split the ethnicities into different categories, and obtain four different age groups, 0-24, 25-49, 50-74, and 75+. However, in our data set there only existed values within the 25-49, 50-74 and 75+. groups. Hence in further results they will be known as age_1, age_2 and age_3 respectively. Furthermore, we take the bmi of the patient, which is taken from the initial nurse consultation. Figure 3.7 discussed later showcases the distributions of these variables.

3.2.3 Tumour Details

The tumours within the system are categorised with the traditional TNM system[21]. The T value refers to the size of the primary tumour. The T values range from TX and T0 for unmeasured and not found respectively, upwards towards 4 depending on the size and extent of the primary tumour [22]. For simplification and ensure of suitably sized tumour groups, we took only the full stage, and not further deviations, for example T3b would be recorded just as T3.

N refers to the nearby lymph nodes and whether the cancer has travelled to them, which is follows a similar scale as to T values. The M category denotes whether there is distant metastasis. M0 indicates no distant metastasis, while M1 indicates that metastasis has occurred. All values are these are taken as of last update prior to being medically cleared. We also took the laterality of the tumour, being on the left, right or both breasts.

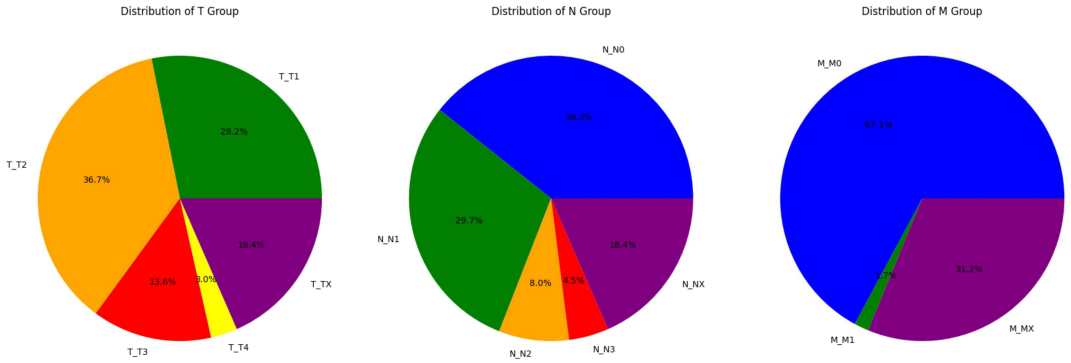

FIGURE 3.2: Pie Charts showcasing the TNM Stages

3.2.4 Chemotherapy Drugs

The chemotherapy drugs section is a boolean list of commonly used chemotherapy drugs. If the treatment is used over the period of their treatment for breast cancer, it is 1 for true, else 0. This was collected by sorting for most commonly used drugs, and filtering the database for these. These were largely different sizes between drug groups, with Cyclophosphamide having only 24.0, Carboplatin having small with 56, Docetaxel a larger group of 114 and then a huge proportion given to Paclitaxel, 465. This showcases the common drug usage within the trust.

3.2.5 Blood Tests

The blood tests section consists of various routine blood checks, taken at various points doing the treatment. These include the likes of white blood cell counts, alkaline phosphate levels and other relevant tests. We took the mean values for each of these results in the blood tests. Figure 3.3 showcases a pairplot of the blood tests. We see each blood test is normally distributed, and notice some interesting correlations, such as between Urea and RBC.

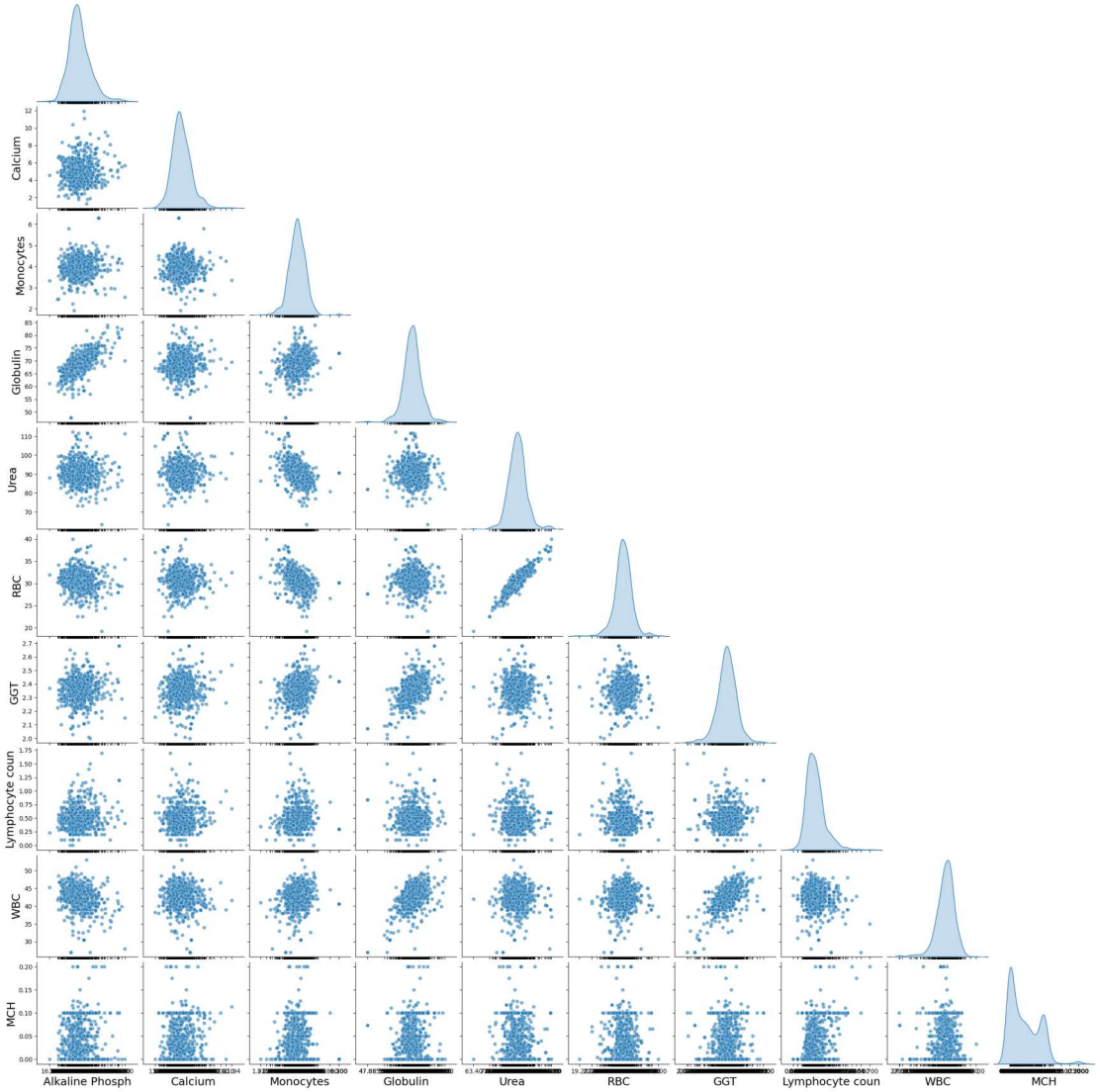

FIGURE 3.3: Pairplot of Blood Tests

3.2.6 Side Effects

During chemotherapy and other treatments, patients are regularly assessed for various side effects, such as the severity of nausea and the degree of alopecia , recorded on a scale from 0 to 4, with 0 indicating no experience of the side effect and 4 indicating the most severe experience. To gain a comprehensive understanding of how these side effects impact patients, I calculated the mean (μ), standard deviation (SD), and variance (σ^2) for each side effect.

The mean (μ) is computed as:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

where x_i represents each individual score and N is the total number of scores. The mean provides a central summary, showing the typical severity level of each side effect across all patients, helping to identify which symptoms are most prevalent and may require more attention.

The standard deviation (SD) measures the average distance of each data point from the mean and is calculated as:

$$\text{SD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

where $(x_i - \mu)^2$ is the squared deviation of each score from the mean. The SD provides insight into how much patients' experiences vary from this average, highlighting whether side effects are experienced consistently or if there is a wide range of responses among patients. A high standard deviation indicates significant variability.

Similarly, the variance (σ^2), which is the square of the standard deviation, is calculated as:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Variance underscores the degree of dispersion in the side effect scores, which is essential for understanding the overall diversity of patient experiences. High variance indicates a wide spread of side effect severities, helping to tailor treatment plans more effectively. By recording both the spread and the average, we gain more information on the impact on the patient.

3.2.7 Cleaning Variables

The dataset contained several issues with the continuous variables. As illustrated in Figure 3.6, these variables are often manually entered by staff, making them prone to errors. For instance, in the case of Alkaline Phosphate levels, some values were recorded on the order of e^{11}, which is clearly erroneous. Such mistakes needed to be corrected to ensure the integrity of the data.

To identify and remove outliers in our dataset, we utilized the Interquartile Range (IQR) method. First, we calculated the first quartile (Q_1) and the third quartile (Q_3) of the data. The IQR was then computed as the difference between the third and first quartile:

$$\text{IQR} = Q_3 - Q_1$$

A data point x was classified as an outlier if it fell outside the following bounds:

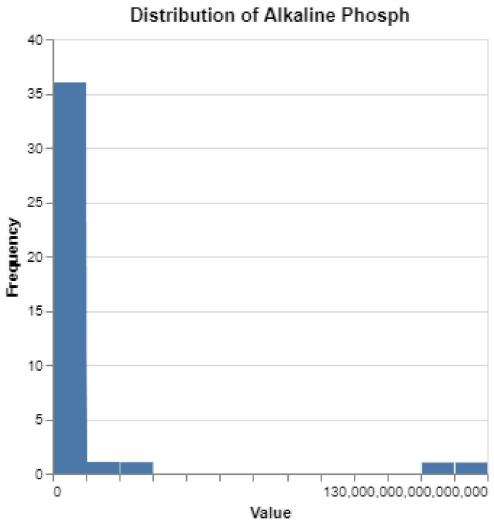

FIGURE 3.4: Pre Clean

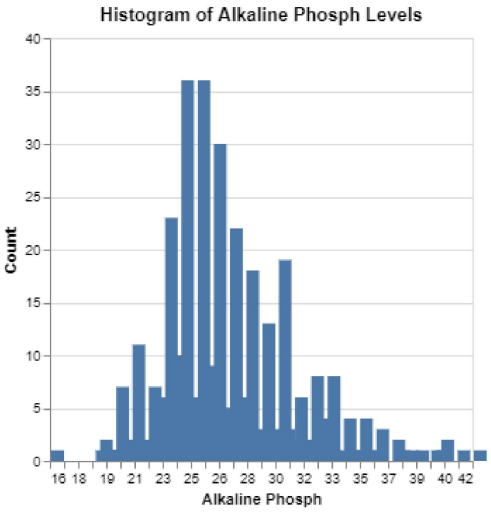

FIGURE 3.5: Post Cleaning

FIGURE 3.6: Distribution of Alkaline Phospate pre and post cleaning

$$\text{Lower Bound} = Q_1 - 3 \times \text{IQR}$$

$$\text{Upper Bound} = Q_3 + 3 \times \text{IQR}$$

Therefore, any data point satisfying either of the following conditions was considered an outlier:

$$x < Q_1 - 3 \times \text{IQR}$$

$$x > Q_3 + 3 \times \text{IQR}$$

This was mostly used just to identify the obviously incorrect values.

3.3 Data Transformations

One hot encoding is a technique used in machine learning to convert categorical variables into a binary matrix, where each column represents a unique category and contains a 1 for the presence of that category and 0 otherwise. This method ensures that categorical data is in a numerical format suitable for model training. Given that many models, including those used in GTM, rely on Gaussian basis functions, one hot encoding is particularly appropriate. It avoids implying any ordinal relationship or distance between the categories, aligning well with the Gaussian assumptions inherent in GTM. This

ensures that the encoded data accurately reflects the non-numeric nature of the original categorical variables, maintaining the integrity of the model's assumptions

The data will be scaled using the Standard Scaler function available in sci-kit learn package [42]. This scales the data using the following transformation

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

Where: x is the original feature value μ is the mean of the feature σ is the standard deviation of the feature z is the standardized value

This happens to each feature individually, rather than the dataset as a whole. This is particularly useful the GTM model, as it ensure each data point is distributed on a Gaussian with mean 0 and variance 1.

To minimize noise in the model, I decided to remove categorical variables that had entries for fewer than 1% of the data. This resulted in a significant reduction in the number of variables, particularly affecting the co-morbidities and treatment drugs categories. It was observed that co-morbidities were generally poorly recorded across the entire dataset. Additionally, there was a long list of drugs that were only used by one or two patients, contributing to the removal of many variables. I could have chosen to map these to a general variable but I decided not to pursue this approach, as it could dilute the specific effects of individual drugs and co-morbidities.

Category	**Columns Removed**
Patient Demographics	Sex, bmi_0
Co-Morbidities	COPD, Stomach_Ulcers, Hepatitis, HIV, LEARNING D
Drugs	Methotrexate, Fluorouracil, Doxorubicin, Cisplatin, Etoposide, Vincristine, Bleomycin, Epirubicin, Ifosfamide, Gemcitabine, Vinblastine
Other	T_T0

TABLE 3.1: Columns Removed

3.4 Ethics

In conducting this study, it is important to address several ethical considerations inherent in the use of real health care data. The dataset utilized is derived from actual patient records, which made sure to ensure the data was anonymous. Furthermore, it is crucial to recognize the demographic limitations of the model. Specifically, the data predominantly consists of over 99% white or unknown patients and is exclusively female.

Whilst it is known to be a female dominated disease, there still exists male patients, with 2800 diagnosis a year in the United States [48]. Furthermore, there exists no patients under the age of 25. Figure 3.7 highlights the disparities in the demographic categories. This significant demographic skew poses notable challenges. Particuarly, the under representation of non-white and male patients in the dataset introduces potential biases and limits the generalizability of the model's findings. When making the model, we were aware of these issues, and such any results obtained should not be applied to these groups.

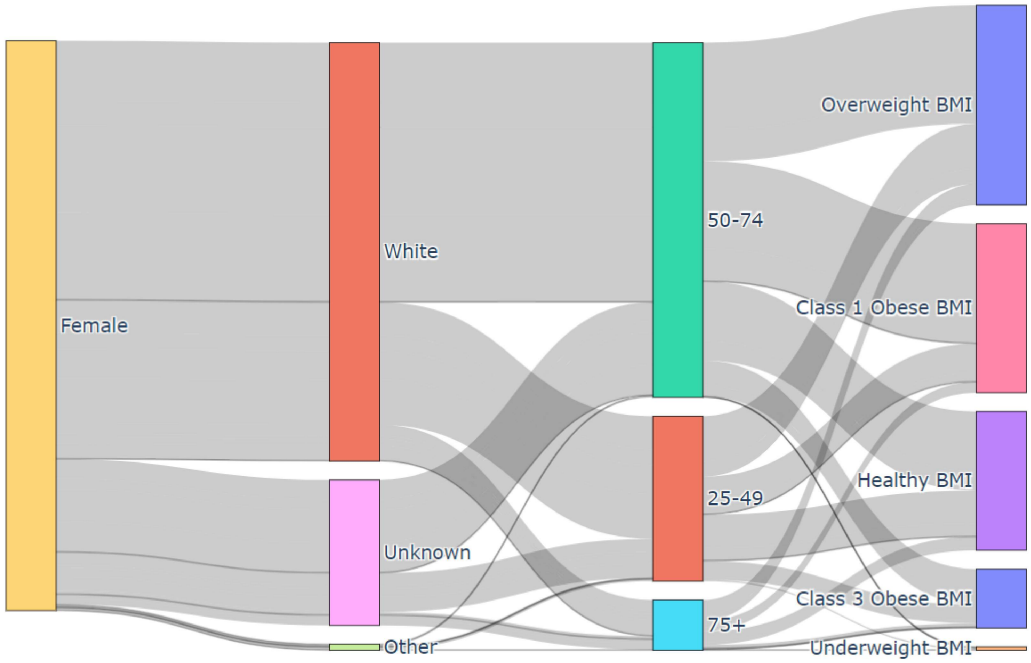

FIGURE 3.7: A Sankey Diagram Showcasing demographic categories

Chapter 4

Methodology

4.1 Overview

Figure 4.1 shows an overview of the methodology used in the research. We begin by pulling the data from the database, using a selective and thorough cleaning and transformation process, discussed in the Chapter 3. After preparing the data, we will fi tthe GTM model, and cluster the gtm using the Ward's linkage method. This will give us a number of unique clusters. We will analyse the clusters, comparing the distribution of all variables, particularly the death and recurrence outcome variables. We will use these to find our distinct subgroups of patients. From here, random forest models will be trained on each cluster to determine recall rates and predictive accuracy, whilst also obtaining the importance scores for different features. Finally, we will risk stratify using a combination of our models, and gain risk scores for each individual. This chapter will further expand on this methodology.

During the study, we used Python as the primary programming language, due to the extensive number of libraries available. For data manipulation, such as scaling and processing, we used the packages Pandas [56] and NumPy [24]. For machine learning tasks, such as developing our random forest model, we implemented Scikit-learn [42]. To build our GTM, we used the uGTM package, which allowed us to easily implement and adjust our GTM. Finally, for all plots we used a mixture of Matplotlib [25]and Altair [55], creating a variety of informative figures.

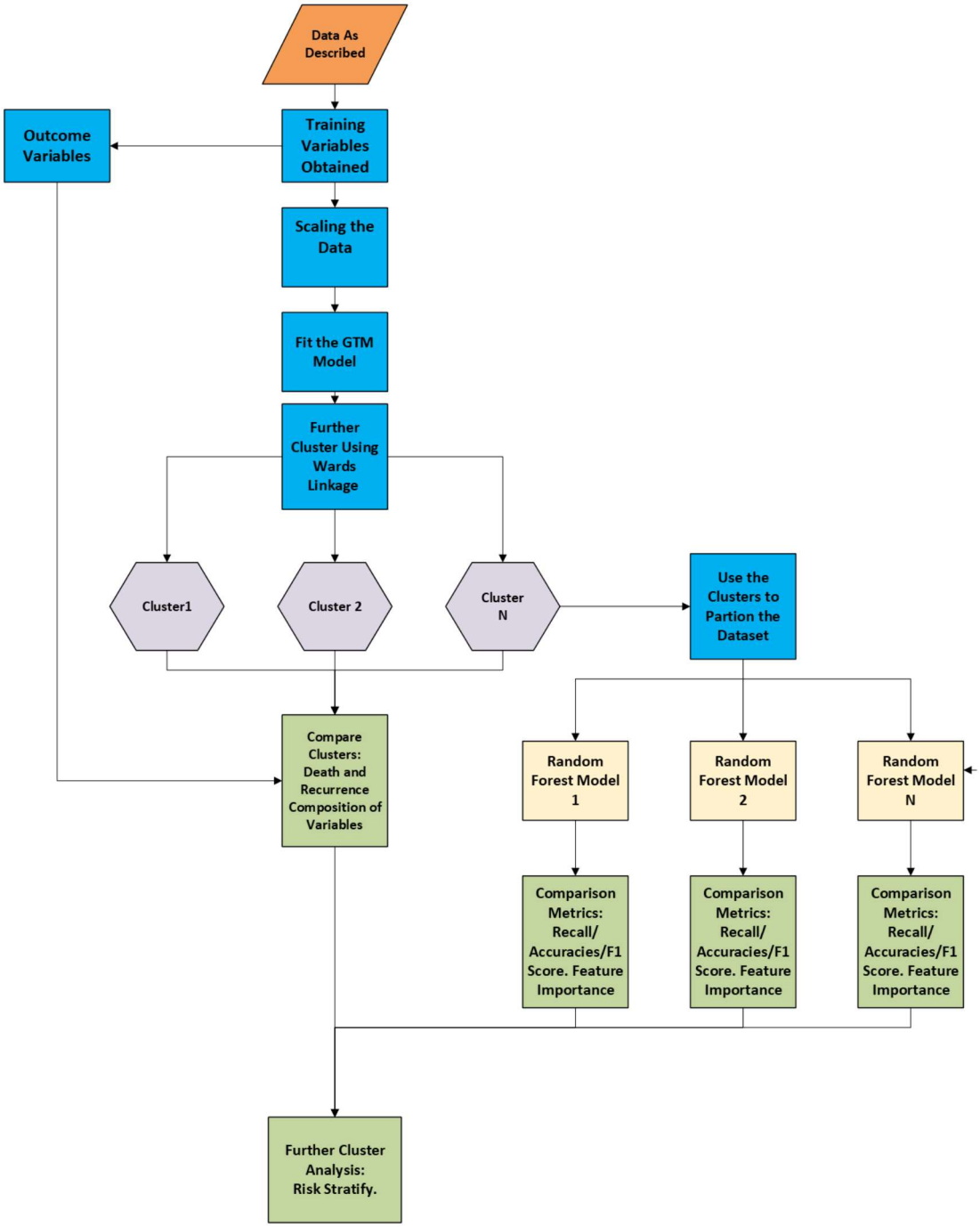

FIGURE 4.1: Methodology Flow Diagram

4.2 Generative Topographic Mapping

Generative Topographic Mapping (GTM) is a sophisticated dimensionality reduction technique. Unlike other method of reduction, such as Principal Component Analysis [57] and t-Distributed Stochastic Neighbor Embedding[54], GTM employs a probabilistic approach to map high-dimensional data onto a lower-dimensional latent space. This not

only preserves the structure of the data, but gives a probalistic framework, which can quantify any uncertainties in the map.

To find a non-linear manifold embedding of K latent variables in a low-dimensional latent space such that $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\} \in \mathbb{R}^L$. These represent a grid of nodes in the latent space. These nodes map to data points in the training set $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N\} \in \mathbb{R}^D$ in a higher-dimensional data space, where $L \ll D$. The GTM algorithm performs a parameterized non-linear mapping $f : \mathbb{R}^L \to \mathbb{R}^D$ from $\mathbf{X}$ to $\mathbf{T}$ consisting of a linear combination of radial basis functions ($\boldsymbol{\phi}$) with weighting coefficients $\mathbf{W}$. This mapping is given as follows:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\phi}(\mathbf{x}) \tag{4.1}$$

After the initial mapping, the algorithm estimates the probability density between these mapped points $\mathbf{y}_k$ and the data points $\mathbf{x}_n$ using a Gaussian noise model. This model assumes a symmetric Gaussian Distribution centred on each point (x_k) in the grid of nodes. This distribution has variance β^{-1}. The probability density $p(\mathbf{t}_n|\mathbf{x}_k)$ is defined by the following Gaussian distribution:

$$N(\mathbf{t}_n|\mathbf{x}_n, \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left(-\frac{\beta}{2}\|\mathbf{t}_n - \mathbf{x}_n\|^2\right) \tag{4.2}$$

By integrating the latent variables, we obtain the following expressed as a function of B and W:

If we choose $p(x)$ to have the form of equally weighted delta functions, we obtain the following sums:

$$p(\mathbf{t}|W, \beta) = \int p(\mathbf{t}|\mathbf{x}, W, \beta)p(\mathbf{x})\, d\mathbf{x} \tag{4.3}$$

$$p(x) = \frac{1}{K}\sum_{k}^{K} \delta(x - x_k), \tag{4.4}$$

$$p(\mathbf{t}|\mathbf{x}, W, \beta) = \frac{1}{K}\sum_{k}^{K} p(\mathbf{t}|\mathbf{x}_k, W, \beta), \tag{4.5}$$

Finally, the GTM is then fitted to the data using the log likelihood function:

$$l(W, \beta) = \sum_{n=1}^{N} \ln\left(\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{t}_n|\mathbf{x}_k, W, \beta)\right) \tag{4.6}$$

[7]

By implementing this model, we will get a membership map of the dataset. This 2D membership map will be on a n by n grid, with each node having varying members. The n will be decided by us, in which we choose the parameter for mapping the grid in order to accurately and easily interpret the plot.

4.2.0.1 Hierarchical Clustering

Given the nodes generated in the GLM discussed prior, we will further cluster using Hierarchical Clustering. Treating each of our nodes in our membership map as a cluster, we can build a hierarchy of clusters by iteratively merging closest pairs based on the euclidean distance $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. We will then reach a specified number of clusters. This will be decided based on sizes of clusters. In this thesis, Ward's linkage, defined as the minimum increase in the total within-cluster sum of squares as a result of merging two clusters. Mathematically, Ward's linkage is expressed as:

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} ||\vec{m}_A - \vec{m}_B||^2 \tag{4.7}$$

where $\Delta(A, B)$ is the distance between clusters A and B, n_A and n_B are the number of points in clusters A and B respectively, and $\vec{m}_A$ and $\vec{m}_B$ are the centres of clusters A and B. By using this linkage method opposed to other ones, we will minimise the within-cluster variance, which should give spherical clusters. It is particularly useful when nodes are similar sizes, which should be suitable for our analysis of GLM nodes [37].

4.2.0.2 Comparing Clusters

Following the creation of the clusters, we will compare the counts and distributions of variables within clusters to each. For categorical variables, we will use the chi-square test. This is a hypothgeis test used to identify if there is a significant difference between observed and expected frequencies. It has the following formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{4.8}$$

where O_i is the observed frequency and E_i is the expected frequency under the null hypothesis. The expected frequencies are calculated assuming no association between the

variables, and the test is valid when the test statistic follows a chi-squared distribution under the null hypothesis[35].

In our case, if the chi squared statistic exceeds the critical value, general set at 0.05 significance, our null hypothesis that the variables have the same distriubtion in each cluster will be rejected, suggesting relationships between the variables and the clusters.

Alternatively, the chi-sqaure is not suited for continuous variables. In those cases, we shall use the Kruskal-Wallis test. This is a statistical test used to determine significant differences between medians of three or more groups. The test statistic H is given by the following equation:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{4.9}$$

where N is the total number of observations across all groups, k is the number of groups, R_i is the sum of the ranks in the ith group, and n_i is the number of observations in the ith group.

The null hypothesis H_0 states that the distributions are identical, implying the medians and IQR ranges should be equal. H_1 on the other hand will show that atleast one of the groups medians differs from the others. If the H value is sufficienly large, the null hypothesis will be rejected, identifying a difference in a cluster.

4.2.1 Random Forest Model

A Random Forest is a ensemble decision tree classification method. By fitting multiple numerous decision trees, and then using a voting algorithm to decide which group it is in. [16].

Given a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ where $x_i \in \mathbb{R}^M$ are the input features and y_i is the target variable. Generate B bootstrap samples, which partitions the data set randomly N times with replacement, $D_1, D_2, \ldots, D_B$, each of size N.

For each sample D_b, construct a decision tree T_b by recursively splitting nodes. The split at node t is based on the feature x_j that maximizes the chosen splitting criterion but only considering a random subset of m features. From here, the outcome is decided by a equally weighted vote.

$$\hat{y} = \text{mode}\left(\{T_b(x) : b = 1, 2, \ldots, B\}\right)$$

In our thesis, we will build numerous random forest. Each cluster will be treated as a separate dataset, and we will make binary random forests predicting a negative outcome vs a clear outcome. We chose to use binary models in the case some individuals have experienced both outcomes.

4.2.1.1 Cross Validation

Rather than using a specified train and test set, we will use K-Fold Cross Validation. This works by partitioning the dataset into k equally sized subsets. The model is trained on $k-1$ of the subsets, with the remaining subset being used as the test set. The model is then trained k times, and we take an average of each model for the performance metrics. This helps ensure every data point is used for training and validating the model, which gives a better estimate of the model's performance on new data, and should help mitigate over-fitting.

4.2.1.2 Grid Search

A Random Forest is built upon various parameters. By using a grid search, we are able to iteratively attempt each combination of parameters. In a random forest,the key parameters to tune include following: the number of trees, maximum depth of trees , minimum samples per leaf , minimum samples per split and maximum features. By selecting which performance metric we wish to tune to, we are able to find the models which are specified for our goals. We will use large range of values for these parameters given, and performance cross-validation for each combination, which in turn should give the optimal set of parameters.

4.2.1.3 Performance Metrics

The common performance metrics are based off the confusion matrix given in Figure 4.2.

In the context of breast cancer risk stratification, where the goal is to identify patients at high risk of death or recurrence, prioritizing recall ($\frac{TP}{TP+FN}$) for these classes is essential. High recall ensures that the majority of high-risk cases are identified, enabling timely interventions. However, this focus on recall may lead to a higher number of false positives (FP), which may lead to wasted resources. Thus, while recall should be maximized, it is

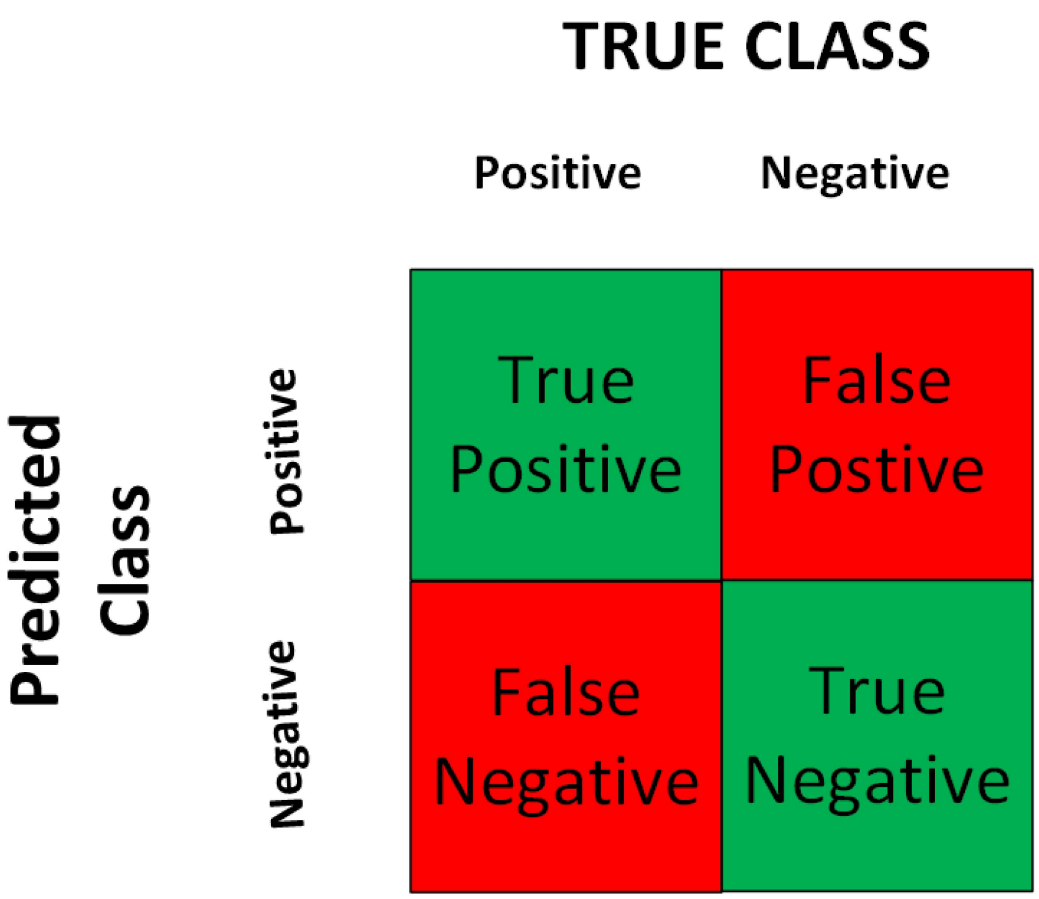

FIGURE 4.2: Confusion Matrix

also important to monitor precision ($\frac{TP}{TP+FP}$) and the overall impact on clinical decision-making. This leads to the importance of the F1 Score, defined as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This is a score which offers a balance between precision and recall. To achieve high recall for these critical classes, the Random Forest model shall be fine-tuned by adjusting the class weights or decision thresholds, emphasizing the importance of correctly predicting the minority high-risk classes. This approach ensures that the model is more sensitive to these cases, even if it means a trade-off in terms of other metrics like overall accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$).

4.2.2 Risk Stratification

The risk stratification model will employ multiple iterations to ensure a comprehensive method for identifying patient risk levels. Patients will be categorized into three broad

risk groups: low risk, medium risk, and high risk. Initially, risk stratification will be derived from the clusters by analyzing the proportions of deaths and recurrences within each cluster to give an initial risk score. Let C_D refer to proportion death within a cluster, and C_R be proportion of recurrence with $W > 1$ a weight given to C_D to show it as a more negative risk, set at 2. Hence the risk of a cluster is given as follows:

$$R_C = WC_D + C_R \tag{4.10}$$

Subsequently, the Random Forest model will be applied to predict individual patient outcomes. This will give us an individual percentage risk for the outcomes, P_R and P_D for recurrence risk and death risk respectively. This gives us the calculation for Total Risk Score R_T:

$$R_T = \frac{R_C \times (WP_D + P_R)}{1 + W} \tag{4.11}$$

Chapter 5

Results

5.1 GTM Results

5.1.1 Fitting The GTM

The visualization in Fig 5.2, which represents the cluster membership map in the GTM latent space of the developed model, shows a mapping of the breast cancer dataset. Each circle represents a cluster, with the number of clusters set to 100, arranged in a 10x10 grid, and the basis functions set to 14. The GTM regularization term was optimized through hyper parameter tuning, with the value of 1 being selected for the lowest error. This is the first model prior to further clustering. We notice a relatively even distribution of amongst the nodes, with participants ranging from between 5 and 15.

5.1.2 Distribution of Variables

5.1.2.1 Chemotherapy Drugs

We observe a high concentration of Paclitaxel, the most commonly used chemotherapy treatment, across all areas of the GTM. In contrast, the distribution of Docetaxel, Carboplatin, and Cyclophosphamide is dichotomous, with these treatments occupying distinct, opposing regions of the GTM.

Figure 5.4 shows the distribution at each cluster point. The pie chart represents the value given to each drug at each point on the latent space. Note that comparatively to the previous plot, it is only comparing the drugs, not all variables that make the point. Furthermore, at points in which they do not make up a significant amount of the

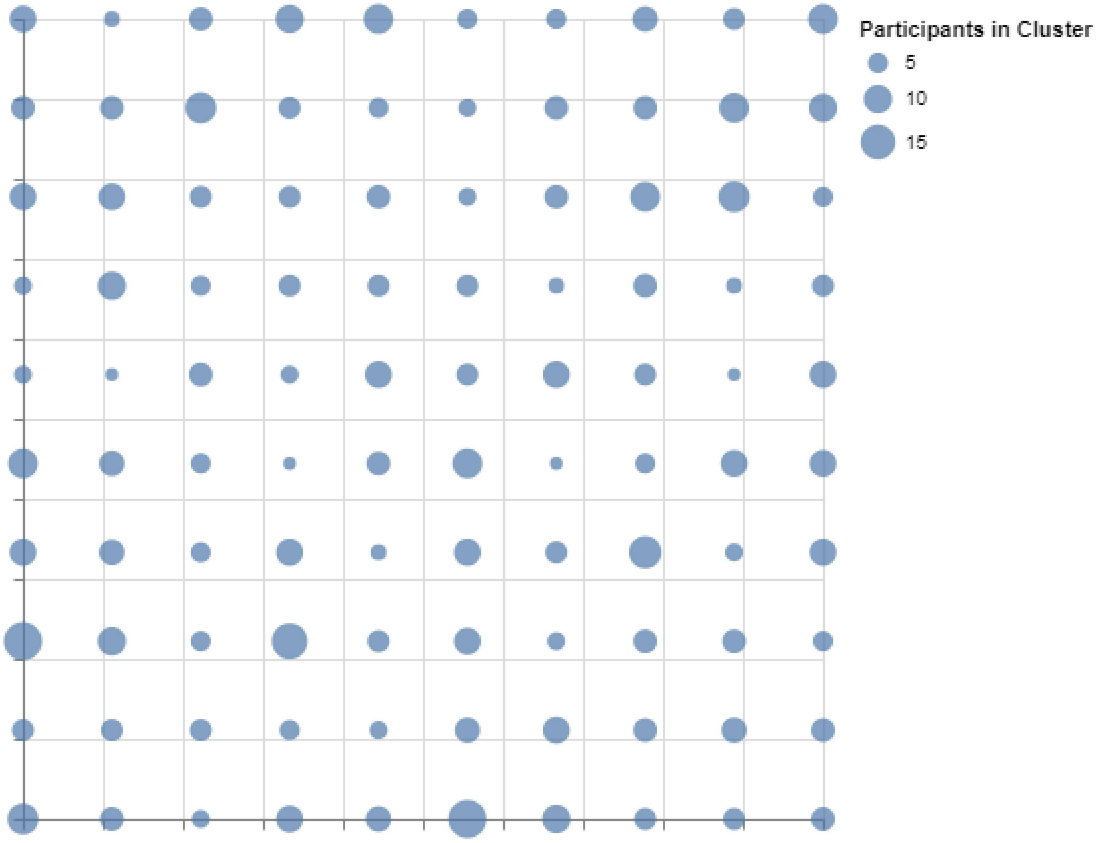

FIGURE 5.1: Membership Map

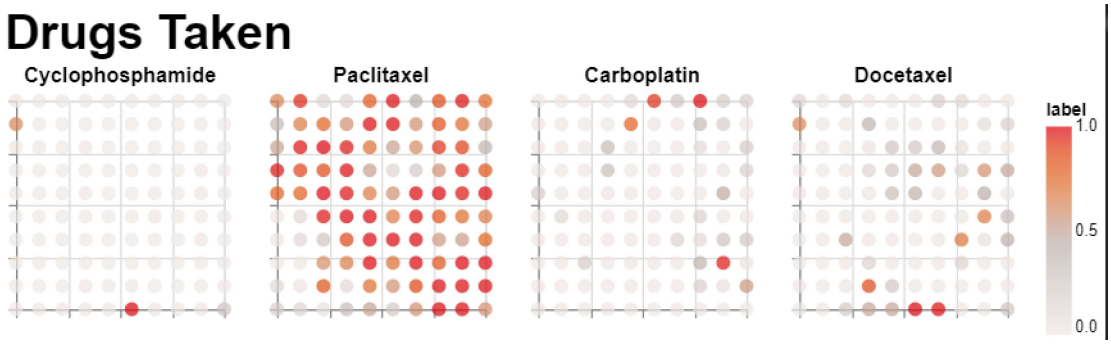

FIGURE 5.2: First GTM

value, I have not included the pie chart. We notice a vast majority with extremely high Paclitaxel, but clustered areas where docetaxel is high in proportion.

5.1.2.2 Demographic Details

In the demographic section, the results are less clear due to the significant imbalance among the Ethnic Origin classes, as previously discussed. The most notable findings are observed between the 'White' and 'Unknown' categories. There is a distinct partition based on age, where age_1 predominantly occupies the right-hand corner, while age_2 is

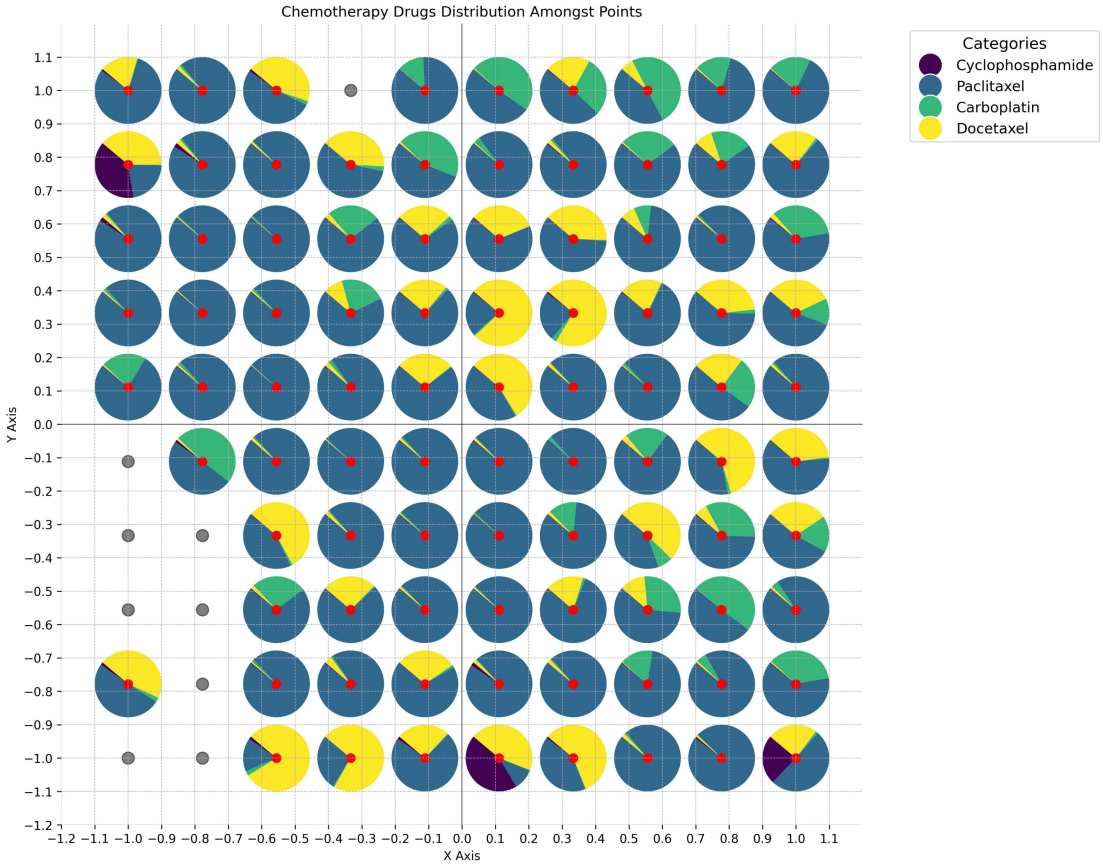

FIGURE 5.3: Distribution of Values of Drugs

concentrated below the diagonal, closer to the center. Age_3 is a much smaller subgroup, taking up only a few clear nodes. We see the bmi values have less of a distinct grouping, with lighter distribution amongst all the nodes.

5.1.2.3 Co-Morbidities

Figure 5.5 highlights the distribution of the co-morbidity variables along the gtm. We see there is few diabetes entries, hence one point is mapped for the diabetes. Mental health and hypertension have small clusters of groups, but seem evenly distributed along the entirety of the GTM.

5.1.2.4 Tumour Details

We see some interesting observations with the tumour details (Figure 5.6). We see small TNM values are mapped onto the left hand diagonal side of the map, with larger values on the right hand diagonal. Particularly, the X values are seen on the top right hand

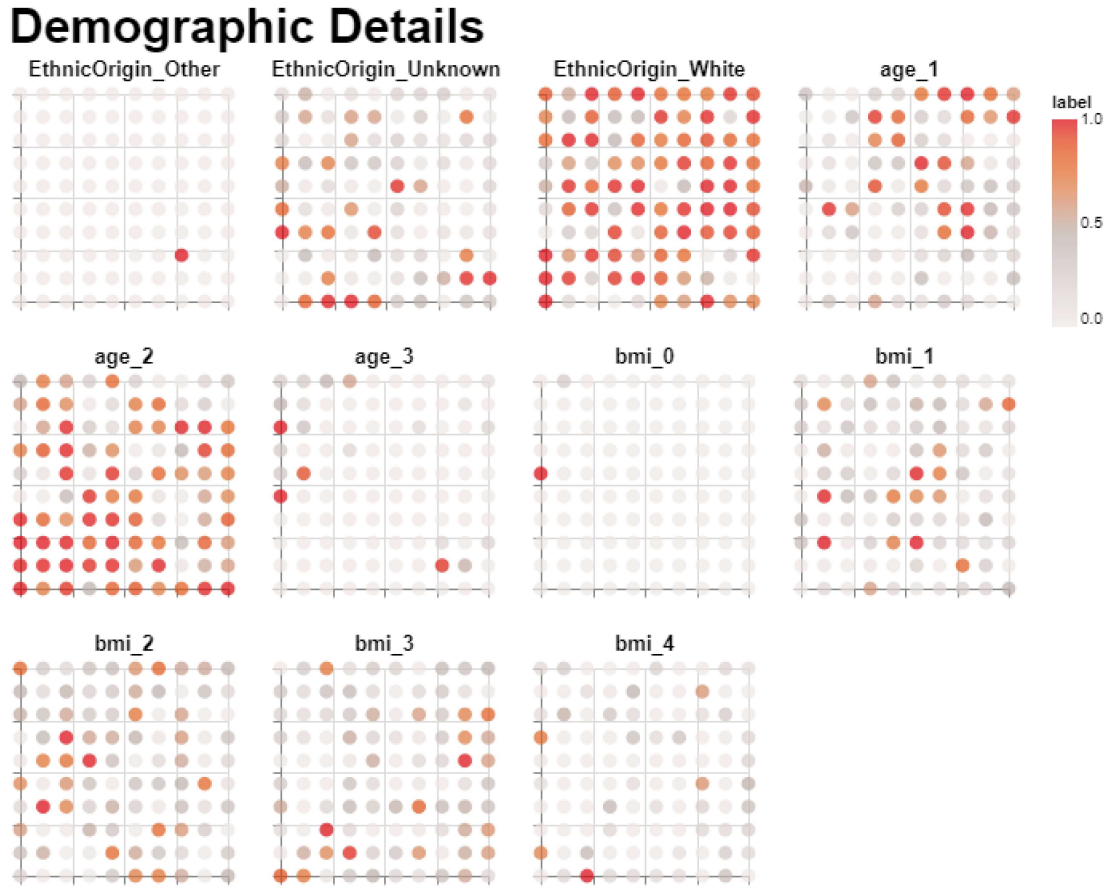

FIGURE 5.4: Distribution of Demographic Variables

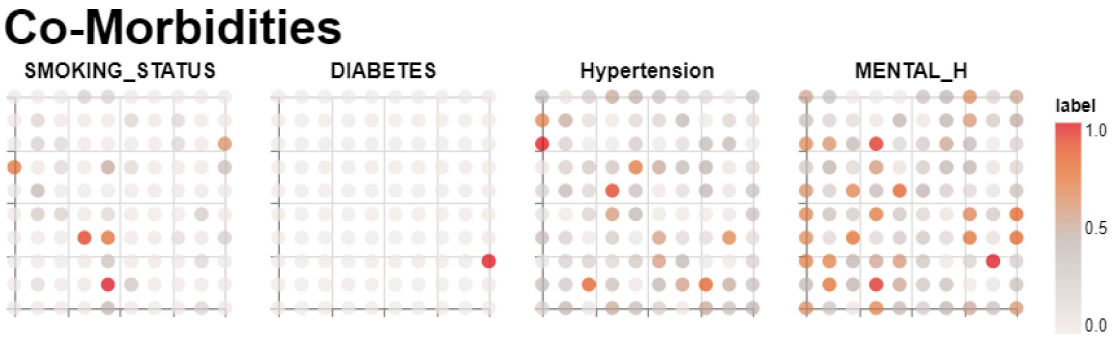

FIGURE 5.5: Distribution of Co-Morbidity

side, suggesting a clustering. Histology grade shows little difference across the spread. Furthermore, PSAtFirstVisit seems to map high values to the lower side of the map.

5.1.2.5 Blood Tests

Blood tests (Figure 3.3) are much harder to interpret. Some tests have some tests which are specifically high in certain areas, such as the Monocytes and Lymphocyte count, which show deep concentrations on individual nodes. Alternatively, Globulin and WBC

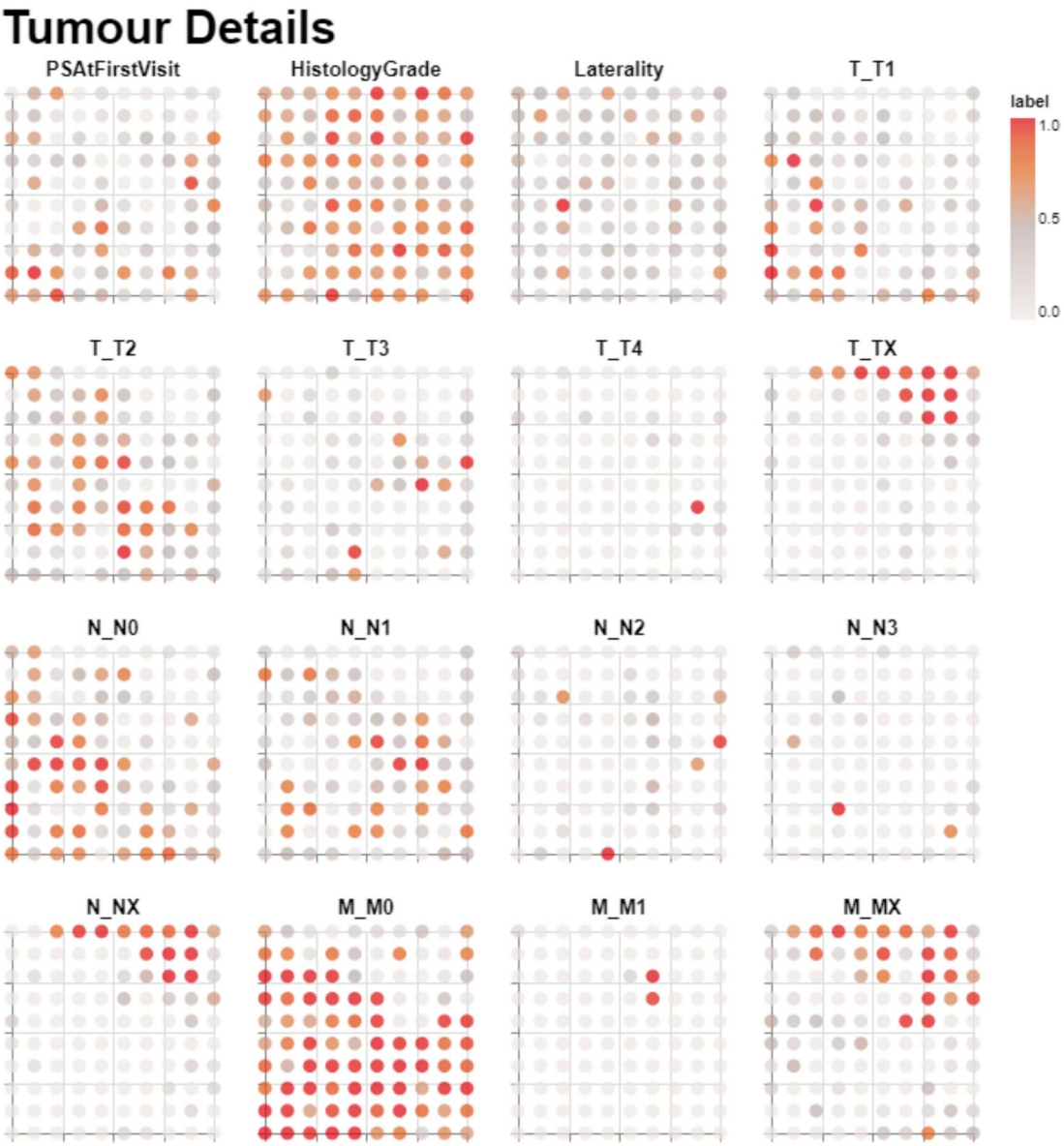

FIGURE 5.6: Tumour Details Mapping

show high concentrations across the entire map, with little clusters where it is further concentrated.

5.1.2.6 Negative Side Effects

With the negative side effects (Figure 5.8), the interesting observations in this differences between those who have high averages, versus large standard deviations. We see generally mirrored values for these, showcasing the differences between patients who have large variance in there side effects versus constant high averages.

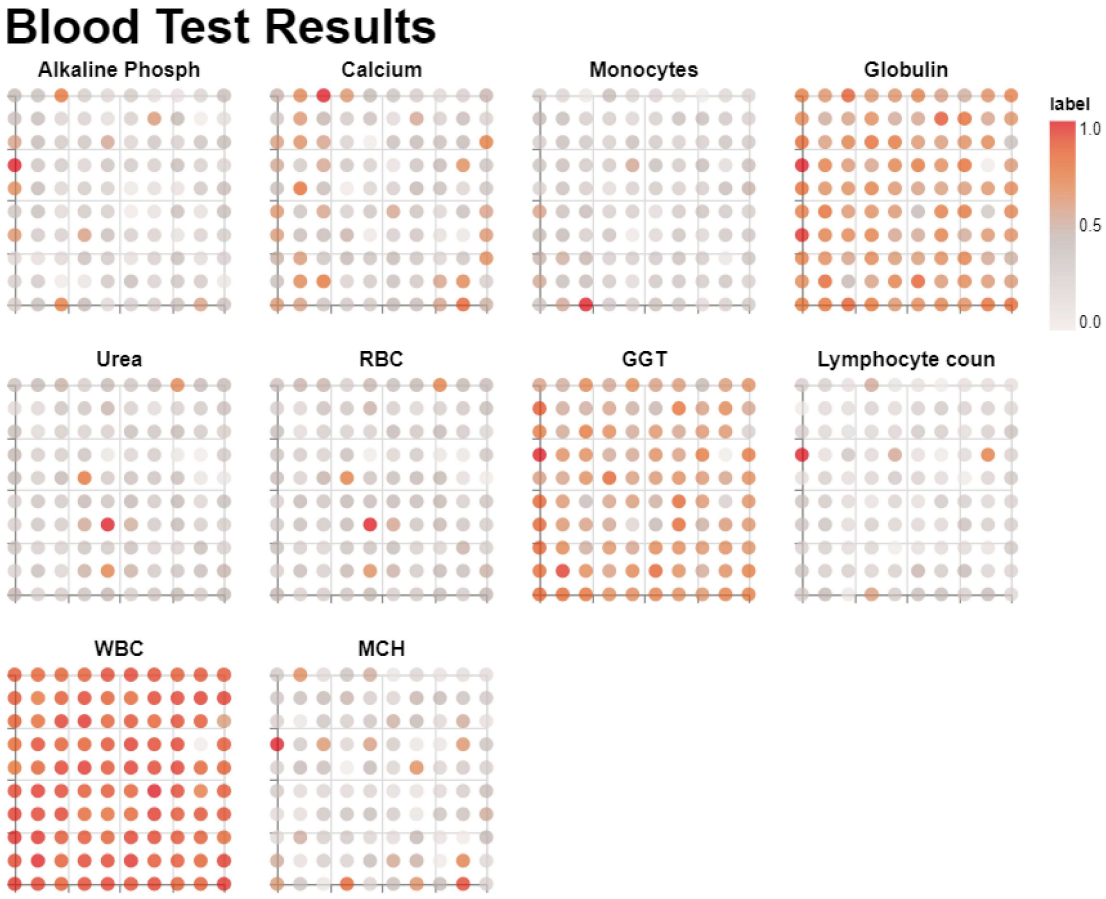

FIGURE 5.7: Blood Test Mapping

5.1.2.7 Outcomes

Figure 5.9 demonstrates the mapping of the Recurrence and Patient Deceased on the developed latent space. Despite not training on these variables, we notice high concentrations in certain areas. We note that there exists areas of extremely high concentration in Patient Deceased on the right hand middle.

5.1.3 Clustering

To perform the clustering of the dataset, we use the hierarchical clustering discussed in chapter 4. By using a limit point of 4 in terms of euclidean distance, we achieve the following Clusters. Figure 5.10 showcases the clusters on the membership map. There exists 6 different clusters within the space, each of varying sizes.

Table 5.1 presents the distribution of deceased and recurrence cases across the clusters. It also details the size of each cluster, including the number of patients and the proportions of each outcome. Several observations stand out.First, we see considerable differences

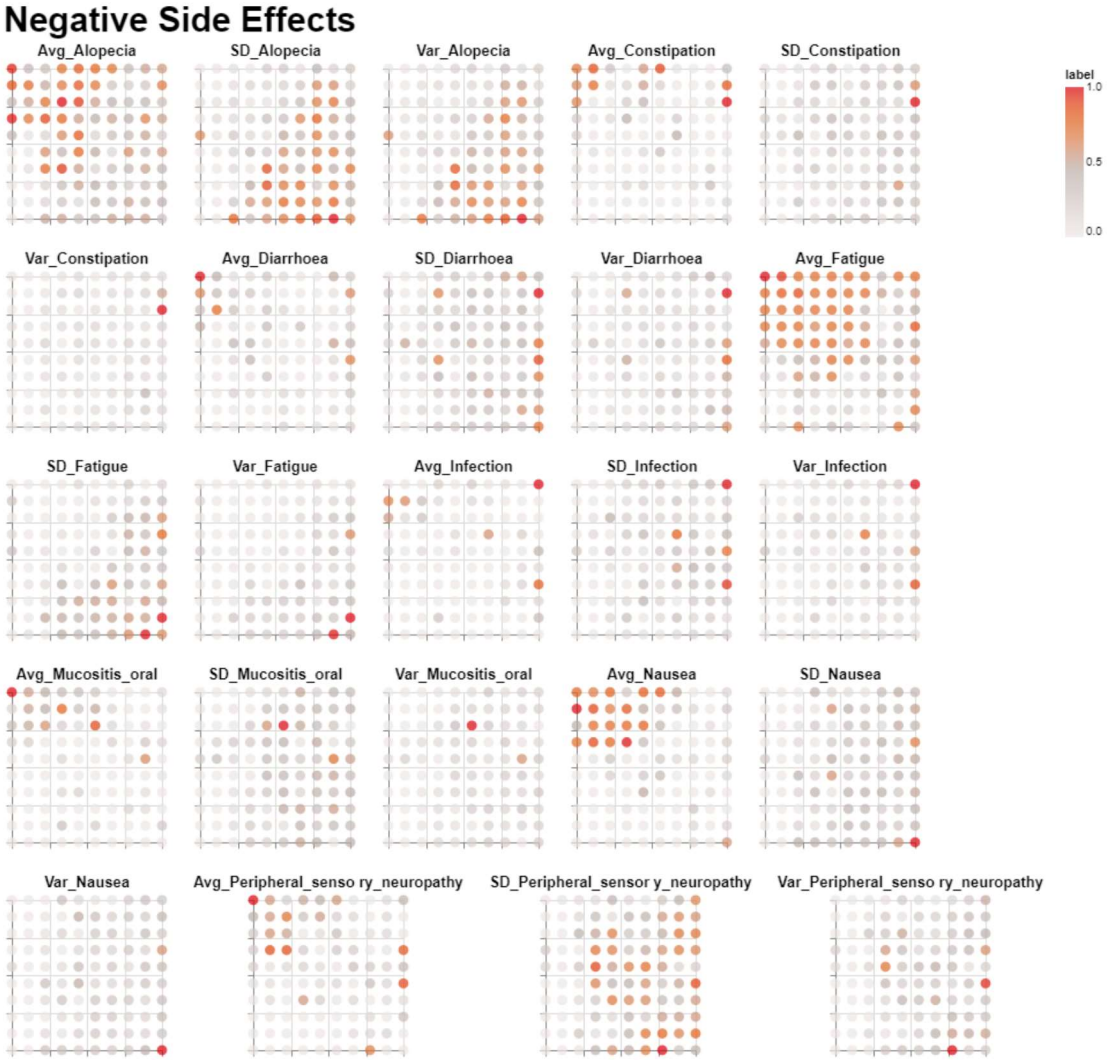

FIGURE 5.8: Negative Effects Mapping

Cluster	Participants in Cluster	Patient Deceased Count	Recurrence Count	Patient Deceased Proportion	Recurrence Proportion
1	146	20	31	0.136986	0.212329
2	115	10	25	0.086957	0.217391
3	139	10	10	0.071942	0.071942
4	95	3	3	0.031579	0.031579
5	106	11	27	0.103774	0.254717
6	124	11	9	0.088710	0.072581

TABLE 5.1: Patient data distribution across clusters.

in the sizes of clusters. With the smallest being 95, and the largest being 146 we note a big difference in the distributions. Furthermore, Cluster 4, which is the smallest cluster, exhibits extremely low proportions of both deceased and recurrence cases, with values of both 0.0316. This is an incredibly low risk outcome rate, significantly lower than general recurrence rates. These issues may introduces challenges when training further models, with not enough samples for a suitable outcome class. Due to this, to enhance risk stratification, any patient identified by the GTM as belonging to cluster 4 shall automatically classified as a low risk patient.

FIGURE 5.9: Outcomes Distribution

In contrast, Clusters 1,2 and 5 all showcased extremely high recurrence rates. With proportions of over 20%, it highlights at risk patients within these groups, and highlights these as at risk groups. Furthermore, we see cluster 1 with a significantly larger proportion of deaths compared to the other groups. This further highlights it as the most at risk class.

To identify whether these are statistically significant results, we can use the Chi-Square statistic discussed earlier. Table 5.2 showcases the results. We see that the deceased is not considered statistically significant using this test, however we see the differences in recurrence are considered significant using the standard 0.05% significance boundary. Whilst the deceased may not be considered significant, we can still use it for the modelling and risk stratification.

Outcome	**Chi-Square Statistic**	**p-value**	**Significance**
Deceased	8.74	0.120	Not Significant
Recurrence	41.62	7.03×10^{-8}	Highly Significant

TABLE 5.2: Chi-Square Results

Based off a sum of negative proportions, we establish the following ranking for a baseline risk (Fig 5.11).This will be used in further modelling of new data.

5.1.4 Distribution of Variables in Clusters

5.1.4.1 Categorical Variables

Using a chi-square test with the null hypothesis (H_0) that there is no significant association between variables and cluster assignment, and the alternative hypothesis (H_1)

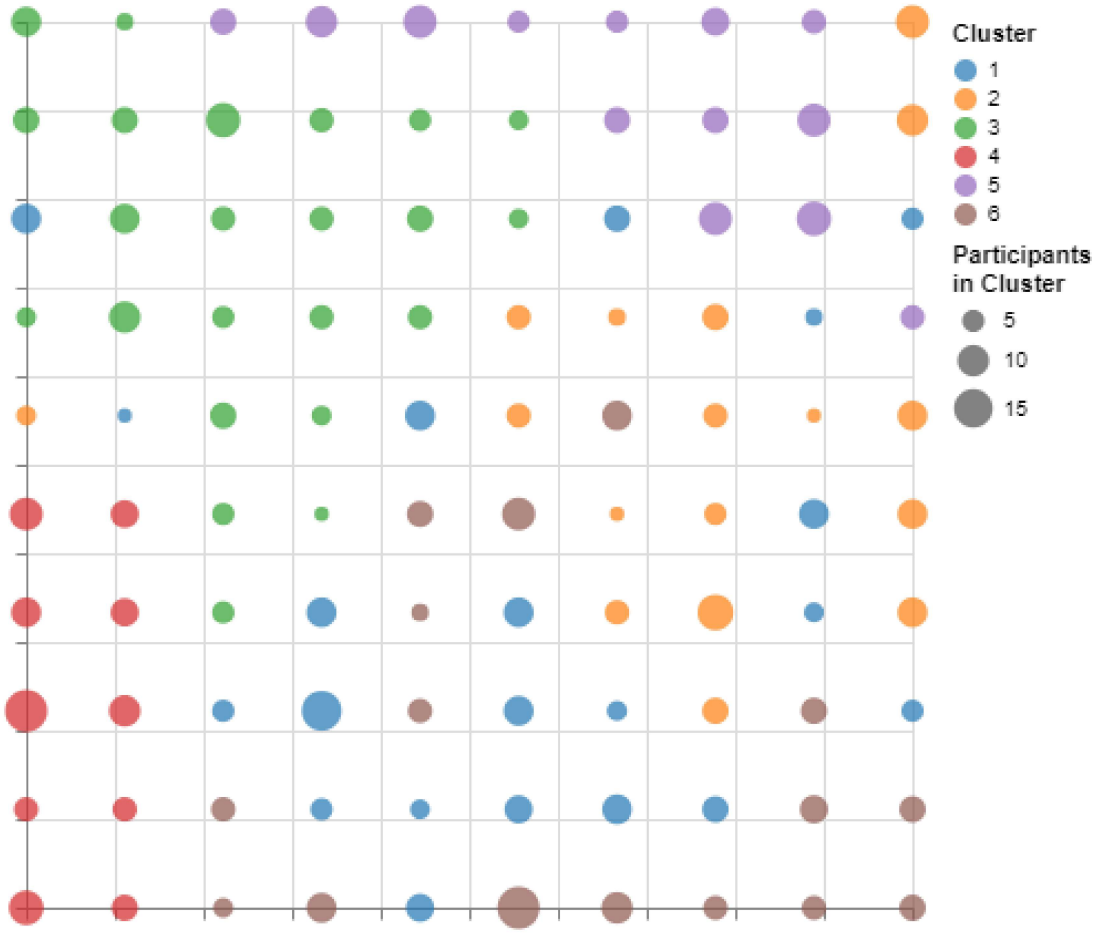

FIGURE 5.10: GTM Clusters

that there is a significant association, we obtained the results presented in Table 5.3. The table is order by the most significant results based on the p values, which in turn will showcase the examples of the biggest differences between classes. We must note that they represent the percentages within groups. For example, in Cluster 5, 94.34% of patients belong the N_{NX} class. However, in smaller sized groups, we will see small percentages across each cluster, such as "Ethnicity Other" is 6.09% of Cluster 2, but that is the only cluster it is mapped to. Hence Cluster 2 is a cluster with exceedingly large "Ethnicity Other".

We see interesting results within the top half of the important variables. We see the T,N,M variables are the most differentiation categorical variables, occupying 6 of the top 10 spots. Cluster 5 particularly contains large percentage make up of the N_NX, M_MX and T_TX, showing a mapping of all the unidentifiable variables to this cluster. Other clusters show big make ups of other N variables, with Cluster 4 having large N and M 0 cases.

When it comes to other important variables, we see the major drug groups occupying a

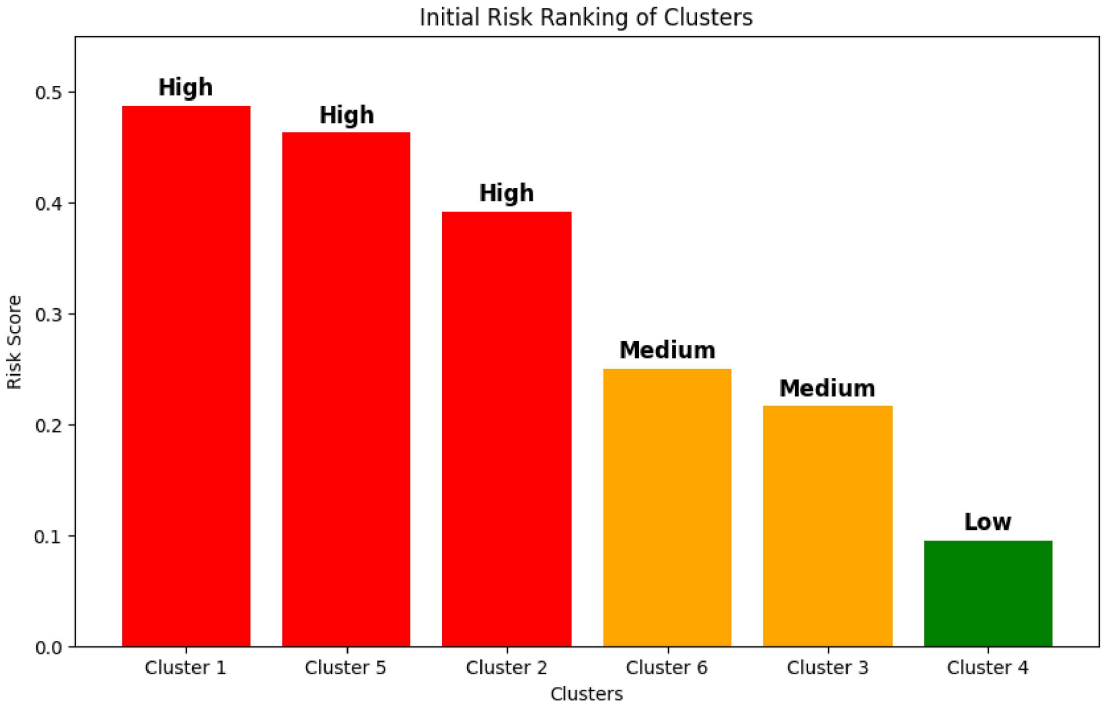

FIGURE 5.11: Risk Ranking

large proportion of the cluster percentages. Most clusters have large "Paclitaxel" usage, which is as expected due to it being the largest class significantly. We further see that Cluster 2 and 6 have large percentage users of "Docetaxel", and Cluster 6 is significantly large for "Cyclophamide". We do see low usage of all drug groups in Cluster 4. This suggests alternative treatment methods which are not recorded, such as radiotherapy or other chemotherapy drugs which we did not keep.

Furthermore, age group shows ability in being able to differentiate between groups. We see age group 2 is large in clusters 1,4 and 6, with age group 1 being large in Clusters 2 and 5. We see less interesting results with age group 3. There exists other interesting categories, such as smokers being in cluster 1 and 3, diabetics in cluster 1 and those who struggle from hyptension generally in cluster 1 as well. Hence we see cluster 1 has large proportions of co-morbidities.

One interesting aspect is that some categories do not show any statistical significance in the distribution amongst clusters. Particularly, 1 which is underweight and 4 which is the largest. We might expect both of these to have impact on progression on outcomes.

Variable	Cluster and Percentage						P Value
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	
N_NX	6.85	13.91	0.72	1.05	94.34	3.23	3.09e-104
T_TX	6.16	13.04	3.60	0.00	92.45	4.03	1.39e-98
M_MX	11.64	24.35	28.78	15.79	83.96	29.03	1.68e-36
Paclitaxel	74.66	69.57	78.42	4.21	70.75	69.35	2.69e-35
M_M0	83.56	73.04	69.78	84.21	16.04	70.97	1.36e-32
N_N0	27.40	24.35	52.52	67.37	0.94	64.52	1.46e-32
age_1	10.96	64.35	33.81	12.63	38.68	16.13	1.39e-24
T_T1	14.38	17.39	39.57	53.68	0.94	45.97	9.28e-24
age_2	72.60	33.04	57.55	75.79	52.83	82.26	2.48e-16
Cyclophosphamide	0.00	0.00	2.88	0.00	0.00	16.13	1.04e-15
N_N1	33.56	54.78	33.09	30.53	2.83	20.97	2.75e-15
T_T2	51.37	43.48	44.60	36.84	2.83	33.87	3.57e-14
Docetaxel	11.64	27.83	12.23	1.05	7.55	31.45	1.37e-11
Ethnicity Unknown	14.38	14.78	32.37	36.84	19.81	47.58	7.42e-11
Ethnicity White	85.62	79.13	67.63	63.16	80.19	52.42	3.35e-09
N_N2	19.86	5.22	10.07	1.05	0.00	6.45	1.00e-08
Ethnicity Other	0.00	6.09	0.00	0.00	0.00	0.00	4.77e-07
T_T4	10.27	2.61	1.44	0.00	0.94	0.81	2.47e-06
Carboplatin	2.74	9.57	5.04	1.05	18.87	9.68	4.47e-06
N_N3	12.33	1.74	3.60	0.00	1.89	4.84	2.32e-05
MENTAL_H	33.56	48.70	30.22	53.68	24.53	40.32	2.54e-05
age_3	16.44	2.61	8.63	11.58	8.49	1.61	9.26e-05
Smoker	16.44	3.48	10.79	3.16	8.49	3.23	1.23e-04
T_T3	17.81	23.48	10.79	9.47	2.83	15.32	1.56e-04
DIABETES	3.42	0.00	0.00	0.00	0.00	0.00	1.27e-03
Hypertension	38.36	19.13	27.34	17.89	29.25	24.19	2.92e-03
bmi_0	0.00	3.48	0.72	0.00	0.00	0.00	5.96e-03
M_M1	4.79	2.61	1.44	0.00	0.00	0.00	1.04e-02
bmi_2	33.56	26.96	46.04	35.79	34.91	29.84	3.02e-02
bmi_3	36.30	27.83	19.42	28.42	31.13	27.42	6.34e-02
bmi_4	6.16	15.65	13.67	6.32	10.38	12.90	8.28e-02
bmi_1	23.97	26.09	20.14	29.47	23.58	29.84	4.76e-01

TABLE 5.3: Categorical Percentages and P-values

5.1.4.2 Integer Variables

Table 5.4 showcases the medians of our continuous integer variables, and again organises by P value, given by the Kruskal-Wallis given by equation 4.9. We see various observations. First the side effects are shown to be the most significant during treatment, with alopecia showing to affect Cluster 1 , 2 and 6. Particularly, we see clusters 1 and 2 showing high signs of all treatment side effects generally across the board. These are likely the patients who struggle the most during treatment. We see big differences i the blood tests also. Cluster 5 has the lowest of all WBC counts at the median , and the largest is cluster 4. Although we do see quite a lot of variables look the same along various clusters, as we are only interpreting the medians.

Figure 5.12 showcases a general demographic of the patients in each cluster. This is a simplified version, but showcases how individuals may be distributed amongst clusters.

Variable	Cluster and Median						P Value
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	
Var_Alopecia	0.333	0.500	0.000	0.000	0.000	0.333	2.14e-52
SD_Alopecia	0.577	0.707	0.000	0.000	0.000	0.577	2.14e-52
Avg_Nausea	0.000	0.000	1.000	0.000	0.000	0.000	2.66e-43
Var_Fatigue	0.276	0.250	0.000	0.000	0.000	0.000	4.69e-43
SD_Fatigue	0.525	0.500	0.000	0.000	0.000	0.000	4.69e-43
Avg_Alopecia	1.000	1.000	0.000	0.000	1.000	0.000	1.22e-40
Avg_Fatigue	0.000	1.000	1.000	1.000	1.000	0.000	6.18e-40
SD_Nausea	0.500	0.518	0.000	0.000	0.000	0.468	2.88e-33
Var_Nausea	0.250	0.268	0.000	0.000	0.000	0.219	2.88e-33
Var_Peripheral_sensory_neuropathy	0.200	0.286	0.000	0.000	0.000	0.244	5.00e-26
SD_Peripheral_sensory_neuropathy	0.447	0.535	0.000	0.000	0.000	0.494	5.00e-26
SD_Mucositis_oral	0.500	0.548	0.000	0.000	0.000	0.444	2.70e-25
Var_Mucositis_oral	0.250	0.300	0.000	0.000	0.000	0.197	2.70e-25
SD_Infection	0.000	0.548	0.000	0.000	0.000	0.000	7.68e-25
Var_Infection	0.000	0.300	0.000	0.000	0.000	0.000	7.70e-25
SD_Diarrhoea	0.189	0.578	0.000	0.000	0.447	0.000	8.47e-25
Var_Diarrhoea	0.071	0.333	0.000	0.000	0.200	0.000	8.47e-25
Avg_Mucositis_oral	0.000	0.000	0.000	0.000	0.000	0.000	6.12e-23
Avg_Peripheral_sensory_neuropathy	0.000	0.000	0.000	0.000	1.000	0.000	1.33e-19
SD_Constipation	0.189	0.500	0.000	0.000	0.408	0.000	2.41e-19
Var_Constipation	0.071	0.250	0.000	0.000	0.167	0.000	2.41e-19
Monocytes	3.918	3.960	3.977	4.270	3.505	3.871	4.33e-15
Avg_Diarrhoea	0.000	0.000	0.000	0.000	0.000	0.000	9.67e-14
HistologyGrade	2.000	3.000	3.000	3.000	3.000	3.000	7.37e-13
WBC	42.161	43.545	42.400	44.200	41.535	42.958	4.49e-11
Avg_Infection	0.000	0.000	0.000	0.000	0.000	0.000	6.80e-11
GGT	2.346	2.345	2.338	2.359	2.361	2.364	8.17e-10
PSAtFirstVisit	0.000	0.000	0.000	0.000	0.000	0.000	9.42e-08
Avg_Constipation	0.000	0.000	0.000	0.000	0.000	0.000	9.73e-07
Globulin	68.090	69.000	69.000	70.500	68.463	67.838	5.04e-06
Calcium	4.726	4.583	4.700	5.300	4.004	4.933	7.01e-05
Lymphocyte_coun	0.468	0.450	0.467	0.550	0.410	0.464	3.14e-04
Laterality	0.000	0.000	1.000	1.000	0.000	1.000	1.73e-02
Alkaline_Phosph	26.000	25.111	26.375	25.670	25.500	24.621	4.03e-02
MCH	0.033	0.020	0.040	0.029	0.029	0.029	1.09e-02
Urea	91.494	90.367	90.814	90.054	95.563	91.214	2.39e-01
RBC	30.750	30.750	30.650	30.888	29.720	30.764	4.33e-01

TABLE 5.4: Median and P-values for clusters.

5.2 Random Forest Model

Initially, we trained each cluster and outcome on a random forest model. This was to obtain feature importance scores. Using this, for each model we then select the top 30 features with the highest importance scores. This reduced the dimensional of the This step helps reduce the dimensionality of the dataset, removing the potential bias introduced by the low importance features brought. Furthermore, this is important in visually determining which features are important, and allows for comparison between models. This will be discussed later in the chapter.

Following this, a grid search is computed for each of the clusters and outcomes. Given the selection of parameters detailed in Chapter 4, the grid search was used to find the optimal ones for each cluster. The grid search is chosen to prioritize recall of the negative outcome class, rather than the F1 Score or Accuracy. However, we noticed no differences

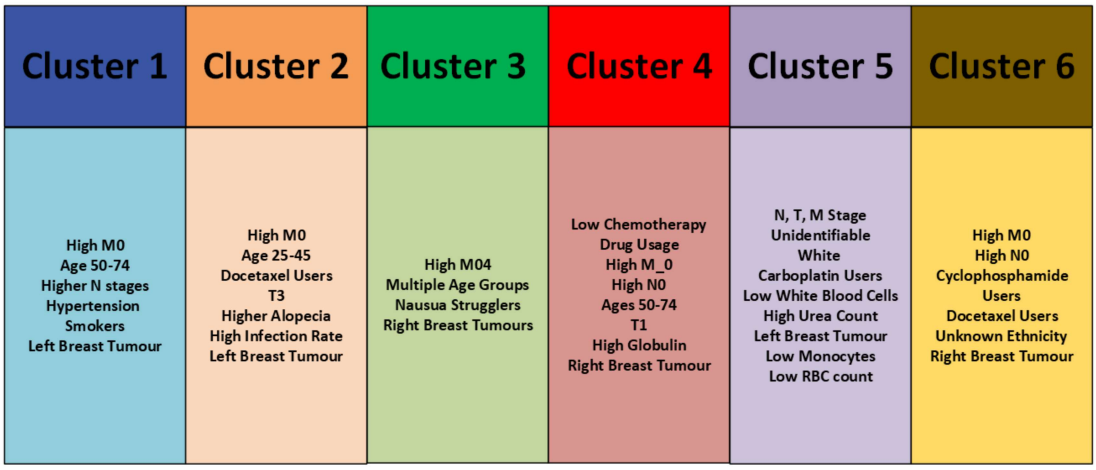

FIGURE 5.12: Cluster Phenotypes

in the parameters between each cluster, only between the separate outcomes. This is likely due to being extremely similar.

Outcome	n_estimators	max_depth	min_samples_split	min_samples_leaf
PatientDeceased	25	None	2	1
Recurrence	25	None	2	1

TABLE 5.5: Grid Search Selections

From here, we finally computed the models with a cross validation of 5, giving the results presented in the following sections. Note that for Cluster 4, I decided to not do predictive modelling as the class imbalances where to big to make a suitable test set. Hence we will set the risk score how it is calculated for cluster risk. The **Mean** is a weighted mean using the size of the clusters to give weights.

5.2.1 Recurrence

TABLE 5.6: Recurrence Classification Reports

Cluster	Clear			Recurrence			Accuracy
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
1	0.904	0.841	0.870	0.559	0.699	0.609	0.827
2	0.881	0.806	0.867	0.575	0.660	0.539	0.756
3	0.9956	0.9084	0.9621	0.1111	0.1667	0.1333	0.9218
4	-	-	-	-	-	-	-
5	0.8526	0.7275	0.7580	0.3667	0.5197	0.4875	0.6852
6	0.9328	0.9667	0.9485	0.3333	0.0667	0.1111	0.9102
Mean	**0.920**	**0.856**	**0.887**	**0.386**	**0.422**	**0.376**	**0.828**

We see the clear case has high predictive F1 Score along all clusters, although it is smaller in Cluster 5. For the high risk clusters as described (1,2,5) we see a good recall rate across these clusters. Particularly, clusters 1 and 2 have recalls of 0.699 and 0.660 respectively. The proportion of recurrence in these classes is higher than the proportion of patients deceased, hence we expect to see better predictive accuracy. We might note that on these cases we see quite low precision scores. This is to be expected with how we chose to approach the modelling.

Comparatively, we see the model really struggles with the "Medium Risk" clusters. This is due to the low proportions making the model struggle with predictive accuracy. We see extremely low scores, with 0.1667 and 0.0667 for recall respectively. However, we should note the low percentage of recurrence within these groups. As shown in Table 5.1. They both have 0.07% recurrence, which is an extremely small amount. Hence we would expect to see a poor predictive accuracy.

For using the model as an ensemble method for total predictive accuracy's, we see a total accuracy of 0.828 percent. However, we note a total recall of the negative class of 0.422. We could conclude the usage of the model in case of the high risk patients, but further data required for the medium and low risk patients.

5.2.2 Patient Deceased

TABLE 5.7: PatientDeceased Classification Reports

Cluster	Class 0			Class 1			Accuracy
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
1	0.929	0.939	0.922	0.573	0.388	0.409	0.882
2	0.940	0.956	0.945	0.389	0.222	0.278	0.904
3	0.9618	0.9362	0.9468	0.1944	0.4167	0.2219	0.9066
4	-	-	-	-	-	-	-
5	0.6877	0.8877	0.7373	0.3888	0.5333	0.4335	0.8652
6	0.9193	0.8963	0.9053	0.3889	0.1500	0.1778	0.8241
Mean	**0.892**	**0.927**	**0.899**	**0.389**	**0.340**	**0.300**	**0.880**

We see similar results with the patient deceased groups. We see generally higher results in the high risk groups, although Cluster 3 shows significantly better performance in this model. Cluster 3 has a higher proportion of deceased than it does recurrence, hence this makes logical sense.

5.2.3 Feature Importance

Using the function of the random forest, we are able to visually interpret which variables are mostly impacting the predictive power of the random forest. The following graphs are stacked feature importance bar charts for each cluster. We should note that each cluster had very similar importance graphs in terms of the variables included in the top 30.

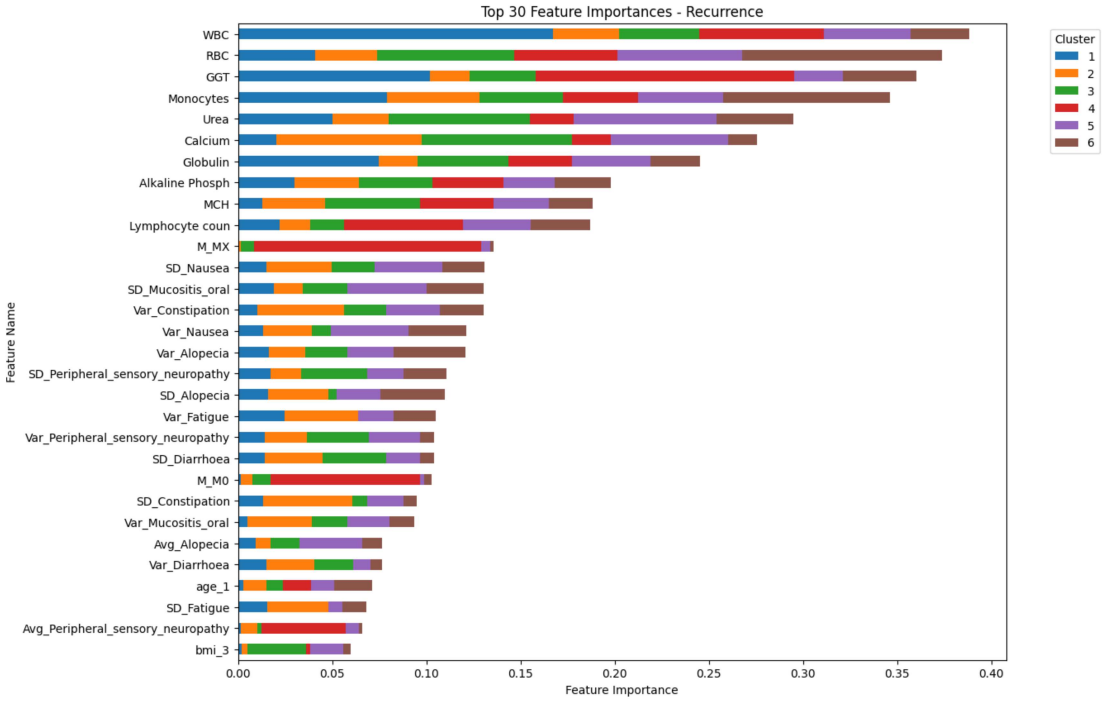

FIGURE 5.13: Feature Importance of Recurrence Models.

For Recurrence, we see interesting results. We note the most important features are given to the blood test results, such as the white blood cell count. This is consistent for all clusters, however we see particularly large importance given to these by the first Cluster. Cluster 1 has majority of importance given to only a few variables, where as for example Cluster 6 seems more evenly distributed amongst the variables. Similarly, the distribution of variables is quite similar for patient deceased outcomes. We see consistently the blood test results given high predictive powers, along with side effects during treatment, the histology grade, and age values.

5.3 Mapping New Data

Given the subset of data which did not meet the three year follow up period described in Chapter 3, we shall map this data onto the membership map to show the suitability

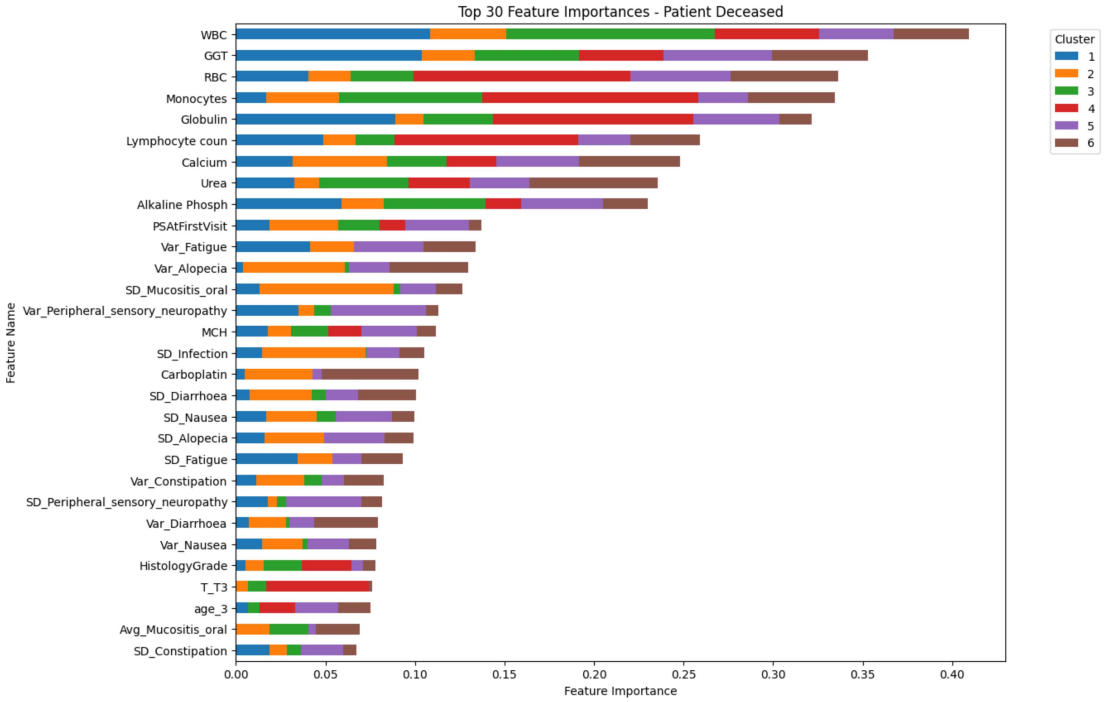

FIGURE 5.14: Feature Importance of Patient Deceased.

for further data. Following this, the risk scores for these shall be calculated to highlight the potential of the pipeline.

By ensuring the same methods of scaling were used, we obtained the following membership map:

We see a large amount being mapped towards 1,3 and 5, with points with varying amounts of members. We see lots of the map is considered empty. From this, we could establish the preliminary risk ranking as discussed earlier. However, using the random forest models established, we shall compute the percentage chance it belongs to the negative groups. Using the equation for total risk, Equation 4.94.11, we can calculate the total risk. The following bar chart shows the total risks for the patients in this test set.

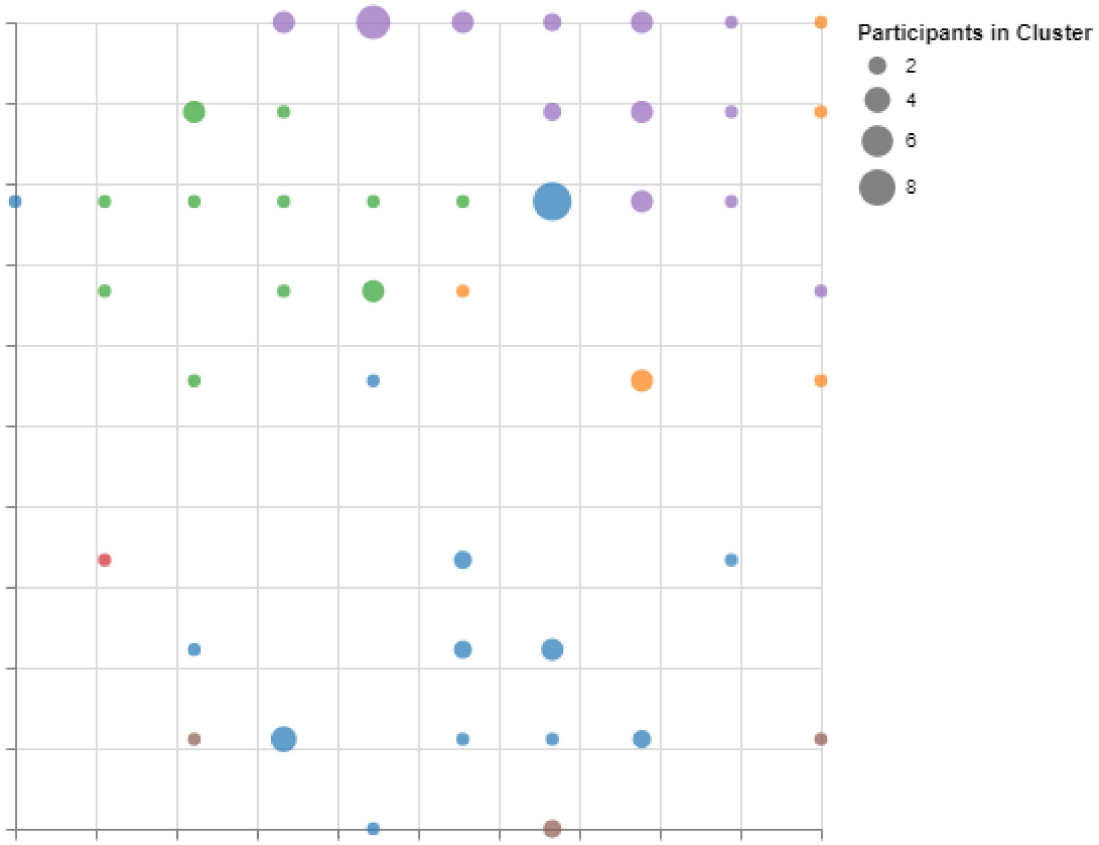

FIGURE 5.15: Membership Map for New Patients

FIGURE 5.16: Risk Map for New Patients

Chapter 6

Discussion

6.0.1 Comparing Clusters

Our analysis has demonstrated effective clustering, revealing distinct groups and showcasing the ability to map new data to these clusters. In particular, we saw key differences based on ages, the TNM staging, chemotherapy drugs taken, and side effects of treatment. This study represents the first instance of using GTM for clustering a breast cancer dataset. When compared to other clustering techniques, the results are comparable. For example, while Markey et al [34] identified a greater number of clusters, the clustering patterns were similar, grouping data based on tumour shape, size, and patient age, although without using TNM staging. This also aligns with other studies on the matter discussed in the literature review [5] Hence the findings align with established effective methods. This highlights the potential of GTM within this domain.

In the higher risk clusters, we noticed a large proportion of unidentifiable tumour markers, higher T stages, (T_3 and T_4), higher N stages. Furthermore, Cluster 1 represents a large proportion of older patients, with the highest rate of death. This compares to commonly found risk assessments.

We found lower amount of clusters compared to prior literature discussed earlier. However, this was a cluster cut off point set by us, and there was definite ability to further cluster, which we decided not to do based on the data limitations. Despite this, we saw good differences between the clusters. If more data was to be collected, it may be worth experimenting with the cut off points used in model.

6.0.2 Comparing Predictions

The predictive accuracy of models trained on these clusters varied, with smaller clusters exhibiting lower predictive power due to data limitations. This highlights the needs to focus on acquiring more data, and ensures the need for data completeness as discussed prior. Despite this, comparing to other research in recurrence, we see better results. One example is a recent study by Gonzales et al [20] showcased the predictive ability of a variety of models on recurrence. Despite only focusing on the accuracy's and other metrics, we find high accuracy's with our models, and better performance on the recall of recurrence. This suggests good performance with the variables and methodology we selected, and gives need for further work with these.

6.0.3 Important Results

In the study, we notice that high feature importance was given largely towards the blood tests undergone during treatment. Whilst clustering at this point has already been done, which potentially eliminates the effect of various variables in this model, as patients within clusters are similar. However, it does highlight a need for further research into the impact of blood and other bio markers. Notably, white blood cell count emerged as the most important feature for a range of models. White blood cell counts are recognised as potential markers for different cancers, and tend to decrease with age, and have shown in some studies to have predictive value in all-cause mortality [10, 15, 29]. Furthermore, high feature importance was assigned to the treatment side effects. This highlights the importance of monitoring patient well being, and opens avenues for further research in how patient health impacts. Similarly, we saw significance statistical differences in clusters based on the treatment drug of choice. This is under researched within the data science and bio-informatics community, despite being known the impacts of various drugs on outcomes, as discussed earlier.

6.0.4 Implications of Work

One important of this study was the potential implications in medicine and patient care. Using our risk stratification model, we were able to successfully assign risk values to individuals. Our models have showcased good recall, particularly for those of high risk. This showcases significant potential in enhancing decision making by clinicians, enabling high risk patients to potential have more treatment, and allowing low risks to avoid any unnecessary ones. In particular, we have defined a subset of follow up data, who did not meet our inclusion criteria. Our results suggest that those highlighted by

the follow up may be at risk, hence consultation with clinicians, and following up these patients would potentially highlight issues or positives with our model. Furthermore, we hope for this to contribute to developing a more thorough way of personalised treatments for patients, to improves outcomes. We note that the model will need work prior to any implication in practice, including further data, changing weights of parameters in the risk model, and improving precision and accuracy rates to ensure resources are not wasted, but still serves purpose as a preliminary model.

Chapter 7

Conclusion

In this thesis, we approached three main research questions:

- To showcase the utility of GTM in clustering the breast cancer dataset, and analyse the clusters to showcase the difference groups of patients.
- Perform Risk stratification based on the clustering.
- Use Random Forest for predictive modelling.

This thesis successfully implemented the GTM on the breast cancer dataset, finding six distinct clusters. These clusters shown significant differences in not only the distributions of the modelling variables used to train, but also negative outcomes, noting statistical significance in recurrence. The predictive accuracy of models trained on these clusters varied, with low risk clusters exhibiting lower predictive power due to data limitations. This highlights the needs to focus on acquiring more data, and ensures the need for data completeness as discussed prior. Despite this, comparing to other research in recurrence, we see better results. The risk stratification method has been implemented and serves as a basis for future work.

7.1 Future Work

7.1.1 Limitations

One significant limitation of this study revolves around the data. Not only is the sample size less than ideal, due to restriction inclusion criteria and other challenges, we would also face issues in applying the model to new data sources, as variables may be unable

to be found. Different countries have different ways of recording data, and different treatment practise, and with the ethical considerations of the demographics of our data, would prove it hard to apply. Further work would need to be done in generalising the model, in order to gain more data and hopefully improve predictive accuracy.

One further limitation of the study is the short follow-up period. We use a period of three years, which is lower than standard duration's typically used, such as five or ten year follow ups. This shortened follow-up may not capture the full extend of various long term outcomes of the disease, hence could lead to invalid results. We decided to use three years due to data limitations, and were aware of potential issues going into the modelling. A longer follow up period would provide a more comprehensive understanding of the outcomes post treatment, and potential improve the robustness of our models. We also saw more difficulty in predicting mortality than we did with recurrence. One issue which may have contributed to this was the usage of all-cause mortality, which in turn encompasses deaths from potentially unrelated sources. This may have introduced bias to the model, in turn disrupting the predictive accuracy of the models.

7.1.2 Further Research

I have established a baseline method that has proven effective for breast cancer patients. One such further research would be as stated in the discussion, to follow up on identified patients and see how the model performs. Alternatively,

models, and experimented more, however this felt too much a task given my time frame. Ultimately, I am proud of my work.

To conclude, despite encountering numerous hurdles during my thesis, I gained a huge wealth of research skills and made a strong attempt at a difficult project. Whilst not everything went as I wished, I enjoyed the process and I am very thankful for the opportunity to work with real world cancer data provided to me by Clatterbridge and my supervisors alike. This has not only deepened my interest in research, but reinforced my passion for bio-informatics.

Bibliography

[1] Abreu, P. H., Santos, M. S., Abreu, M. H., Andrade, B., and Silva, D. C. (2016). Predicting breast cancer recurrence using machine learning techniques: A systematic review. *ACM Comput. Surv.*, 49(3).

[2] Ahmad, L. G., Eshlaghy, A., Poorebrahimi, A., Ebrahimi, M., Razavi, A., et al. (2013). Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform*, 4(124):3.

[3] Andrade, A. O., Nasuto, S., Kyberd, P., and Sweeney-Reed, C. M. (2005). Generative topographic mapping applied to clustering and visualization of motor unit action potentials. *Biosystems*, 82(3):273–284.

[4] Asri, H., Mousannif, H., Al Moatassime, H., and Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, 83:1064–1069.

[5] Belciug, S., Salem, A.-B., Gorunescu, F., and Gorunescu, M. (2010). Clustering-based approach for detecting breast cancer recurrence. In *2010 10th International Conference on Intelligent Systems Design and Applications*, pages 533–538.

[6] Bergom, C., Bradley, J. A., Ng, A. K., Samson, P., Robinson, C., Lopez-Mattei, J., and Mitchell, J. D. (2021). Past, present, and future of radiation-induced cardiotoxicity: refinements in targeting, surveillance, and risk stratification. *Cardio Oncology*, 3(3):343–359.

[7] Bishop, C. M., Svensén, M., and Williams, C. K. (1998). Gtm: The generative topographic mapping. *Neural computation*, 10(1):215–234.

[8] Boehm, K. M., Aherne, E. A., Ellenson, L., Nikolovski, I., Alghamdi, M., Vázquez-García, I., Zamarin, D., Long Roche, K., Liu, Y., Patel, D., et al. (2022). Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nature cancer*, 3(6):723–733.

[9] Cancer Research UK (2023). Mental health and cancer — coping with cancer. Accessed: 2024-08-13.

[10] Chen, J., Luo, Y., Xi, X., Li, H., Li, S., Zheng, L., Yang, D., and Cai, Z. (2022). Circulating tumor cell associated white blood cell cluster as a biomarker for metastasis and recurrence in hepatocellular carcinoma. *Frontiers in Oncology*, 12:931140.

[11] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the royal society interface*, 15(141):20170387.

[12] Clatterbridge Cancer Centre (2024). Patient records for breast cancer treatment. *Internal Data Source*. Collected July 2024. Consists of records between 2016 to present. Merseyside, UK. Authorized Personnel Only.

[13] Cossrow, N. and Falkner, B. (2004). Race/Ethnic Issues in Obesity and Obesity-Related Comorbidities. *The Journal of Clinical Endocrinology Metabolism*, 89(6):2590–2594.

[14] Cvetković, R. S. and Scott, L. J. (2005). Dexrazoxane: a review of its use for cardioprotection during anthracycline chemotherapy. *Drugs*, 65(7):1005–24.

[15] De Labry, L. O., Campion, E. W., Glynn, R. J., and Vokonas, P. S. (1990). White blood cell count as a predictor of mortality: results over 18 years from the normative aging study. *Journal of clinical epidemiology*, 43(2):153–157.

[16] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

[17] Enewold, L., Parsons, H., Zhao, L., Bott, D., Rivera, D. R., Barrett, M. J., Virnig, B. A., and Warren, J. L. (2020). Updated Overview of the SEER-Medicare Data: Enhanced Content and Applications. *JNCI Monographs*, 2020(55):3–13.

[18] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

[19] Gaspar, H. A., Hübel, C., and Breen, G. (2019). Biological pathways and drug gene-sets: analysis and visualization. *European Neuropsychopharmacology*, 29:S834.

[20] González-Castro, L., Chávez, M., Duflot, P., Bleret, V., Martin, A. G., Zobel, M., Nateqi, J., Lin, S., Pazos-Arias, J. J., Del Fiol, G., et al. (2023). Machine learning algorithms to predict breast cancer recurrence using structured and unstructured sources from electronic health records. *Cancers*, 15(10):2741.

[21] Greene, F. L., Balch, C. M., Fleming, I. D., Fritz, A., Haller, D. G., Morrow, M., and Page, D. L. (2002). *AJCC cancer staging handbook: TNM classification of malignant tumors*. Springer Science & Business Media.

[22] Gress, D. M., Edge, S. B., Greene, F. L., Washington, M. K., Asare, E. A., Brierley, J. D., Byrd, D. R., Compton, C. C., Jessup, J. M., Winchester, D. P., et al. (2017). Principles of cancer staging. *AJCC cancer staging manual*, 8:3–30.

[23] Group, C. R. (2022). Machine learning-based risk stratification model for mortality prediction in atezolizumab-treated cancer patients. *Cancer Medicine*, 11(5):1234–1245.

[24] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

[25] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

[26] James, F., Wootton, S., Jackson, A., Wiseman, M., Copson, E., and Cutress, R. (2015). Obesity in breast cancer–what is the risk factor? *European journal of cancer*, 51(6):705–720.

[27] Kerckhove, N., Collin, A., Condé, S., Chaleteix, C., Pezet, D., and Balayssac, D. (2017). Long-term effects, pathophysiological mechanisms, and risk factors of chemotherapy-induced peripheral neuropathies: a comprehensive literature review. *Frontiers in pharmacology*, 8:86.

[28] Klastersky, J., Paesmans, M., Rubenstein, E. B., Boyer, M., Elting, L., Feld, R., Gallagher, J., Herrstedt, J., Rapoport, B., Rolston, K., et al. (2000). The multinational association for supportive care in cancer risk index: a multinational scoring system for identifying low-risk febrile neutropenic cancer patients. *Journal of clinical oncology*, 18(16):3038–3051.

[29] Kruse, A. L., Luebbers, H. T., and Grätz, K. W. (2011). Evaluation of white blood cell count as a possible prognostic marker for oral cancer. *Head & neck oncology*, 3:1–5.

[30] Kushi, L. H., Doyle, C., McCullough, M., Rock, C. L., Demark-Wahnefried, W., Bandera, E. V., Gapstur, S., Patel, A. V., Andrews, K., Gansler, T., Nutrition, A.

C. S. ., and Committee, P. A. G. A. (2012). American cancer society guidelines on nutrition and physical activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity. *CA Cancer J Clin*, 62:30–67.

[31] López, N. C., García-Ordás, M. T., Vitelli-Storelli, F., Fernández-Navarro, P., Palazuelos, C., and Alaiz-Rodríguez, R. (2021). Evaluation of feature selection techniques for breast cancer risk prediction. *International Journal of Environmental Research and Public Health*, 18(20):10670.

[32] Macaulay, B. O., Aribisala, B. S., Akande, S. A., Akinnuwesi, B. A., and Olabanjo, O. A. (2021). Breast cancer risk prediction in african women using random forest classifier. *Cancer Treatment and Research Communications*, 28:100396.

[33] Maps, S.-O. and Secaucus, N. (2001). Self-organizing maps. *Inc.: Springer-Verlag*.

[34] Markey, M. K., Lo, J. Y., Tourassi, G. D., and Floyd Jr, C. E. (2003). Self-organizing map for cluster analysis of a breast cancer database. *Artificial Intelligence in Medicine*, 27(2):113–127.

[35] McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2):143–149.

[36] Mehta, L. S., W. K. E. B. A. B. T. M. and et al. (2018). Cardiovascular disease and breast cancer: where these entities intersect: a scientific statement from the american heart association. *Circulation*, 137:e30–e66.

[37] Miyamoto, S., Abe, R., Endo, Y., and Takeshita, J.-i. (2015). Ward method of hierarchical clustering for non-euclidean similarity measures. In *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pages 60–63.

[38] Osman, M. M. et al. (2022). Radiomic-based machine learning models for prostate cancer risk stratification using mri. *Journal of Medical Imaging*, 9(1):011001.

[39] Parikh, R. B., Manz, C., Chivers, C., Regli, S. H., Braun, J., Draugelis, M. E., Schuchter, L. M., Shulman, L. N., Navathe, A. S., Patel, M. S., and O'Connor, N. R. (2019). Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. *JAMA Network Open*, 2(10):e1915997–e1915997.

[40] Patnaik, J. L., Byers, T., DiGuiseppi, C., and et al. (2011). Cardiovascular disease competes with breast cancer as the leading cause of death for older females diagnosed with breast cancer: a retrospective cohort study. *Breast Cancer Res*, 13:R64.

[41] Pedersen, R. N., Esen, B. , Mellemkjær, L., Christiansen, P., Ejlertsen, B., Lash, T. L., Nørgaard, M., and Cronin-Fenton, D. (2021). The Incidence of Breast Cancer

Recurrence 10-32 Years After Primary Diagnosis. *JNCI: Journal of the National Cancer Institute*, 114(3):391–399.

[42] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[43] Poulter, N. (2003). Global risk of cardiovascular disease. *Heart*, 89:ii2–ii5.

[44] Sehhati, M., Tabatabaiefar, M. A., Gholami, A. H., and Sattari, M. (2022). Using classification and k-means methods to predict breast cancer recurrence in gene expression data. *Journal of Medical Signals & Sensors*, 12(2):122–126.

[45] Shakir, D. K. and Rasul, K. I. (2009). Chemotherapy induced cardiomyopathy: pathogenesis, monitoring and management. *J Clin Med Res*, 1(1):8–12.

[46] Shi, Y., Olsson, L. T., Hoadley, K. A., Calhoun, B. C., Marron, J., Geradts, J., Niethammer, M., and Troester, M. A. (2023). Predicting early breast cancer recurrence from histopathological images in the carolina breast cancer study. *NPJ Breast Cancer*, 9(1):92.

[47] Shukla, N., Hagenbuchner, M., Win, K. T., and Yang, J. (2018). Breast cancer data analysis for survivability studies and prediction. *Computer Methods and Programs in Biomedicine*, 155:199–208.

[48] Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. (2023). Cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(1).

[49] Smigal, C., Jemal, A., Ward, E., Cokkinides, V., Smith, R., Howe, H. L., and Thun, M. (2006). Trends in breast cancer by race and ethnicity: update 2006. *CA: a cancer journal for clinicians*, 56(3):168–183.

[50] Sublime, J., Grozavu, N., Cabanes, G., Bennani, Y., and Cornuéjols, A. (2015). From horizontal to vertical collaborative clustering using generative topographic maps. *International journal of hybrid intelligent systems*, 12(4):245–256.

[51] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globcan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, 71:209–49.

[52] Team, B. O. R. (2023). Pan-cancer risk prediction model using routine health check-up data: A machine learning approach. *BMJ Oncology*, 2(3):e000123.

[53] Urquhart, R., Cordoba, W., Bender, J., Cuthbert, C., Easley, J., Howell, D., Kaal, J., Kendell, C., Radford, S., and Sussman, J. (2022). Risk stratification and cancer follow-up: towards more personalized post-treatment care in canada. *Current Oncology*, 29(5):3215–3223.

[54] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[55] VanderPlas, J., Granger, B., Heer, J., Moritz, D., Wongsuphasawat, K., Satyanarayan, A., Lees, E., Timofeev, I., Welsh, B., and Sievert, S. (2018). Altair: Interactive statistical visualizations for python. *Journal of Open Source Software*, 3(32):1057.

[56] Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

[57] Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

[58] Wu, R. and et al. (2023). Evaluation of machine learning algorithms for the prognosis of breast cancer from the surveillance, epidemiology, and end results database. *PLOS ONE*, 18(1):e0280340.

[59] Yu, J., Gross, C., Wilson, L. D., and Smith, B. (2009). Nci seer public-use data: Applications and limitations in oncology research. *Oncology*, 23(3):288–295.

[60] Zajaczkowska, R., Kocot-Kepska, M., Leppert, W., Wrzosek, A., Mika, J., and Wordliczek, J. (2019). Mechanisms of chemotherapy-induced peripheral neuropathy. *International journal of molecular sciences*, 20(6):1451.