Exploring the Topology of Progressive Disease Data

A thesis submitted for the degree of
Doctor of Philosophy

By

Seyed Erfan Sajjadi

Brunel University London
College of Engineering, Design and Physical Sciences
Department of Computer Science

Table of Contents

List of Figures

List of Tables

Abstract

This thesis aims to investigate the crucial objective of improving the comprehension of clinical data structure, acknowledging its increasing importance. We initiate our investigation by exploring the historical context of topological data analysis, a fundamental approach that facilitates the extraction of the inherent topological structure of data. This methodology reveals discrete segments within the dataset, wherein specific segments may indicate the presence of diseases in their initial stages, while other segments may correspond to different subtypes of advanced diseases. The identification of areas has significant significance for clinicians as it enables a deeper understanding of patients' symptoms within the disease topology and facilitates the implementation of personalised treatments.

In the following section, we will go into the domain of Pseudo Time techniques, which enable the creation of temporal models from non-temporal cross-sectional data. These approaches provide useful insights by deducing temporal aspects of diseases. Nevertheless, the effectiveness of these methods relies heavily on the selection of suitable distance measures and labelling schemes that may effectively direct the process of trajectory modelling. The utilisation of clinical staging data, namely the categorisation of patients into "early stage" and "advanced stage," plays a crucial role in limiting the potential biases of pseudo-time models, hence guaranteeing the accurate representation of disease progression patterns.

The advancement of our inquiry involves the use of two separate methodologies in constructing temporal phenotypes using data topology analysis: topological data analysis and pseudo time-series. Using data on type 2 diabetes, we give evidence that topological data analysis can effectively identify trajectories that reflect various temporal phenotypes. Additionally, we show that pseudo-time series analysis can be used to infer a state space model that exhibits transitions between hidden states, each representing discrete temporal abnormalities. Significantly, both approaches emphasise the importance of lipid profiles in identifying these symptoms.

Our research presents the innovative TDA-PTS algorithm, which combines pseudo temporal and topological data analysis. The efficacy of the combined method is assessed on three different datasets, namely simulated data, diabetic data, and genomic data. This

evaluation demonstrates how the system effectively identifies unique temporal phenotypes in each disease by considering various trajectories throughout the progression of the disease.

Moreover, we explore the use of clinical staging data in order to construct robust and realistic trajectories. In this study, we utilise simulated data to showcase the accuracy attained in estimating the fundamental transition parameters using limited pseudo time approaches, which effectively mitigate the occurrence of unrealistic transitions. In the context of breast cancer pseudo time models, the trajectories are constrained by using the uniformity of cell size as a proxy of disease staging. This constraint leads to the development of models that more accurately depict the progressive increase in symptoms over time.

Finally, we employ these techniques to actual glaucoma data, therefore confirming the efficacy of the algorithm in accurately representing the advancement and categorisation of the condition. This study provides a thorough examination of illness dynamics within clinical datasets, presenting information in a chronological order that spans from background information to methods and outcomes. The findings of this research make a substantial contribution to the field, enhancing our comprehension and modelling of disease dynamics.

Acknowledgements

I want to firstly, and most importantly, thank my supervisor, Dr Allan Tucker for all his constructive supervision throughout my PhD study. He has been incredibly understanding and extremely supportive throughout a tough Covid-19 hit doctoral research. Dr Tucker has provided continuous encouragement, guidance and advice from our very first meeting, both academically and personally. Without his direction and provision, I would not have been able to complete this thesis. It has been a privilege to work with and to learn from him, invaluable lessons which I will cherish and utilise in future endeavours.

I am honoured to have Dr Stephen Swift as my second supervisor and my Research Development Advisor. He has been incredibly helpful and caring throughout all my reviews and has given me continued guidance throughout my research.

I would like to also thank my colleagues and all staff members at Brunel University that have helped, supported, and pushed me along to finishing my doctoral research.

Finally, I owe an immense gratitude to my wife, my newborn son, my parents and my family for their love and emotional support, their patience and continued encouragement throughout my entire PhD study. It can't have been easy to see and deal with the rollercoaster highs and lows during my research.

It has been a unique and enjoyable journey, and I am tremendously grateful for everything I have gained through out these five years.

Publications

Publications from the research presented in this thesis:

1. Dagliati, A., Geifman, N., Peek, N., Holmes, J., Sacchi, L., **Sajjadi, S**. and Tucker, A., 2019. Inferring Temporal Phenotypes with Topological Data Analysis and Pseudo Time-Series. *Artificial Intelligence in Medicine*, pp.399-409.
2. Dagliati, A., Geifman, N., Peek, N., Holmes, J., Sacchi, L., Bellazzi, R., **Sajjadi, S**. and Tucker, A., 2020. Using topological data analysis and pseudo time series to infer temporal phenotypes from electronic health records. *Artificial Intelligence in Medicine*, p.101930.
3. **Sajjadi, S**., Draghi, B., Sacchi, L., Dagliani, A., Holmes, J. and Tucker, A., 2020. Building Trajectories Over Topology with TDA-PTS: An Application in Modelling Temporal Phenotypes of Disease. *ECML PKDD 2020 Workshops*, pp.48-61.
4. **Sajjadi, S**. and Tucker, A., 2021. Exploiting Clinical Staging Data to Constrain Pseudo-Time Modelling of Disease Progression. *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pp.241-246.

Chapter 1 Introduction

1.1. Motivation

The National Health Service (NHS) states that the '*one size fits all*' approach to the care and treatment of patients is ineffective as everyone responds to diseases differently. Therefore, the need for personalised medicine arises, to better manage patients' health and to adapt treatment to the patient's genomic data. Through the 100,000 Genomes Project conducted by the NHS, the human genome in patients with diseases and cancer are being decoded to aid the future development of diagnosis and treatment of the diseases based on the genetic information. The goal is to characterise these genomic data to allow best-fit targeted therapies to be offered to patients. The human genome may contain 20,000 to 23,000 genes which provides a vast variety of information that needs to be characterised in order to be used [1]. Subtle gene differences cause large differences in health. Understanding these differences can lead to an improved method of diagnosing, preventing, and treating many types of health conditions [2]. Cancer is an example of a health problem, which is caused by the division of abnormal cell populations that can be fatal when interfering with an individual's body function [3]. Cancer and many other systemic health conditions are as a result of genetic changes in the genome. Hence, it makes sense to consider the analysis of the human genome to find the exact genetic changes responsible. The human DNA can be sequenced at a much lower cost, which results in the classification of a larger cohort of genome data. The vast data available from the individual's genome allows for a better understanding of the genetic risks that the individual may possess. Subsequently, this will allow a more precise diagnosis and permit the use of personalised treatments [4]. Once the cancer genome is sequenced, it can be compared with a "normal" genome (from blood or saliva) to find changes that occur with the goal to finding an effective drug. Hereafter, the patient can be vaccinated against the change that appear in their genome.

Owing to the affordability and accessibility of sequencing the human genome, research in the bioinformatics field has been dominated by an overwhelming amount of

experimental data [5]. Tan & Gilbert state that due to the vast amount of complex data, machine learning and artificial intelligence techniques must be used to discover and mine the information from these databases. Baldi & Brunak state "*As a result, the need for computer/statistical/machine learning techniques is today stronger rather than weaker.*" [6]. Machine learning has a valuable nature of thriving in domains where we have an enormous data set with a limited amount of theory, which is what we encounter in bioinformatics [8]. Molecular biology is an ideal field for the use of machine learning techniques [7].

Machine Learning uses algorithms that can learn from a dataset and experience with interacting with the dataset in terms of tasks and a performance measure [9]. Clinical trials can be used to capture information that may be used to demonstrate hidden characteristics of health problems, diseases processes or to discover how a disease progresses. Cross-sectional studies allow us to see a snapshot of a disease process for a large population but with the limitation of not enabling the temporal nature of the disease to be modelled. However, longitudinal studies overcome this limitation and allows for the development of the disease process to be investigated over time. This process can become time consuming and expensive, especially if conducted for a large population, with the disadvantage of capturing a small window within the disease process; hence the need for developing an effective method to understand and discover insightful information about an unknown dataset.

Seeing the effects of covid-19 reemphasises the demand for progressive models to enhance the understanding of the underlying processes within relatively unknown diseases. Disease progression varies in complexity and can form different trajectories from healthy to many intermediate and eventually advanced stages depending on the disease type and the individuals. Understanding risk factors leading to certain diseases and gaining insight to how these diseases progress will give clinicians the ability to provide a more accurate diagnosis and prognosis. As a result, this valuable information can be used for earlier detection and more effective interventions.

This thesis focusses on Machine Learning methodologies for modelling disease progression and dealing with complications that are inherent with this type of modelling.

It will look to exploit advantages of Topological Data Analysis (TDA) and Pseudo-Time Series (PTS) methods to create a novel method of modelling the temporal nature of disease progression. This will be done along with hidden Markov models (HMMs) in order to model disease and find key regions in the disease trajectory, which can be tested on both real and simulated biomedical data.

1.2. Aims and Objectives

The aim of this thesis is to create a novel method to learn disease progression trajectories to allow intermediate stages to be identified and insights to the nature of the progression to be discovered. Furthermore, these trajectories will be constraint based on clinicians' knowledge, enabling a semi-supervised machine learning to be adopted. The objectives of this thesis are as follows:

1. **Development of a Novel Methodology:**

   Formulate and develop an innovative method for learning disease progression trajectories, aiming to uncover intermediate stages and gain nuanced insights into the nature of disease progression.

2. **Integration of Clinician Knowledge:**

   Integrate domain-specific knowledge from clinicians to establish constraints on disease trajectories, fostering a semi-supervised machine learning approach. This incorporation ensures alignment with expert insights, enhancing the clinical relevance and interpretability of the proposed model.

3. **Topology Mapping through TDA:**

   Employ TDA to systematically map out the underlying topology of diseases. This objective seeks to provide a structural understanding of the data, laying the foundation for subsequent trajectory construction.

4. **Trajectory Construction with Multifaceted Approaches:**

   Utilise a synergistic combination of bootstrapping, hidden Markov models, and pseudo-time series to construct disease progression

trajectories within the mapped topology. This multifaceted approach aims to capture the complexity and dynamics inherent in disease progression.

5. **Identification of Intermediate Disease Stages:**

   Design the methodology to specifically identify intermediate stages within disease trajectories. This objective addresses the challenge of recognising subtle transitions in disease progression, contributing to a more nuanced and comprehensive representation.

6. **Insight Discovery through Trajectory Analysis:**

   Perform in-depth analysis of the constructed trajectories to uncover valuable insights into the nature of disease progression. This exploration aims to reveal patterns, trends, and relationships that may not be apparent through traditional analytical methods.

7. **Validation and Evaluation:**

   Rigorously validate and evaluate the proposed methodology using both simulated and real-world biomedical datasets. This objective ensures the reliability and generalisability of the developed model, establishing its efficacy in diverse clinical contexts.

8. **Documentation and Communication of Findings:**

   Document the entire process, from methodology development to results analysis, in a clear and comprehensive manner. Effectively communicate the findings through academic publications, contributing to the broader scientific community's understanding of disease progression modelling.

By addressing these objectives, the thesis endeavours to make a significant contribution to the field by advancing the methodology for learning disease progression trajectories and providing valuable insights that can inform clinical decision-making.

1.3. Thesis contributions

This subsection recounts the thesis's main contributions, describing a transforming journey through innovative technique. This section highlights the various factors that help determine disease development trajectories, revealing new insights and correlating with clinical expertise.

1. **Advancing Methodology for Disease Understanding:**

   Presented a comprehensive exploration of Topological Data Analysis and Pseudo-Time Series, showcasing their utility in constructing temporal phenotypes and modelling disease progression.

2. **Innovative Algorithm Development:**

   Introduced a pioneering combined TDA-PTS algorithm, providing a formal definition and pseudo code for the novel approach in data analysis.

3. **Constructing Informative Disease Models:**

   Developed topological models from cross-sectional data, enabling the visualisation of data shapes, identification of intermediate disease stages, and enhanced understanding of disease progression dynamics.

4. **Unveiling Disease Trajectories:**

   Applied Pseudo-Time Series to simulated and real-world biomedical datasets, constructing disease progression trajectories. Mapped trajectory transitions through dataset topology, offering insights validated against the dataset's true state for model reliability.

5. **Enhancing Model Robustness:**

   Improved the model by introducing constraints in CBPTS, utilising prior disease knowledge to guide trajectories. This refinement, demonstrated on simulated and real-world breast datasets, ensures robustness and reliability.

6. **Application to Progressive Diseases:**

   Applied the novel methods to three distinct glaucoma datasets, demonstrating the effectiveness of the model in understanding and representing clinically proven progressive and irreversible diseases.

1.4. Thesis outline

This thesis has the following layout:

- Chapter 2 presents a literature review of how machine learning is used in medicine and provides a background on how it can be utilised to investigate disease progression. Different methods of machine learning techniques are explored to assess their capabilities for this research, which will be used to form the objectives of this thesis.

- Chapter 3 focuses on key concepts such as Topological Data Analysis and Pseudo-Time Series, which sets a foundation of the work that this thesis aims to build upon. These methods of machine learning will be explored and how it intends to build reliable models of disease progression will be demonstrated and explained.

- Chapter 4 explores the novel TDA-PTS algorithm by initially describing the algorithm along with defining the pseudo-code. Subsequently, the different datasets are presented and how the experiments are conducted will be outlined. Finally, the results producing disease progression trajectories are presented, analysed, and discussed.

- Chapter 5 introduces the use of constraints to improve the disease progression trajectories with the implementation of CBPTS. Similar to the previous chapter, the method is described, and results are presented, analysed, and discussed.

- Chapter 6 applies the TDA-PTS and CBPTS methods to real-world cross-sectional datasets. The results are presented and discussed in a clinical setting

- Chapter 7 concludes the thesis by summarising the main accomplishments in this study, limitations and potential areas for further research that can be explored.

Chapter 2 Literature Review

2.1. Chapter Outline

The objective of this chapter is to review the relevant literature, which contributes to the motivation behind this thesis. This chapter aims to provide an explanation and analysis of the primary machine learning approaches that are frequently employed, in order to provide a solid foundation for the present research. This chapter is organised as follows: Section 2.2 provides an introduction, by discussing supervised and unsupervised machine learning. Section 2.3 describes classification and examines the main types of classification techniques. Section 2.4 describes clustering and examines the main types of clustering techniques. Section 2.5 delves into Topological Data Analysis by breaking it down into the different methods within this machine learning technique. Section 2.6 describes Pseudo-Time Series Analysis. Section 2.7 depicts the uses and limitations of cross-sectional and longitudinal datasets. Section 2.8 analyses the core model performance evaluation techniques. Section 2.9 provides a summary.

2.2. Introduction

There are two types of machine learning techniques that can be used for data mining: supervised and unsupervised learning. Supervised learning is when the system or the learner has some previous knowledge of the data it will deal with, also known as where the output is a priori. Unsupervised learning is where the system or the learner has been given no information about the data it will be dealing with or the output, this is known as a posteriori [5].

Initial research has shown that an analytical technique can be used that relies on the learned Bayesian networks to identify the gene bands that will be affected by a certain treatment [10]. Using a Bayesian method will provide a platform for sequential learning and can take a ‘learn as we go’ approach. Bayesian can be used as an adaptive approach, which is helpful for outcome adaptive randomisation, to allocate further subject patients into a more effective treatment method as there are an increased number of data sets. It

can also be used as an interim monitoring technique for early diagnosis due to futility. Finally, Bayesian can be used as an adaptive sample size estimation through calculation of the probability for a successful trial to reach a definitive result. Results should demonstrate data integration between genome, protein, and drug via a machine learning method [11].

Machine learning techniques can be used for temporal phenotyping, which is used to distinguish clinically expressive information from patient data over time. Being able to identify temporal phenotypes of patient data and enabling different sub-groups of diseases to be linked can aid researchers and clinicians to make informed diagnoses and personalised procedures. Furthermore, being able to generate this clinically expressive information will help in better understanding the diseases, their links, and their progressions. Temporal graphs obtained from electronic health records are used for temporal phenotyping but the interest in this research is to identify trajectories through the different data sets with the aim of finding links and discovering disease progression in time [12], [13]. There are many variations of machine learning techniques, each with their own benefits and limitations. These machine learning techniques must be critically analysed, and the most effective technique selected. However, to find the most effective and certainly novel approach, a combination of methods can be used so therefore, the study becomes versatile and reflective. Feio et al used a combination of techniques from Support Vector Machines (SVM), Multi-layer Perceptron and K-Nearest Neighbour (KNN) for their studies classification task. They have stated that the combined use of the three machine learning techniques gave more effective results in prediction tasks compared to if they were applied individually [14].

2.2.1. Supervised and Unsupervised Learning

Supervised or unsupervised learning can be further categorised to include reinforcement learning. The main difference in choosing which type of learning is applicable, is the existence of labels within the dataset that will be used to train the machine learning algorithm. Supervised learning requires a certain kind of

predetermination of the output attributes along with a prior understanding of the input attributes [56]. A supervised machine learning algorithm tries to predict and classify the predetermined attributes as well as their accuracies, misclassification, and other performance measures. When the algorithm reaches a satisfactory performance level, it will halt the learning process [57]. This type of learning algorithm will initially use the training dataset to perform analytical tasks and then it will build contingent functions so that it can map new instances of the attribute [58]. Commonly the dataset is split and roughly 66% is used for the training set as it helps to achieve the desired result but also accommodating for the amount of computational time it requires to avoid being too demanding [59]. Whilst using an unsupervised approach, all the variables within the dataset are usually used in the analysis as inputs so that patterns can be discovered without requiring a target attribute. This method makes unsupervised learning suitable for clustering and association mining techniques. On the other hand, during a supervised learning approach, a target attribute and prior knowledge about what outcome the user wants to achieve is available. Hence, certain variables can be omitted as they may have little significance to the analysis, and it will help reduce dimensionality, computational time, and complexity. This makes supervised learning appropriate for classification and regression analysis [8], [56]. The two methods of learning can be combined by using unsupervised learning for creating labels or identifying rules that accurately represent relationships between attributes within the data, which then can be used to implement supervised learning tasks [60]. The decision on which type of learning to use does not need to be binary. A combination of both methods can be used in a semi-supervised learning technique, which can exploit benefits from both approaches. In the end, the approach taken depends on what the user is trying to achieve and most importantly, what data is being used and how is it structured.

2.3. Classification

2.3.1. k-Nearest Neighbour

The nearest neighbour classifier is seen as a simple but extremely powerful tool for classification tasks. It can classify unlabelled data point or sections of the data by assigning them the class of similar labelled data points or sections of data [15]. This technique is successfully used to identify patterns in genetic data, computer vision applications and can be used in a commercial environment such as advertising. The nearest neighbour approach to classification is illustrated by the k-nearest neighbour algorithm (k-NN). Although considered the simplest machine learning algorithm, it is very widely used.

Similar to any machine learning techniques, k-NN comes with its own advantages and disadvantages. As stated before, the k-NN algorithm is simple and effective, which is an attractive trait when dealing with straight forward and simple data. K-NN makes no assumptions about the underlying data distribution, allowing an unbiased approach; with a fast-training phase of building the machine which is a huge advantage. However, the classification phase is relatively slow, and the nominal features and missing data require additional processing, which creates more delays. Another weakness of the k-NN algorithm is that it does not produce a model, limiting the ability to understand how the features are related to the class. Finally, an appropriate "k" value must be selected, which can cause some problems known as *bias-variance trade-off*. Selecting a large "k" value will diminish the impact caused by noisy data, but this can bias the learner as it will create the risk of ignoring small patterns that can be very important to the data [15].

To get a better understanding of the effects of selecting different "k" values, we can assume a very large value is set, as large as the total number of observations in the training dataset. When assigning labels, the most common class will get the majority vote. Consequently, the model will always predict the majority class and would disregard the nearest neighbours. However, if we assume the opposite extreme and set a very small "k" value, the unlabelled data will be influenced by a single nearest neighbour. Subsequently, if the single nearest neighbour to the unlabelled example is incorrectly labelled, then the

model will predict an incorrect class for that data point, even if there are several other nearest neighbours that would have voted for a different class assignment.

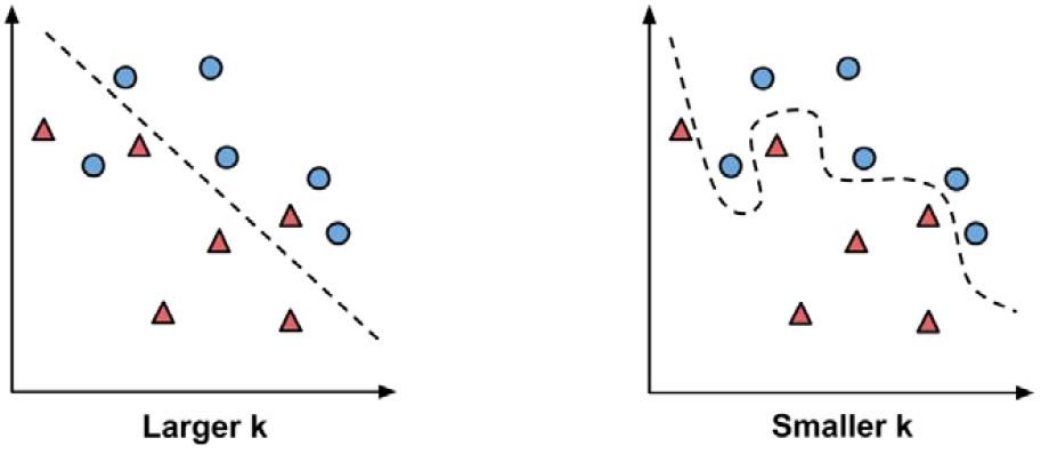


Figure 1: Visual representation of effects of choosing a large or small "k" value

Figure 1 illustrates the effects of selecting both extremes for values of “k”, where the boundary decision is depicted by a dashed line. It can be seen that a smaller value will be able to generate a more complex boundary decision that will be sensitive to all the data points in the training set. However, the issue lies with not knowing whether the curved or straight boundary will represent the true underlying concept of the entire data better. A common approach is to start with a “k” value equal to the square root of the number of training examples and then this value can be altered by testing several “k” values in a trial-and-error method and choosing the value that delivers the most effective classification performance for the dataset.

K-NN relies heavily on the training phase, which is merely storing the training data verbatim instead of actually learning. This method speeds up the training phase, which was one of the strengths of this algorithm, but the decision-making process is relatively slow in comparison to the training. This type of “lazy” learning is known as instance-based learning. Instance-based learners use a non-parametric learning approach, which means no parameters are learned about the data. Non-parametric methods do not allow the user to understand how the data is used by the classifier. However, not knowing any

parameters will give the learner the opportunity to discover natural patterns in the data instead of forcing the data into a potentially biased state.

The k-NN algorithm is applied through numerous steps:

1. **Data collection:** The data must be collected in a format that is useable for the researcher and stored in a legible format by the programming language where the k-NN algorithm will be applied.
2. **Data exploration and preparation:** This is the most vital step for implementing any machine learning approach. The data that has been collected is almost never ready to be used by the machine learning approach without first being explored and prepared. From this exploration we can find out characteristics about the data and shine some light on the relationships. After the data is explored, "errors" and imperfections within the data should be visible. Subsequently these "errors" and imperfections will need to be dealt with before starting the next step. Some examples of issues to deal with are, data normalisation, anomalies, missing data and different scaled data. Finally, the data can now be prepared by creating training and test datasets. Determining the split proportion for the training and test datasets can be tricky as there is no perfect proportion. The user must consider enough data points so that the data can be trained but not too much so that the test dataset is limited. Furthermore, a method for splitting must be determined, will it be random, or will there be some sort of restraints applied before splitting.
3. **Model training on the data:** Now that the data is prepared and split into training and test datasets, model building can commence on the training dataset. However, as explained before, the training phase for the k-NN algorithm does not involve building a model and instead it simply involves storing the input data in a structured format.
4. **Model performance evaluation:** After the model has been trained, the next step of the process is to evaluate how effectively the model predicts the classes of the test dataset points by comparing the prediction to the actual classification of the data points. Evaluation of accuracy, sensitivity and specificity can be done via a

receiver operating characteristic curve, also known as ROC curve, to see how well the model performs in terms of accuracy, sensitivity, and specificity. From this evaluation, a confusion matrix (CM) is achieved that shows the outcome in a table format:

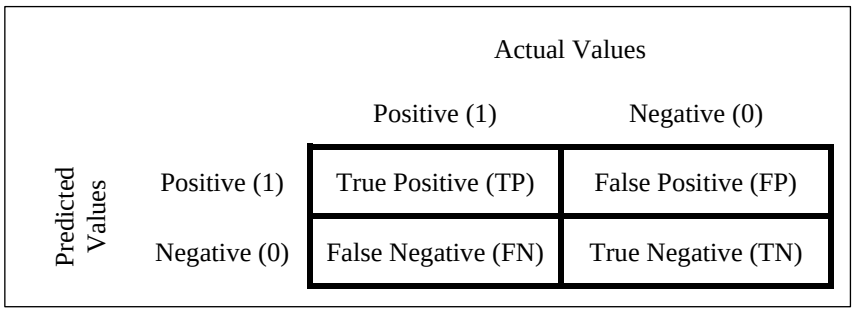


Figure 2: Example of a Confusion Matrix

Accuracy is the measure of how often the model predicts correctly. It is obvious that a high accuracy is preferable when evaluating any model, but we must consider if the model actually predicted the accuracy correctly or if it "cheated." Therefore, a model that has a 100% accuracy rate raises a warning as this is highly unlikely to be achieved so further investigation would be necessary.

Sensitivity is the percentage of positive outcomes that the model has predicted correctly. This value can be determined from the confusion matrix with the following formula:

$$Sensitivity = \frac{TP}{(TP + FN)}$$

Specificity is the percentage of negative outcomes that the model has predicted correctly. This value can be determined from the confusion matrix with the following formula:

$$Specificity = \frac{TN}{(TN + FP)}$$

In a medical and disease prediction setting, sensitivity and specificity are the two most widely discussed measures of biomarker efficacy. A negative result

from a highly sensitive test (>90%) provides substantial evidence against the presence of the target disease. Tests with a high degree of specificity (>90%) provide few false positives, therefore a positive result is highly suggestive of the presence of the disease or condition of interest. One major benefit of sensitivity and specificity is that they are unaffected by the illness prevalence, or the percentage of cases in the study population at a given period. However, they can be biased in one of two ways: either the sensitivity of the test decreases as the severity of the disease decreases, or the specificity decreases because there are other plausible causes for a positive test result. In addition, the diagnostic threshold chosen causes an inverse correlation between the two metrics, so that increasing the threshold increases specificity at the expense of decreasing sensitivity. Greater cumulative sensitivity and specificity may be achieved than with single molecules by combining individual biomarkers into panels that better represent the complexity of disease.

It is important to remember that sensitivity and specificity convey contradictory clinical information. For example, if a person has the target disease, then the need to answer the question of what is the likelihood that the test result is negative (specificity) or positive (sensitivity) would seem obsolete. Having knowledge of the presence of a disease would negate the need for a diagnostic test designed to detect that disease.

5. **Model performance improvement:** During this stage the model is altered with the aim of improving its prediction capabilities suited towards the researcher's need. Many attributes of the model can be changed such as the using a different method for dealing with anomalies or missing values, alternative methods for scaling and rescaling the numeric features, feature selection and in the case of k-NN algorithm, testing different "k" values.
6. **Final model testing:** The optimised model will then be tested against the benchmark that the initial model achieved. Stages 4-6 can be repeated n number of times depending on what the cut-off point would be for the researcher and when the most effective model has been achieved.

To summarise, k-NN does not learn anything as it simply stores training data verbatim. The unlabelled test examples are consequently assigned to the most similar records in the training set using a type of distance function. Then, the unlabelled example will be assigned to the label of its nearest neighbour. Even though this approach to machine learning is a quite simple, it has the capability of undertaking very complex tasks.

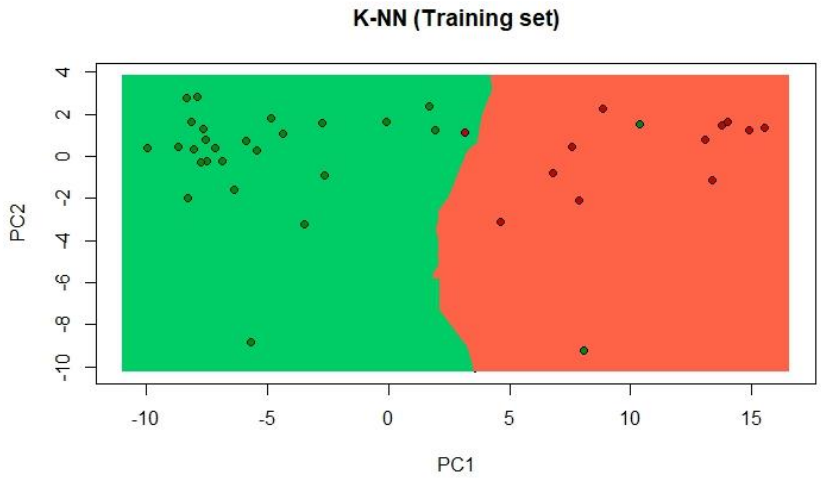


Figure 3: K-NN classification of cancer patients in sample training dataset (Green points=benign patients, Green area=benign classification area) (Red points=malignant patients, Red area=malignant classification area)

Figure 3 shows that the classifier performs well except for 3 data points. Even though the model was not 100% accurate, it performed reasonably well. The classifier predicted 2 data points that were benign as malignant, which is a false positive, the effects of which is not problematic. However, the issue arises when 1 malignant patient has been identified as benign, also known as a false negative, which can obviously be very problematic. The model needs further evaluation to see the implications of this false negative results and if the model should be tweaked.

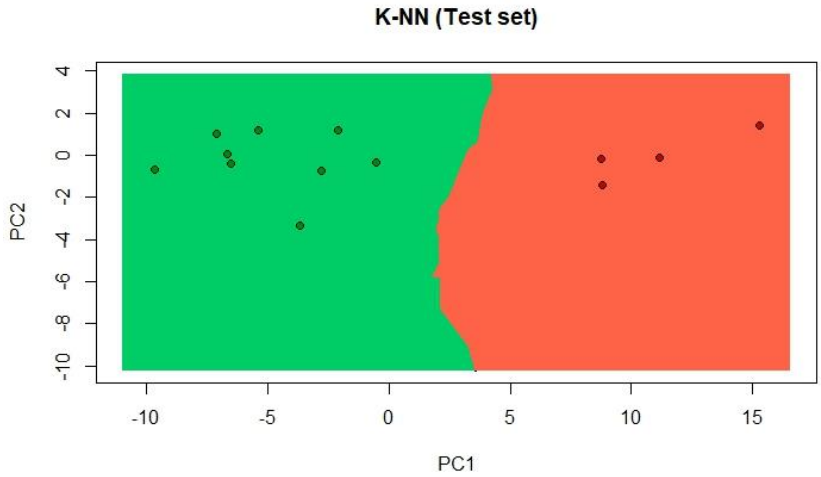


Figure 4: K-NN classification of cancer patients in sample test dataset

On the other hand, results from the classifier on the test dataset in Figure 4 shows that the patients have be classified accurately and the model may not require significant changes or tweaks.

2.3.2. Bayesian Network

Bayesian Networks are probabilistic graphical models that show the probability relationships between a set of variables. A graphical structure consisting of nodes with conditional probability distributions can be used to represent this network quantitatively or qualitatively. Probabilistic queries can be solved by using Bayesian Networks. Bayesian Networks are a popular machine learning method, where it is used for computational modelling of knowledge. These networks are present within the fields of medicine, bioinformatics, and any decision-making system. Being able to transparently model the relationship among variables makes this classification tool very effective as well as being able to obtain the uncertainty in the knowledge and data [16].

Bayesian Networks are very popular in the medical field due to the networks' ability to combine data with clinical expertise. An informative prior can be used by constructing and regularly updating a model via Bayesian techniques as soon as more data becomes available, which will result in a posterior model [17]. Bayesian Networks have the advantage to obtaining and dealing with uncertainty effectively. This makes it useful for

microarray data analysis even when the data is un-normalised, which makes it effective against gene microarray data. Additionally, this classifier can be used to deliver valuable information about biological pathways when analysing differential gene expression [18].

Bayesian Networks can work with different data types as well as being able to model combinations of different types of data in a single model, which makes this technique very desirable in medicine as it can combine clinical and gene expression data. Bayesian Networks are very good at presenting the model at a given time with a system that is in an equilibrium and stationary state. However, systems can change over time and obtaining knowledge about how the system evolves over time can be very insightful, meaning another tool must be used with the ability to model dynamic systems.

A dynamic Bayesian Network (DBN) is a Bayesian Network extended with additional mechanisms that are capable of modelling influences over time also known as time-series or sequential data [19]. The temporal addition of Bayesian Networks does not mean that the parameters or structure of the network changes dynamically, but it simple means that a dynamic system is modelled. In other words, the underlying process, modelled by a dynamic Bayesian Network, is stationary, which is a model of a random process.

Dynamic Bayesian Networks can have many applications in medicine such as aiding the prediction of early presence of Osteoarthritis: a degenerative knee condition that can cause pain, disability, and reduction in bone mass. As this is a gradually degenerative condition, dynamic Bayesian Networks are very effective at analysing the progression of the condition over time [20]. This Machine Learning technique can be used to identify how treatment affects the condition, whilst modelling disease progression as well as detecting the development of any complications [21].

Simple Bayesian Networks can be applied to model the relationships between Hydration and Dialysis Sessions on Dry Weight when trying to monitor the treatment of renal failure patients. Subsequently, the effect of past events affecting the patient's present state can be explored by dynamically modelling the treatment with dynamic Bayesian Networks [22].

Glaucoma is a common eye condition where the optic nerve connecting the eye to the brain, becomes damaged. The progression of the disease can be modelled by using a

dynamic Bayesian Network that can cluster time series sections alongside learning the networks' structure and parameters [23]. Dynamic Bayesian Networks can also be implemented to model Neuronal Interactivity for activation patterns in the brain as well as being used to optimise treatment in intensive care units [24], [25]. It is evident that Bayesian Networks have lots of effective abilities but are often slower than other deterministic methods, often making them unsuitable for large models and/or large data sets.

2.3.3. Neural Network

Neural Networks (NNs), which are also known as Artificial Neural Networks (ANNs) or Simulated Neural Networks (SNNs), are a subset of machine learning that is the centre of deep learning algorithms. This machine learning technique is inspired by the human brain by mimicking the structure of neurons in the brain and how they signal to one another [26]. Neural Networks are effective at dealing with high level of complexity in data that is obtained experimentally, which consists of nodes or neurons that can receive, process, and transmit signals to one another. Simple Neural Networks has an advantage of being able to learn from previous networks, something that is not easily possible in other classifiers. Two layers are present in a simple network: the first an input layer and the second an output layer . Artificial Neural Networks also consist of node layers resembling the simple network, but also including one or more hidden layers between the input and output layer. Individual nodes or artificial neurons are connected to one another along with an associated weight and threshold. The data from one layer will be sent to the next layer of the network if any individual node is above the specified threshold value. If the threshold is not met, no data is transferred to the next layer of the network. Neural Networks depend on training data and previous examples to be able to learn and improve their future classification accuracy over time. Consequently, when the model has been fine-tuned, it can become a very powerful tool in machine learning by allowing classification tasks to be done rapidly, something evident in Google's search algorithm. Neural Networks are popular with solving various real-world problems in business,

education, economics, healthcare and many other sectors as it benefits from factors such as accuracy, performance, fault tolerance, processing speed, latency, volume, scalability and convergence [27], [28]. The layers within Neural Networks are independent from one another, which contain numerous bias nodes that are always set to one. These bias nodes act like the offset in linear regression. Along with the normal inputs that are received by the network node, a bias function is utilised to offer the node with a constant value that is trainable. These bias values are key to allowing the activation function to move either left or right, which can be vital for the network's success in training. The input and output nodes will mirror the features that are inputted as well as the output classes [29]. Neural Networks are good at dealing with problems that have many parameters as well as being able to classify objects that are spread within complex highly dimensional spaces [30]. An Artificial Neural Network has been effectively used in a protein study to be able to analyse the development of drug resistance in human immunodeficiency virus 1 (HIV-1) [31]. Neural Networks have the advantage of being an effective decision-making tool for diagnosis and prognosis due to being able to handle highly dimensional data, a clear benefit for healthcare and medicine [32]. This benefit is further evident when trying to achieve predictions on DNA sequence level, protein sequence level and protein structure level [33]. Assessing medical outcomes and utilisation of resources in intensive care units can be undertaken with Neural Networks [34].

Similar to all machine learning techniques, Neural Networks also have limitations. Single layered Neural Networks can only classify signals that are linearly separable, and their limitation is apparent when trying to accomplish complex mappings [35]. To overcome this, several hidden layers can be implemented between the input and output layer. This approach is known as Multilayer Neural Networks (MNNs), which is a fully connected network with all nodes from one layer connecting to the next layer, allowing it to be extended with any number of hidden layers and finally connected to the output layers. The number of hidden layers and the nodes that needs to be present in order to achieve the desired output mapping depends on the complexity of the input pattern space to be partitioned [35]. Multilayer Neural Networks use backpropagation as the main aspect of their algorithm. However, this can be seen as a limitation due to it being

vulnerable to overfitting the training data at the cost of decreasing the generalisation accuracy over other new data [9]. In the field of medicine, Neural Networks are very prominent such as, modelling prognosis in breast cancer patients [36]–[38] adaptive control of mean arterial blood pressure [39], supporting the diagnosis of heart diseases by using clinical records to train and test the model [40], modelling surgical decisions on traumatic brain injury patients [41] and low back pain classification [42] are just some examples of Neural Networks in medicine.

Neural Networks, especially Artificial Neural Networks, are an effective tool for predictive tasks but compared to Bayesian Networks they can be slow in the training and application phases [43]. As mentioned before, Neural Networks are susceptible to overfitting, where the model focuses on a subsection of the data that is irrelevant to the classification due to having too many parameters.

2.3.4. Decision Tree (Random Forest)

The foundation of a random forest classifier comes from the general decision tree classifier. simply put, a decision tree can be thought of a series of yes or no questions that are asked about the data with the aim of eventually leading these questions to predicting the class that should be assigned to the data. This model is very interpretable as it classifies data similarly to how we approach an unknown data or problem by asking an order of questions about the unknown data or problem until we eventually arrive to a decision or a solution.

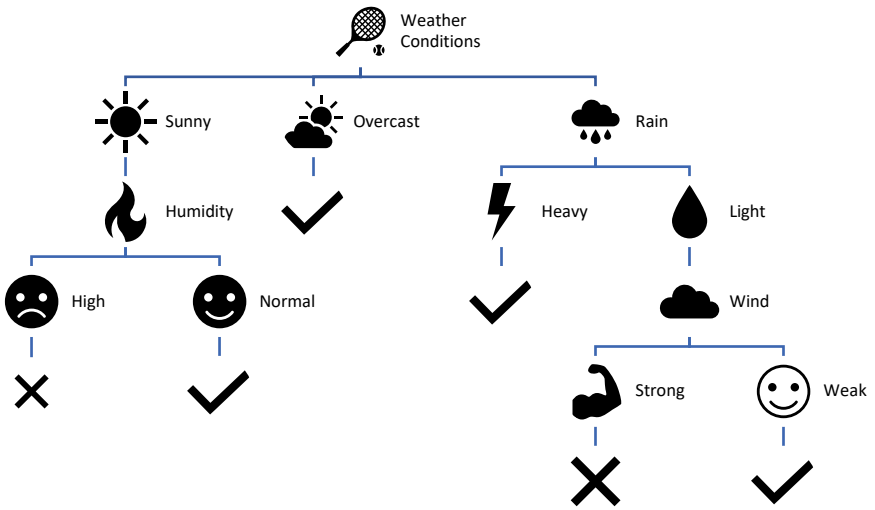


Figure 5: Decision tree for deciding whether to play tennis depending on weather conditions.

Random forest classifier is the growing of a group of trees and allowing them to vote for the most popular class. To be able to grow these groups of trees, random vectors are produced which control the growth of every tree in this group [44]. This approach uses the base principles of bagging with random feature selection with the aim of adding diversity to the model compared to standard decision tree models [15]. The beauty of this machine learning technique is the power and versatility that random forests possess. The group of trees only use a small but randomly selected feature set, which means it can easily handle very large datasets avoiding the high-dimensionality issues. Random forest is one of the strongest decision tree sectors and a strong candidate for being one of the most effective classifiers. As stated, the capability of handling large datasets with large numbers of features or examples makes it very desirable in the classification field. Furthermore, it can handle missing or noisy data as well as continuous or categorical features and is excellent at selecting the most important features. Therefore, random forest is considered an all-purpose model that will perform adequately for a wide range of problems. However, random forest is not easily interpretable like the decision tree, and it may need some extra work so that the model is finetuned to the dataset used. Nonetheless,

with these minor weaknesses, random forest is one of the most popular machine learning methods.

Like any classification tool, the data must be pre-processed and prepared so that the random forest classifier can be applied. During the training process each tree in the random forest will learn from a random sample of the dataset. Bootstrapping method is used, which means that some of the samples may be used multiple times in a single tree. The aim with using this method is so that each tree is trained on a different sample of the dataset. Subsequently, each tree may have a high variance with respect to a specific part of the training data, but as a whole, the forest will have a lower variance and avoiding the increase of any bias. Now that the random forest is trained, the model can be tested. During the testing phase, predictions are made from averaging the predictions that are generated by each of the trees. This type of training individual trees and testing them through a bootstrap method is known as bootstrap aggregating or bagging for short [45].

Classification methods that have been investigated, namely k-nearest neighbour, Bayesian network, neural network, and decision tree, it becomes evident that rigorous data preparation is a crucial factor for achieving optimal results. Every technique undergoes a distinct training procedure, whether it is based on proximity-based neighbours, probabilistic relationships, complicated neural structures, or hierarchical decision rules. Notwithstanding their varied characteristics, it is evident that the fundamental premise remains unambiguous: the achievement of effective categorisation is contingent upon meticulous preprocessing and customised learning methodologies. As we go to the subsequent segment about clustering, these fundamental understandings lay the groundwork for revealing intrinsic patterns inside datasets.

2.4. Clustering

The focus is redirected from the process of classification to the exploration of clustering approaches that demonstrate exceptional ability in identifying intrinsic structures within datasets. The exploration highlights the significance of K-means, hierarchical, Gaussian mixture, fuzzy c-means, and DBSCAN algorithms. Each method

offers a unique technique for categorising data points by identifying similarities and patterns, revealing underlying structures that go beyond predetermined labels. In the process of elucidating the complexities inherent in clustering approaches, our objective is to reveal the subtleties associated with their ability to interpret intricate datasets and enhance our comprehension of fundamental patterns and interconnections.

Clustering is essential a simple method for machine learning, whereby the data points or population is divided into several different groups. Within these groups there are data points or sectors of the population that will have similar attributes that the researcher will set. The aim is to simply take the dataset, segregate them into groups with similar traits and assign them into clusters. Clustering is an unsupervised machine learning technique, which means that it clusters the data without being told how the groups should be like beforehand. This type of machine learning is very beneficial when the researcher has no prior knowledge of the data and does not know what they are looking for, which makes this technique excellent for knowledge discovery rather than prediction. Clustering has many applications such as segmenting customers into groups with similar buying patterns so that advertising and marketing can be tailored towards them to maximise sales and profits. It can also be used for spotting irregular behaviour, such as unauthorised network intrusions, by finding patterns of use that are not like existing clusters. Generally, clustering is very useful when the dataset is varied and diverse and they can be segregated into smaller groups, which in turn provides meaningful data structures that can be used as insight to patterns and relationships within the dataset.

2.4.1. K-means

The k-means clustering algorithm is the most commonly used clustering technique [15]. Due to its popularity, it has been used for many years and it has been used as a benchmark to create branches on more complexing clustering techniques. Like k-NN, the k-means clustering algorithm consists of a “k” value which will be determined by the researcher. The algorithm will assign each of the examples or data point to one of the “k” clusters, with the aim of trying to reduce the number of differences that the data points

have within each of the cluster and also trying to increase the differences between each of the cluster [46]. There may be a lack of feasibility to create optimal clusters across all the population if the “k” value and the amount of data points or examples are extremely small. In these cases, the k-means algorithm will use a heuristic approach with the aim of finding locally optimal solutions to the problem. In other words, the algorithm will start to guess how the clusters can be assigned and then it will slightly adjust how the clusters are assigned to assess if the alterations will improve the cluster homogeneity. Therefore, the algorithm will set an initial “k” cluster and then updating the boundaries of the clusters based upon the datapoints that are within the cluster. These minor alterations occur many times until the cluster assignment can no longer be improved, resulting in the process stopping and the clusters reaching their optimal assignment. The final result of the clustering will change when slight adjustments are made to the initial conditions of the clustering due to the heuristic nature of the k-means algorithm. However, the final result should not change drastically otherwise this is an indication that there is an underlying issue. An issue that could arise is that the “k” value has been selected badly or the case could be that the data does not group naturally, which implies a weak dataset. To avoid such major issues, it is recommended that clustering is applied several times so see how robust the dataset is in relation to the results.

This type of algorithm is a relatively old approach, which still widely used. The main strength of k-means is performing well with many real-world use cases as the unsupervised approach, making it extremely desirable. K-means has a high level of flexibility and can be enhanced with slight tweaks, addressing a lot of the present limitations. Furthermore k-means uses simple principles, which makes it accessible for non-statistical data and users. However, the simplicity and unsupervised learning does come at the cost of not being able to guarantee the discovery of an optimal set of clusters. Subsequently, for effective clustering to occur, there should be a rough understanding of the number of natural clusters existent within the data. Finally, the k-means algorithm lacks the ability to deal with non-spherical clusters or clusters that have a wide range of density.

The initial step for the k-means algorithm starts by the allocation of "k" points within the feature space, which will act as the cluster centres. With the cluster centres defined, the rest of the population of the data can start falling into their clusters. The cluster centres are selected randomly within the training dataset, which means that it may not necessarily choose a point that is central within the cluster. As the k-means algorithm is quite sensitive to where the starting point is set to, assigning the centre clusters randomly could have a significant effect on the final set of clusters. This issue can be solved by implementing different methods of centre cluster assignment rather than allowing random selection, which will in turn improve the k-means algorithm. Alternatively, another option could be to completely skip the centre cluster selection by allowing each data point to be assigned to a cluster randomly, resulting in the algorithm jumping straight to the update phase and starting to tweak the clusters until it finds an optimal arrangement. Any approach that is taken will result in biases being introduced to the final arrangement of clusters, either improving or hindering the final outcome. Once the initial cluster centres have been selected, the remaining population of the data will be assigned to the nearest cluster centre based on a specific distance function. The k-means algorithm uses the Euclidean distance as a default, but other distances such as Minkowski or Manhattan distances are used occasionally.

It is evident that the k-means algorithm is very sensitive to how the centre clusters are assigned, to the extent that if a different combination is selected initially the clusters that are obtained in the final result would be completely different. Furthermore, the k-means algorithm is also sensitive to the number of clusters that have been instructed. If there is a small "k" number then the number of clusters will be small, which would risk losing valuable information or creating meaningless clusters. On the other hand, having a very large "k" number would enhance the homogeneity of the clusters but that will also increase the risk of the data being overfitted. Therefore, selecting the right balance as well as assigning the optimal starting point will have a significant effect on the final clusters. It is advisable to have some sort of prior knowledge regarding the selection of the number of clusters.

The k-means clustering algorithm is applied through numerous steps:

1. **Data collection:** As mentioned before, the data must be collected like any machine learning technique.
2. **Data exploration and preparation:** The data must be explored to see what it looks like and what kind of attributes it consists of. A common issue that must be dealt with like any other machine learning method is when there are missing values in the data. Usually, the way to solve this is to omit the missing datapoint but applying this method to every missing value could accumulate to a fair chunk of the data being omitted, resulting in a shrunken dataset. An alternative solution would be to include these datapoints but assign them with another label. An example would be if the data gathered consisted of gender; the datapoints that had a missing gender value could be assigned "unknown" so that the datapoint's other present values are still used. However, this approach is very subjective and is mainly for categorical variables. For certain datasets with non-categorical data such as gene expression it may not be possible to apply this solution, and omitting the missing values would be the best approach. Numeric missing values, such as age, can't be given an unknown categorical value as mentioned before but instead an imputation approach can be taken. This approach consists of filling the missing values of the data with an educated guess as to what the true value could be if the data collection was more effective [15]. In some cases, the simplest and most effective approach would be to take an average of existing values, which would allow the imputation to not impact the entire data, but the other values of the data can be used and therefore valuable information would not be lost.
3. **Model training on the data:** Once the missing issue has been dealt with, the data can now be used for model training. The data will be split into training and test dataset with a ratio set randomly by the researcher and randomly or other approaches that have been discussed before. The model will have access to the training data as well as what the clusters should be

like so that it can learn how to cluster additional data points that will be introduced.

4. **Model performance evaluation:** After the model has been trained and the researcher is satisfied with the level of clustering, the model can be introduced to the test dataset. However, the evaluation of the clustering results may be subjective as the failure or success of the k-means clustering model relies on what is the purpose for the clustering. Whether the clustering result answers the question or the aim that was set by the researcher and therefore, every evaluation will be different. Generally, the simplest way of evaluating the effectiveness of the clustering would be to examine the clusters and see what portion of the datapoints have been assigned to the correct cluster. However, this would mean that a priori knowledge about the data may be necessary.

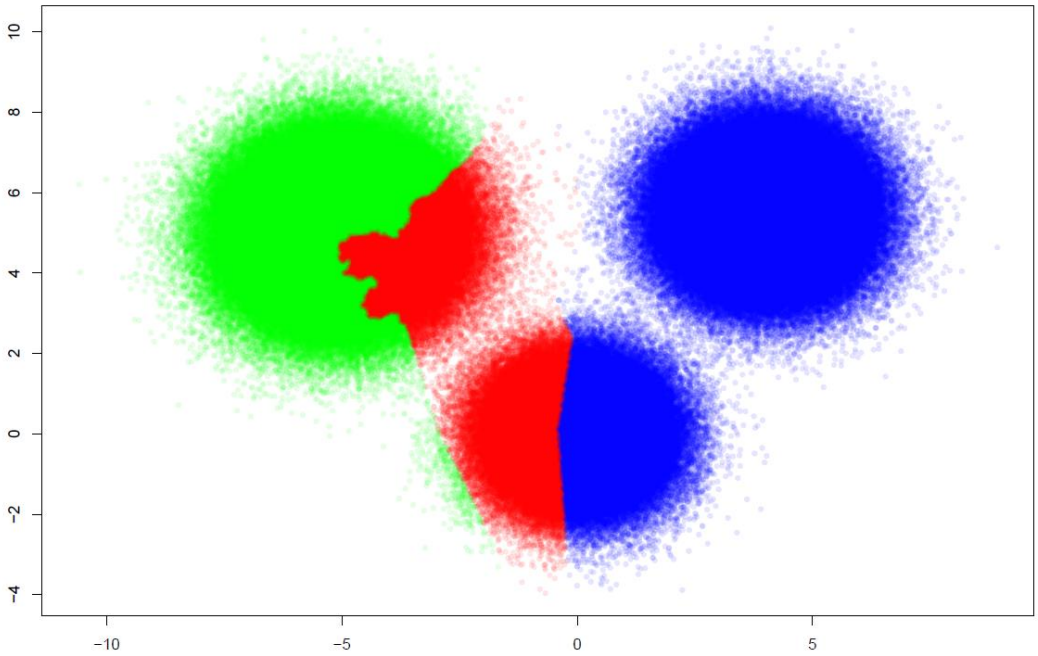


Figure 6: K-means clustering on prostate cancer sample data

Figure 6 shows how the k-means clustering algorithm performs unsupervised machine learning on prostate cancer sample data. From this result it is evident that the unsupervised

approach does not give very useful information as the aim of the machine learning was to cluster the benign and malignant patients in the dataset. However, further investigation can give us more insight on why the data has been clustered in this way, but the aim of the analysis was not satisfied.

5. **Model performance improvement:** From the evaluation process we can see that after the clustering process, new information is produced by the model. How well the model has performed relies on the quality of the clustering and what is done with the information that is given. If the information is meaningful and gives the researcher the required solution then improvement may not be intensive or even needed. In this case, the attention can be placed on taking this new information and creating actions accordingly. However, if the information is not very insightful or accurate, then the parameters of the clustering should be tweaked or altered, and the training process should be repeated.
6. **Final model testing:** If the model is altered to improve the performance, once the model reaches a satisfactory level determined by the researcher then the model can be used to solve real-world problems.

The k-means clustering algorithm is a very sturdy machine learning approach, which has been used as the foundation for further sophisticated models. Other variants of the algorithm introduce unique biases and heuristics that can improve the performance based on the desired requirements.

2.4.2. Hierarchical

Hierarchical clustering is one of the most popular and easy to understand methods used, where the data is presented in the form of a hierarchy over an entity that is set. Some features can be placed in the same hierarchy or in a separate hierarchy depending on the needs of the researcher. The approach of hierarchical clustering can be divided into agglomerative clustering and divisive clustering. Agglomerative clustering is a technique where a hierarchy is created in a bottom-up fashion, starting with each data point as a

separate cluster and sequentially merging them into similar clusters until a final cluster is formed. Divisive clustering works in an opposite way by creating a hierarchy in a top-to-bottom fashion, considering all the data points as one big cluster and after each iteration splitting the data points that are not similar until a certain number of dissimilar clusters are formed. It is simply evident that the agglomerative clustering approach would take a lot more computational power especially if there is a high dimensional data being used. However, this approach will result in a more accurate clustering result and parameters of similarity can be set so that the model knows when to stop the iterations [47].

Agglomerative clustering is a relatively simple and basic approach. Firstly, the proximity matrix is computed, then each datapoint is allowed to be recognised as a cluster. The main section of this clustering approach is merging the nearest 2 clusters based on a specific distance function. This step is repeated, and the proximity matrix is updated accordingly. Once a single cluster is formed the process finishes and the result can be evaluated and analysed. The result of the hierarchical clustering technique can be visualised in the form of a dendrogram, which is a tree like diagram that shows how and where each cluster splits until reaching the final single data points.

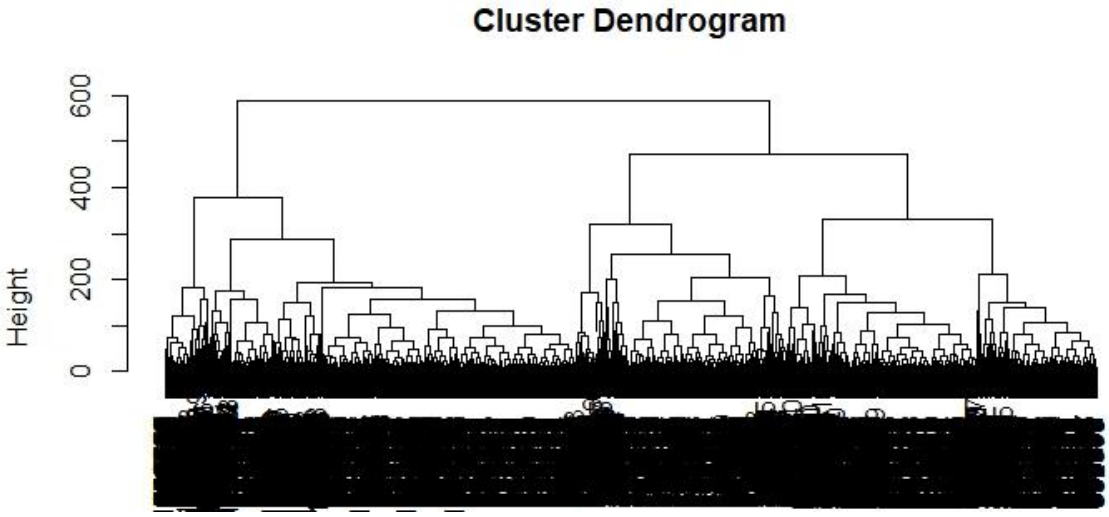


Figure 7: Hierarchical clustering of the sample prostate cancer data

Figure 7 shows the outcome of hierarchical clustering analysis on the sample prostate cancer. The blurred black section on the bottom of the dendrogram is the label for each

data point. It is evident from the dendrogram that the entire data splits into 2 clear main clusters, which is what was expected as the cancer data contains 2 groups of benign and malignant patients. Additionally, the right-hand side main cluster divides into 2 more clear clusters. This gives us extra insight and motive to further investigate the underlying reason for the cluster being split in that way. In some cases, other types of analysis may need to be carried out to enable the extraction of this information from the dendrogram.

The most important part of hierarchical clustering is calculating the similarity between 2 clusters and at what threshold should they be merged. There are a few approaches to calculate this:

- **Single linkage algorithm** (MIN), is simply when the two closest data points are selected such that one data point lies in cluster 1 and the other data point lies in cluster 2, taking their similarity and declaring it as the similarity between the two clusters. This approach is very simple and can be used to split non-elliptical shapes but the gap between the 2 clusters has to be large. However, this does mean that this approach cannot separate clusters if there is noise within the data or at the cluster boundaries.
- **Complete linkage algorithm** (MAX), works in the complete opposite way to the single linkage algorithm. The algorithm picks the two farthest data points such that one point lies in cluster 1 and the other point lies in cluster 2, taking their similarity and declaring it as the similarity between two clusters. This approach copes well with separating clusters within a dataset that has noisy cluster boundaries. However, the approach usually breaks large clusters into smaller ones as it could falsely detect noise.
- **Group average** takes every single pair of points, and it figures out their similarities and finally calculates the average of all the similarities. Similar to the complete linkage algorithm, the group average approach can cope well with separating clusters that have noisy boundaries. However, this approach is very biased towards data that contains globular clusters, rendering it not be useful for certain datasets.

- **Centroid distance** assigns centroid data points within 2 clusters and then takes the similarity between 2 centroid data points as the similarity between the entire cluster.
- **Ward's method** calculates similarities between 2 clusters like the group average approach, but the Ward's method calculates the sum of the square of the distances. Ward's method copes very well with noisy cluster boundaries but as like the group average approach it is very biased towards globular clusters.

There is no perfect approach so therefore the researcher must select an approach that will be the most effective for the dataset that needs to be analysed.

2.4.3. Gaussian Mixture

Gaussian Mixture clustering is another popular clustering algorithm for high dimensional data used in the biomedical domain. This type of clustering utilises a mixture model to represent the probability distribution of observations and subsequently, cluster the dataset. This is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. It can be seen as a way to generalise k-means clustering so that it can include information about the structure covariance of the data along with the centres of the latent Gaussians [48].

Each component in a Gaussian Mixture is modelled by a multivariate normal distribution. The parameters of component k include the mean vector μ_k and the covariance matrix Σ_k, which gives the probability density function of:

$$f_k(A_i|\mu_k, \Sigma_k) = \frac{exp\left\{-\frac{1}{2}(A_i - \mu_k^T)\Sigma_k^{-1}(A_i^T - \mu_k)\right\}}{|2\pi\Sigma_k|^{1/2}}$$

K is defined as the number of components in the mixture and τ_ks is the mixing proportions [49]:

$$0 < \tau_k < 1, \quad \Sigma_k \tau_k = 1$$

Gaussian Mixture clustering provides a big advantage of not assuming that clusters take up a particular sort of geometry, which means that it works well with non-linear

datasets and non-linear geometric distributions. Subsequently, it will not create cluster size bias, which can be evident in k-means. However, this technique uses all the components that it has access to, hence, it could lead to data high dimensionality when it comes to creating clusters. Consequently, the model can be over complicated and prove difficult to interpret.

2.4.4. Fuzzy C-means

Fuzzy C-means clustering is also a data clustering technique popular in the biomedical domain. This clustering method groups the dataset into N clusters, where every data point or observation in the dataset belongs to every cluster to a certain degree. This allows one piece of data to belong to two or more clusters and is frequently used in pattern recognition. Fuzzy C-means uses fuzzy logic principles to cluster multidimensional data by assigning each point a membership in each cluster centre or centroid from 0 to 100%, which is more powerful compared to the traditional hard-threshold clustering. The algorithm will assign membership to each data point corresponding to each centroid based on the distance between them. The closer the data point is to the centroid, the higher the membership value towards that centroid, resulting in the sum of all memberships of each point equalling to 1. Fuzzy C-means clustering is an unsupervised learning method that allows a fuzzy partition to build from the data. The algorithm relies on the value '*m*', corresponding to the degree of fuzziness of the solution. The larger the '*m*' value, the more blurred the classes will be, which can make all the elements to belong to all clusters. Hence, the effective selection of the parameter '*m*' is important as it will lead to different partitions of the data. Fuzzy C-means can be used for classifying of oral cancer cell data [50] and a modified version of the algorithm can be used to estimate field bias and to segment magnetic resonance imaging data [51].

The researcher assumes and assigns a fixed number of clusters, *c*, and then the fuzziness exponent, *m*, and termination tolerance, ε, are specified before the model can be constructed. The algorithm follows the following steps:

1. Randomly select cluster centre.

2. Initialise $U = [u_{ij}]$ matrix, $U^{(0)}$. u_{ij} can be calculated using:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

3. At k-step: calculate the centres vectors $C^{(k)} = [c_j]$ with $U^{(k)}$.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

4. Update $U^{(k)}$, $U^{(k+1)}$.

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

5. If $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$ or the minimum *J* is achieved, then STOP; otherwise return to step 3.

Clustering in this way gives very good results for datasets that are overlapping, especially when compared with k-means. A data point is assigned membership to each cluster centre and therefore each data point may belong to more than one centroid, which is why it performs better than k-means for overlapping data. However, choosing the number of clusters is a priori approach, which can cause a lot of trial and error. Subsequently, the performance of the algorithm depends on the selection of the initial cluster centroid and the initial membership value [52], [53].

2.4.5. Density-Based Spatial Clustering of Applications with Noise

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a well-known data clustering algorithm and very commonly used in data mining and machine learning. Being a density-based clustering, a set of points given in space are grouped together based on their distal proximity with many neighbouring points based on a distance metric such as Euclidean distance. With outliers, this method clusters together points in low-density regions with their nearest neighbours being far from each other.

Two parameters are required to build this model, "ε" specifying how close two points need to be to one another to be considered as a part of the same cluster. If the distance metric between the two points is lower or equal to the ε value, they will be considered as neighbours. The other parameter is "*minPts*" which is the number of points that is required to form a dense region. When choosing these parameters, it is necessary to understand how they will be used and some prior knowledge about the dataset will be helpful. Using too small of an ε value will cause a large part of the data to not be clustered. This is due to the model considering them as outliers as they don't satisfy the number of points needed to create a dense region. However, choosing too large of an ε value will cause clusters to merge, resulting in most of the dataset's objects being clustered together, which will hinder the effectiveness of the model. Selecting an optimal ε value based on what the desired outcome of the model should be, will increase the model's robustness, which can be done based on the overall distances of the dataset or by using a k-distance graph. The "*minPts*" parameter can be derived from a number of dimensions *D* in the dataset, where $minPts \geq D + 1$. Using larger values for this parameter is usually better for noisy datasets as it will form more informative clusters. The minimum value for the *minPts* should be 3, but the larger the dataset, the larger the *minPts* value should be [54]. One of the biggest advantages of DBSCAN is that it does not require a predetermination of the number of clusters in the dataset as opposed to other methods such as k-means. DBSCAN is very effective at finding arbitrarily shaped clusters, where a cluster can be completely surrounded by a different cluster but have no connection to it. DBSCAN has a notion of noise, is very robust to handling outliers and only needs two parameters - which can easily be set by the user that has prior knowledge about the dataset. It is mostly insensitive to the ordering of the points in the dataset. Conversely, DBSCAN does not handle highly dimensional data well, as the quality of the model is dependent on the distance metrics, which is useless due to the "curse of dimensionality" making it very difficult in selecting an effective ε value. Due to inappropriate selection of *minPts*-ε combination, DBSCAN cannot cluster datasets with large differences in densities successfully [55]. Finally, if

there is poor or no prior knowledge about the structure and scale of the dataset, choosing a meaningful ε value becomes difficult.

2.5. Topological Data Analysis

Topological Data Analysis (TDA) is an effective method at inferring networks within a data set and connecting data points that show similarities in terms of an observable proxy such as biomarkers. TDA is a relatively novel but increasingly popular approach to data analysis. This approach attempts to analyse datasets by studying the shape of the data to either reduce the dimensionality or better understand the underlying structure. This type of analysis can allow clinicians to gain a better understanding of potential sub-groups of patients for personalising interventions. TDA can be used to map complex clinical data sets as a non-dimensional network graph by clustering similar data points [61]–[63]. Topology is generally a mathematical and statistical method for inferring and analysing topological and geometric shapes. Singh, et al introduced the Mapper algorithm as a geometrical tool to analyse and visualise the topology of datasets [64] in the form of graphical structures. These mappers are used to discover the shape characteristics in the data set by means of specific filter functions [65], [66] and partial clustering of the dataset.

One benefit of this method is the independence from a specific clustering technique or algorithm, giving freedom and flexibility to the user to implement any appropriate clustering technique that best suits the data under analysis. The TDA Mapper has been used in many different studies such as the detection of topics in twitter [67], text representation for natural language processing and mining [68]–[70] and various clinical data such as breast cancer patients [65], spinal cord and brain injury [62], protein interaction networks [71], RNA-sequencing analysis [72] and analysing high and low functioning neuro-phenotypes within fragile X syndrome [73].

Firstly, the points in the data set are characterised with a similarity metric, which will measure the distance between the data points in the space. Secondly, a filter function must be defined on the data in that space to describe the data distribution. Thirdly, several overlapping bins should be defined by the resolution and the amount these bins can be

overlapped. Fourthly, the data points within each bin are clustered and one node (vertex) is generated for each cluster to define the geometric scale of the shape based on the number of clusters in each bin. Finally, for each pair of clusters, an edge is connected between the vertices to generate a mapping graph. The graph nodes and edges can then be coloured based on a specific feature of interest to allow the visualisation of the data set. Figure 8 is an example of TDA showing the patient trajectories [74].

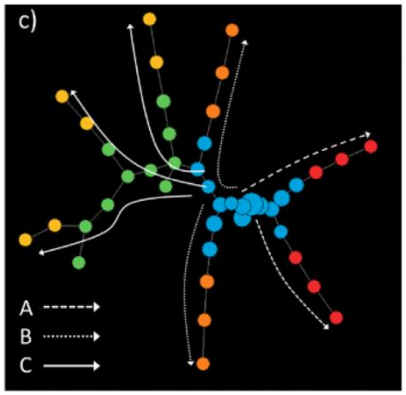


Figure 8: Minimum Spanning Tree identifying trajectories of patients with coloured nodes based on clustering membership.

2.5.1. Principal Component Analysis

As the field of data science is developing, larger datasets are becoming increasingly widespread, which are complex and highly dimensional. Interpreting these datasets require considerable computational effort and capabilities, requiring resources and increasingly more time consuming. To handle these datasets, their dimensionality needs to be significantly reduced to make it more manageable, whilst preserving most of the information in the dataset. This process is important to allow TDA to be applied effectively. Various dimensionality-reduction methods have been developed but Principal Component Analysis (PCA) is one of the most established and commonly used techniques. The idea is to transform the large set of variables into a smaller one that still preserve as much variability of the data as possible. To do this, new variables need to be found which are linear functions of the original dataset. These linear functions need to successfully maximise the data's variance and must not be correlated to each other [75].

PCA can be broken down into five steps. The first step is standardisation, which aims to standardise the range of the continuous initial variables so that each one of them contributes equally to the analysis. This initial standardisation step is crucial because if this step is omitted, there will be large differences between the ranges of the initial variables. Hence, the variables with larger ranges will dominate over those with smaller ranges, which will lead to biased results. Transforming the data to comparable scales can prevent this issue. The following mathematical formula can be used to standardise the variables to the same scale:

$$StandardisedValue = \frac{OriginalValue - Mean}{StandardDeviation}$$

The second step is covariance matrix computation, which aims to understand the relationship between how the input variables are varying from the mean with respect to each other. The output matrix displays the covariances of each pair of variables, which shows the distribution direction and magnitude of the multivariate data in the multidimensional space. How the data is distributed in two dimensions can be determined by controlling these values. If the output covariance matrix is positive, then the two variables are correlated and if it is negative, then they are inversely correlated.

The third step aims to compute the eigenvectors and eigenvalues of the covariance matrix so that the principal components can be identified. Eigenvalues represent the magnitude of the distribution of the two variables and eigenvectors show the direction. Both values help determine the principal components of the data. Principal components are new variables that are constructed as linear combination of the initial variables. These principal components are constructed so that they are uncorrelated and most of the information from the initial variables are compressed into the first component. Subsequently, the maximum remaining information is compressed into the second component and then this process continues until the information is spread into all principal components. The number of principal components is determined by the number of dimensions in the data. The first principal component accounts for the largest possible variance in the dataset and can be determined by ranking the eigenvectors in order of their

eigenvalues from highest to lowest, which gives the order of significance of the principal components.

The fourth step is to choose which principal components to use and which ones to discard due to their lower eigenvalues (lesser significance to the dataset). Establishing this will allow a Feature Vector to be constructed, which is a key step to dimensionality reduction as it shows the eigenvectors of the components that are being preserved. The decision of which components to keep or discard lays with the user and depends on the aim of the analysis. Generally, the first two principal components are selected as this provides an optimal dimensionality to variance trade-off. Finally, the data needs to be recast along the principal components axes with the use of the feature vector. This can be done by using the following formula:

$$FinalDataSet = FeatureVector^T \times StandardisedOriginalDataset^T$$

Many adaptations of PCA have been proposed and used for a variety of datasets from simple binary data to other complex time series data, expressing the significant benefits of this technique for dimensionality reduction [75].

2.5.2. Multidimensional Scaling

Multidimensional scaling (MDS) is a statistical tool that aims to represent distances or dissimilarities between sets of objects visually. These objects can be a variety of variables within the dataset, where the user wants to discover the hidden structure of the dataset, which is an important step for TDA [76]. Objects that have shorter distances between them or have more similarities are closer on graphs compared to objects that are further away from each other or are less similar. MDS can be used as a tool for dimensionality reduction as well as interpreting dissimilarities as distances on a graph [77]. MDS consists of four basic steps: firstly, a number of points need to be assigned to coordinates in n-dimensional space. Secondly, a mathematical distance such as Euclidean distance is used to calculate the distances for all pairs of points, which will result in a similarity matrix. Thirdly, the similarity matrix must be compared to the original input matrix by evaluating the stress function. Stress is a measure of goodness-of-fit, which is

based on the differences between the actual and predicted distances. Finally, the coordinates can be adjusted if it is necessary to minimise the stress. This technique of dimensionality reduction is similar to PCA but uses a similarity matrix to plot the graph rather than the original data [78].

2.5.3. Network Representation Learning

Network representation learning (NRL) aims to represent vertices within a network in a low-dimensional dense representations, where similar vertices in the networks will be linked through their distances such as Euclidean or cosine distances and labelled as close representations [79]. These representations can be used as features of the vertices such as an adjacency matrix and may be applied to a variety of networks. NRL is effectively a technique to turn network information into low-dimensional dense real-valued vectors and this can be used as the input for other machine learning algorithms [80].

2.5.4. Visualisation

Data visualisation is not only the most important step in TDA, but it is an essential tool for most data analysis. It allows detecting or identifying complex structures and patterns within the dataset via visual elements such as charts, graph, and maps, which may not be revealed in any other way [81], [82]. In the area of Big Data, a variety of visualisation techniques can be vital in enabling the analysis of an immense amount of information and provide the user with tools to make data driven decisions. Using data visualisation on unclassified data can discover hidden trends, outliers and how the data clusters naturally. On the other hand, using classified data with visualisation tools can offer insight on how the data can be separated into different classes and the structure of the classes [81]. Visualising multidimensional data can be classified into five different groups: geometric projection, graph based, hierarchical, icon-based, and pixel-oriented techniques [83]. The benefits of effective data visualisation are obvious as identifying trends and outliers is faster and simpler when the data is displayed visually rather than in

tables of numbers. Data visualisation can be very helpful in pre-processing and data cleaning steps due to the ability to find incorrect, corrupt, or missing values easier.

2.6. Pseudo-Time Series Analysis

Pseudo-time series (PTS) analysis is a successful method to create realistic trajectories through non-time-series data based upon distance metrics and external knowledge on staging within a process (such as disease state). These maps can be used to measure the disease progression between the patients in the data set [84]. When time series data is not available, this will enable the temporal behaviour of measured features to be understood. Forming a relative ordering of the patients in the data set, will allow a series of disease states to be determined, which can be used to map trajectories through disease progression.

To construct a PTS analysis, a distance matrix will be created between all datapoints similar to one created for topological analysis. Several points within the data set will be sampled via a bootstrapping method and a weighted graph can be generated based on a specific distance parameter [85]. This graph is then used to build multiple shortest paths from pre-labelled start points to pre-labelled end points using a combination of resampling and graph theory [86].

2.7. Cross-sectional and Longitudinal Studies

Medical machine learning explores data through cross-sectional and longitudinal studies. Cross-sectional studies, efficient and rapid, offer snapshots of disease processes but lack temporal insights. Longitudinal studies track subjects over time, providing detailed temporal data but at increased cost and time. This subsection delves into the strengths and limitations of each, highlighting the potential for a combined approach to create a more effective study design in medical machine learning research.

Data and experiments can be studied and explored in many ways, with cross-sectional and longitudinal studies being the main categories. Cross-sectional study is a type of observational study, where information is recorded from a sample or entire population of

subjects but by not altering the exposure status or environment [87]. The recorded information could be clinical test results, demographic attributes, or an association of variables. During this study, past or future behaviour is omitted, which means all measurements are made at a single point in time for a particular subject [88] [89]. In the field of medical machine learning, this allows the user to see a snapshot of the disease processes for many patients. Cross-sectional studies enable the creation of subgroups such as BMI or gender for certain diseases especially hereditary diseases. Subsequently, cross-sectional studies have the benefit of being conducted in a fast and inexpensive manner compared to longitudinal studies [88]–[90]. Furthermore, cross-sectional studies can be used for public health planning, monitoring and evaluation. This means that it can be used to monitor the prevalence of certain diseases and could drive future protocols [87]. As mentioned previously, cross-sectional studies take snapshot measures of variables in a single point in time, which evidently makes it difficult to derive casual relationships. Additionally, due to the lack of any measurement of progression over time, the temporal nature of the disease is not captured, which will not allow vital temporal behaviour to be analysed. Consequently, patients that are showing early signs of disease onset will be omitted and not discovered leading to late or misdiagnosis and prognosis. Moreover, this type of study is also prone to certain biases, which may result in misinterpretation of the associations and the direction of associations [87].

Longitudinal studies are another type of observational study, but it is used to investigate disease progression of the subjects over time [91]. Longitudinal data is commonly collected during clinical trials, which will initially attempt to obtain a more precise estimate of the outcome so that it can be used to influence the treatment. Also, this data is used to monitor the clinical variables at a specific time and finally, to evaluate how the treatment is affecting the patients over time [92]. The recording of the clinical test results is usually conducted without altering the study environment. Monitoring the same subjects over long periods of time may be necessary to provide further insight into the cause-end-effect relationships [88]. Recording multiple tests will generate multivariate time-series data, which is very common for patients that are in a high-risk category of the disease and need to be regularly monitored before a diagnosis can be made

and interventions executed. These studies provide the user with the benefit of being able to distinguish slight variations in characteristics of the target subject [91]. Also, temporal details of how the disease is progressing beyond a specific moment in time can be captured [88]. On the other hand, this level of data capturing over periods of time results in a limitation to the study's cohort size due to increasing costs and time.

It is evident that both cross-sectional and longitudinal studies have their respective benefits and limitations but combining them could develop a robust method of study, which is an area where research is progressing [93]–[96]. A cross-sectional study design is effective at determining prevalence and variation in a wide population and it will provide findings without any reasoning or explanation [97]. Consequently, using a longitudinal study design will provide the missing cause-end-effect relationships, which is not present in cross-sectional studies, but only under limited samples. Therefore, combining the strengths and compromising on the limitations can provide a novel and more effective study design.

2.8. Model Performance Evaluation

Evaluating the machine learning algorithm to assess the strengths and weaknesses of the technique is the most important process. Various assessments can be conducted on the algorithm to see how it has coped with the trained data and this will give insight into how it will perform on future data. The different performance measures will be discussed in this section, but the most effective method depends on the researcher needs and the dataset that is being used.

2.8.1. Confusion Matrix

A confusion matrix is a table that classifies the predictions made by machine learning techniques according to whether they match the actual value. This method of performance evaluation has been discussed in the k-nearest neighbour classifier, where the final result can be given as sensitivity and specificity. Two more performance measures can be obtained from the confusion matrix, which are accuracy, and error rate. The accuracy,

also known as the success rate, is the sum of the total number of true positive predictions and the total number true negative predictions, divided by the total number of predictions. Below shows this calculation in an equation format.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The error rate is simply the opposite to the accuracy, and it is the sum of the total number of false positives and the total number of false negatives, divided by the total number of predictions. Subsequently, the error rate can be calculated by one minus the accuracy, which is show in the equation below.

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN} = 1 - accuracy$$

This type of performance measure is relatively simple to calculate and understand but provides valuable insight to the machine learning algorithm's prediction capabilities.

2.8.2. ROC Curves

The Receiver Operating Characteristic (ROC) curve is a measure to inspect the model's relationship between achieving true positives and avoiding false positives. This curve has been in use since World War II to evaluate a receiver's capabilities to determine between true signals and false alarms [15]. The ROC curve is determined by the percentage number of true positives on the y-axis, which is sensitivity value against the percentage amount of the false positives on the x-axis, which is one minus the specificity.

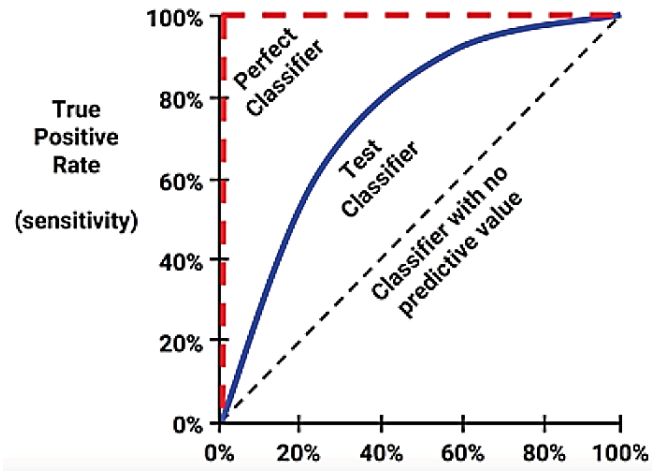


Figure 9: Labelled example of the ROC curve

The ROC curve points give the relationship between how the positive rate is against a varying false positive threshold. The classifier's predictions are arranged by the model's estimated probability of the positive class, which has the largest values sorted to create the curves. With every correct prediction, the curve will go vertically and for every incorrect prediction, the curve will go horizontally. The ROC curve will show a curve closer and closer to a perfect curve as the model becomes better and better at predicting positive values. The area under the ROC curve (AUC) can be used as a measure of how well the model is performing and predicting positive values. The AUC score can be interpreted with the values below.

- A: Outstanding = 0.9 – 1.0
- B: Excellent/good = 0.8 – 0.9
- C: Acceptable/fair = 0.7 – 0.8
- D: Poor = 0.6 – 0.7
- E: No discrimination 0.5 – 0.6

Nevertheless, these interpretations are an average similar to the kappa value and therefore the true values are subjective to the study and the researcher's needs. It must be stated that 2 ROC curves can have the same AUC, but one model could be effective and

the other could be poor. Therefore, the AUC alone cannot be used for evaluation, and it must be linked with its corresponding ROC curve.

Upon examination of the different methods of judging and assessing the performance of different machine learning models, it is evident that there is not a perfect performance measure and that each measure is subjective to the cause. However, majority of the performance measures provide numerical values as feedback on the machine learning technique, but this outcome may need extra computation to be able to use it effectively. Furthermore, from numerical values it is hard to understand how well the machine learning technique is performing, a visual representation of the performance will give a lot more meaning to a wider population. The ROC curve is a great example of how visualisations can help to improve understand and to see aspects that may not be easily evident with numerical results. Nonetheless using simple evaluation methods like confusion matrices will enable a quick assessment of the model performance and allow for improvements to be made to the algorithm. A combination of methods could provide insightful outcomes. Seeing the impact that visualisation has, it has prompted the motivation to research how the data can be analysed visually and to see if new techniques of visualisation can bring new understanding of machine learning, especially in the medical field.

2.8.3. Precision and Recall

Precision and recall are two performance measures that have a lot of similarity to sensitivity and specificity, which measures the compromises that have been made by the machine learning technique. This method of performance evaluation is used often with information retrieval, as this type of statistic gives the researcher insight on how relevant the results of the model are and if the results have been affected by any noise, rendering the results meaningless. Precision, which is also known as the positive predictive value, is simply how often the model makes a correct positive class prediction. Having a high precision value means that the model only predicts the positive class where there is a high chance of the value being positive, hence, making this model dependable. However, there

is a risk of the model becoming too precise and in turn losing its dependable reputation. Precision can be calculated the equation below.

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

The recall performance evaluation is the measure of how complete the results are. Calculating the recall value is the same as calling the sensitivity value, where the total number of true positives is divided by the sum of the total number of true positives and the total number of false negatives.

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

However, the interpretation of the recall value is different to the sensitivity value. A model that consists of a high recall value will have a wide breadth, which means it can gather a large amount of the positive example.

The precision and recall makes it is difficult to create a model that has both a high level of precision with a high level of recall, there must be a trade-off. Building a model with high precision is relatively simple as you aim for the examples that are easy to classify and building a model with a high level of recall is simple too by creating a very wide net, which makes the model excessively persistent in classifying the positive cases. Recall assesses the ability to capture all relevant instances among actual positives, emphasising the avoidance of false negatives. Sensitivity, on the other hand, focuses on the true positive rate, gauging the model's capability to identify positive cases. While both metrics aim to measure a model's performance in capturing positives, their nuanced emphasis leads to distinct interpretations. It is evident that creating a model with both high precision and recall simultaneously is a very problematic task. It is necessary to build and test several different models to ensure the ideal and effective ratio of precision and recall can be achieved based on the needs of the researcher and the dataset.

2.8.4. The kappa statistic

The kappa statistic is a performance measure where it will alter the accuracy by taking the possibility of the model making a correct prediction by chance alone into consideration. A dataset that has a major class imbalance would benefit from this approach as a classifier can achieve a high level of accuracy by simply predicting the most frequent class. The kappa value is calculated and presented with a value between one and zero, which describes how well the relationship is between the real classification and the model's prediction. The closer the kappa value is towards one, the better the predictions the model makes. A benchmark interpretation of the kappa value is as follows:

- Very good = 0.8 – 1.0
- Good = 0.6 – 0.8
- Moderate = 0.4 – 0.6
- Fair = 0.2 – 0.4
- Poor = 0.0 – 0.2

The above interpretations are just an average and the actual interpretation of the kappa value is subjective to the study and the needs of the researcher [98]. The kappa value is calculated with the following equation, where P(e) is the expected agreement between the classifier and the true value, P(a) is the proportion of the actual agreement.

$$kappa\ (\kappa) = \frac{P(a) - P(e)}{1 - P(e)}$$

Table 1 : Overview of the Machine Learning methods discussed.

Machine Learning Method	**Advantages**	**Disadvantages**
k-Nearest Neighbour	Simple, effective for small datasets	Sensitive to irrelevant features, computationally intensive for large datasets
Bayesian Network	Probabilistic framework, handles uncertainty	Requires prior knowledge, complex structure learning

Neural Network	Excellent for complex patterns, adaptable	Prone to overfitting, black-box nature
Decision Tree (Random Forest)	Interpretable, handles non-linearity	Overfitting, biased towards dominant classes
K-means	Simple, efficient for large datasets	Sensitive to initial centroids, assumes equal variance
Hierarchical	Captures complex relationships, visual representation	Computationally expensive, sensitive to noise
Gaussian Mixture	Accommodates different shapes, probabilistic outputs	Sensitive to initialisation, complex optimisation
Fuzzy C-means	Handles overlapping clusters, flexible membership	Sensitive to initialisation, computationally intensive
Density-Based Spatial Clustering	Robust to outliers, discovers arbitrary shapes	Sensitive to density parameter, difficulty handling varying densities
Topological Data Analysis	Captures global structures, handles noise	Computationally expensive, sensitive to scale
Pseudo-Time Series Analysis	Captures temporal dynamics, interpretable	Sensitivity to pseudo-time parameter, data preprocessing challenges

2.9. Summary

Previous and current research in the field of machine learning for biomedical data analysis has been reviewed in this chapter. A single method of analysis leaves limitations in different areas, which prompts the idea of combining methods to create a robust model. Cross-sectional studies have the disadvantage of not allowing for the temporal nature of the disease to be modelled as the time variable is not captured so it only shows a snapshot observation taken at a single fixed point in time. Consequently, longitudinal studies are expensive as each individual patient needs to be monitored and data must be collected

regularly over the entire time of disease or lifetime. This presents issues as many studies will only cover a relatively small window within the disease progression, not being able to capture the significant early or late stages. However, due to the lack of longitudinal data, analysing and building disease progression through cross-sectional data is being explored [86], [99], which is the area that will be explored in this thesis. Furthermore, using clustering techniques can help understand and identify important stages in disease progression, which can be used to model the missing temporal information.

The research within this thesis focuses on building topologies of the cross-sectional dataset, which would create networks and connect data points that show similarities in terms of biomarkers. Trajectories with multiple stages and endpoints can be built through these topologies using and extending the pseudo-time series technique. Additionally, the research will look to improve the robustness of combining and creating this novel analysis method by introducing prior knowledge to constraint the disease progression trajectories.

Chapter 3 Methodological Foundations for Novel Algorithm Development

3.1. Chapter Outline

The objective of this chapter is to express, define and test Topological Data Analysis and Pseudo-Time Series trajectories, which will be the foundation of the novel algorithm introduced in the next chapter. This chapter will discuss the underlying motivation of this research. This chapter is organised as follows: Section 3.2 provides an introduction, by discussing and describing the methodology of both TDA and PTS individually. Section 3.3 presents the diabetes patient data (MOSAIC data) that will be used for the initial testing of TDA and PTS. Section 3.4 describes the methodology used to firstly define the test parameters, then apply the methods. Subsequently, the results are presented accompanied by a clinical assessment. Section 3.5 provides a summary.

3.2. Introduction

In this crucial section of the thesis, we explore the complex domain of TDA and PTS, uncovering the significant insights they provide within the framework of datasets with high dimensions. With the ongoing expansion of the big data era, there is a growing urgency for the development of inventive approaches that can effectively extract significant patterns from intricate and multifaceted information. This chapter provides a thorough examination of the theoretical foundations and practical applications of both TDA and PTS. TDA, which is based on the principles of algebraic topology, offers a unique perspective for revealing the underlying structure of intricate datasets. The chapter provides a comprehensive explanation of the core principles of TDA, demonstrating its ability to capture the geometric properties and interrelationships among data points, surpassing conventional statistical approaches. Furthermore, the examination of the incorporation of TDA with datasets that possess a high number of dimensions is presented, demonstrating its ability to extract essential topological characteristics amongst the intricacy of the data. This investigation is enhanced by a comprehensive

analysis of PTS, which is a dynamic methodology for arranging observations in a data-centric manner based on temporal order. High-dimensional datasets frequently contain temporal elements, and PTS provide a sophisticated approach that allows for the identification of temporal trajectories without the need for explicit temporal annotations. The chapter provides a thorough examination of the rules that govern the generation of pseudo-time series. It highlights the usefulness of these concepts in effectively analysing the complex temporal dynamics found in a wide range of datasets. Importantly, these approaches are applied to highly dimensional datasets to illuminate their practical problems and complexities. Real-world examples and case studies show how TDA and PTS may have the potential to work together to provide deep insights into dimensionality issues, which is the main motivation behind this research.

3.2.1. Topological Data Analysis

Topological Data Analysis (TDA) works with large and complex datasets to allow the discovery of structural phenotypes. This is done by establishing networks to link individual or clusters of datapoints with similarities in biomarkers, clinical and demographical attributes. TDA utilises topology, which provides an analytical method to map complex, highly dimensional omics datasets. Exploiting the distances in the topology will allow qualitative information about the data to be extracted for analysis.

TDA is considered to have three fundamental properties, which are coordinate invariance, deformation invariance and compression. Topology is the study of shapes that doesn't use coordinates. In fact, topological constructions don't depend on which coordinate system is used. Instead, the shape is defined by the distance function. The topological characteristics of a geometric shape are unaffected by any stretching or deformation that may be applied to the shape. For example, considering a letter of the alphabet, the topological information can be gained even if it is deformed as long as the key features remain. Because of the topological nature of how our brains work, it is possible to recognise the letters in any font [100]. Hence, topologists look to TDA as a technique that performs better in noisy environments. TDA can also produce a topological

network with nodes and edges if the letter is treated as a massive data set consisting of millions of datapoints. This compact representation encodes all these connections in a straightforward fashion. These factors make this highly scalable approach a promising tool for analysing extremely large data sets found in the medical and omics field.

Joining relevant data points and constructing topological models as networks, TDA can capture the shape and structure of the data. This enables the visualisation of a "disease space," the understanding of the fundamental structure of the data, and the recognition of pertinent clusters as interconnected parts of the network [101]. TDA has the benefit of being able to construct a continuous shape on top of the data, which allows for the study of patients' conditions as a continuous spectrum, where patients can vary over the disease space, navigating between the network graph's nodes as their conditions change. TDA is dissimilar to cluster analysis as it accurately portrays continuous variation. This method of analysis utilises hierarchical clustering in the construction of its network graph, but it also adds extra precision to the groups that are produced in the process. Despite the fact that local behaviours may be lost or hidden, TDA removes the necessity for clustering approaches that separate items even if they belong together. This may be especially difficult for progressional, and inherently related data sets found in electronic health records. Instead, TDA executes clustering across overlapping areas of the data set, keeping the connections between the mining networks intact [65]. Using linear algebra and geometric parameters, TDA gives comprehensible representations of the derived findings. As such, it provides a straightforward solution to a pressing issue in artificial intelligence today: how to make findings from scientific studies more widely available through user-friendly applications that rely on visual displays of information and user-guided exploration of data [102]. By concentrating on the capture of data shapes, TDA makes it possible to model intricate data sets. TDA employs topology, a mathematical framework for measuring and expressing shapes, to represent complicated real-world and high-dimensional data sets as network graphs, making them easier to understand and analyse. Topological mappers are a class of mathematical tools that are used to figure out the structure of a dataset along predefined filter functions.

Topological data analysis starts by using a similarity metric to represent the points in the dataset by determining the degree of spatial proximity between them. Following this, the filter functions provide a statistical description of the data distribution in a coordinate space onto which the points will be projected. Overlapping bins are used to divide up the projections. Resolution is used to determine how many bins are generated within the projections' range of chosen lens values, and how much overlap there is between bins is defined by the gain. Each of these bins undergoes a clustering process. Subsequently, the number of clusters in each bin is used to determine the shape's geometric scale in this stage. Clusters are then used as the nodes in the network graph, with edges connecting them to other nodes in the graph that contain similar samples. After the graph has been constructed, nodes and edges can be given a colour, utilising a median value of filter functions, or by manually generating a function that fully captures the essence of the variables of interest. These colours can be based on average age of patients, density of points within each bin, different categories of the same disease etc. As a result of these characteristics, TDA is an appropriate tool for depicting temporal phenotypes and gives the ability to model disease progression. When it comes to the temporal requirements of a dynamical system, topology on its own is not sufficient. But topology, and in particular persistent homology, has been considered to handle time delay embedding models in contexts like risk analysis and the forecasting of critical transitions in financial markets [103]. Furthermore, TDA has also been suggested as a way to feature time series without making assumptions about their structure that depend on time [104].

3.2.2. Pseudo-Time Series

Ideally, the study of dynamic or progressive biological behaviour should take place inside a longitudinal framework. This type of framework enables the monitoring of individuals through time, which ultimately results in temporal data. Cohort sizes are typically small in longitudinal research due to logistical and financial constraints. On the other hand, research based on molecular 'omics are more likely to employ cross-sectional surveys of a group of subjects since they are simpler to implement at scale. Pseudo time

computational analysis may help cross-sectional research replicate some features of temporal variation that they miss while observing changes in patient disease characteristics.

A Pseudo-Time Series (PTS) is a collection of time-series measurements that attempts to simulate longitudinal data by constructing multiple trajectories from cross-sectional information. The goal of PTS is to reduce a set of high-dimensional molecular data from a cross-sectional cohort of individuals to a set of one-dimensional quantities called *pseudotimes*. By comparing the *pseudotimes* of different subjects, we can infer how they are progressing (temporal behaviour of measured features) through the biological process of interest such as disease progression, even in the absence of explicit time series data. In order to do this kind of study, it is necessary for members of the cross-sectional cohort to exhibit asynchronous behaviours and be located at distinct stages of development. Accordingly, we may construct a succession of molecular states that together make up a trajectory for the biological process of interest by establishing a progression ordering of the individual subjects [105] [85].

The process of creating PTS involves plotting multiple trajectories through cross-sectional data based upon distances between data points, with the help of prior knowledge of healthy and disease states. To make predictions, these trajectories can be input into approximative temporal models. Multivariate Time-Series (MTS) can be described as time-series data that includes not just one but multiple connected variables. Degenerative diseases, such as, cancer, glaucoma, and Parkinson's, can be predicted through MTS by studying the dynamic relationships between variables over time. Univariate time-series models don't always produce accurate predictions like MTS models do. However, it can be very difficult to learn reliable models that can exploit MTS data. A partial path can be fitted through the entire cross-sectional dataset in order to create an MTS model. This path is discovered using a PQ-tree approach in conjunction with a straightforward hill-climbing procedure, and it is intended to minimise some distance metric between the data, such as the Euclidean or cosine distance. Since there is a potentially huge number of possible orderings, the PQ-tree method is used to merely increase search efficiency. PQ-trees are a tool in graph theory that may be used to describe a partial ordering of points

and reveal which portions of the ordering are strongly supported, known as Q-nodes and which parts are more tentative, known as P-nodes. Children of P-nodes can be rearranged in any order, but children of Q-nodes can only have their order reversed. First, a distance matrix can be developed between all the variables, and then it can be used to construct a minimum spanning tree, which is the basis of how the data is ordered. The diameter path of the tree is utilised as the main Q-node of the PQ-tree, which serves as the skeletal structure for the reorganised ordering. The main Q-node is supplemented by adding the branches of the diameter paths as additional P-nodes and Q-nodes. To that end, the constructed PQ-tree can be seen as only a provisional partial ordering of the data points [86], [106]. Using a hill-climb approach, the PQ-tree's partial ordering can be transformed into a full ordering in order to further reduce the distance while staying within the PQ-tree's constraints. The search can be limited to paths between two predetermined points, one representing the starting health region and the end diseased region in the cross-sectional study [86]. Once this process is complete, a temporal model can be built for predictions.

This pseudo-MTS model is dependent on the size of the dataset along with the number of transition or diseased states it has. If a dataset includes many different disease states, trying to fit a single trajectory through it will produce an unrealistic or impossible trajectory. This will then result in creating poor temporal models. Additionally, inaccurate trajectories can also be due to poorly selecting the start and endpoints as the dataset can be highly dimensional. The method of bootstrapping [107] can be introduced, which is an attempt to deal with the natural variability that exists between the trajectories of disease within a population. This is accomplished by repeatedly resampling data from the dataset and fitting shortest paths between healthy and diseased regions. For the purpose of training a time-series model, each of these paths is used as an ordering to produce a unique pseudo time series. In this case, the Floyd-Warshall algorithm is used to determine the shortest path [108]. In contrast to traditional methods, the bootstrapping method can be used to construct time series models without the pitfalls associated with choosing suitable start and endpoints. Assuming the dataset has properly labelled classes representing

healthy and diseased regions, the algorithm should be able to construct realistic trajectories.

3.3. MOSAIC Data

For the initial testing of TDA and PTS, the dataset used was from the MOSAIC project co-funded by the Seventh Framework Programme of the European Union, which aims to improve the way Type 2 Diabetes Mellitus (T2DM) and its related complications are diagnosed and managed [109]. The collection of medical records from 924 pre-diagnosed T2DM patients yielded 13,623 instances for the dataset. Factors that have been shown to affect T2DM risk are body mass index (BMI), diastolic blood pressure (DBP), glycated haemoglobin (HbA1c), high-density lipoprotein (HDL), smoking habits, systolic blood pressure (SBP), total cholesterol and triglycerides [110]. The experimental results were mined for microvascular comorbidities and variables such as age, BMI, HbA1c, SBP, smoking habit, total cholesterol, and triglycerides are used to construct the topology and pseudo-time series. Firstly, all continuous variables within the dataset were normalised to $-1 \leq x \leq +1$ to avoid data anomalies. To evaluate whether the trajectories accurately model patients' development, the time since first visit was used since several of these patients had variable follow-up measurements. Kaplan Meier visualisation was used to compare the onset of microvascular comorbidities in subjects who belonged to the various trajectories that were discovered. Subsequently, the discovered groups of patients were investigated to see if they could be used as a predictor for the onset of microvascular complications. In order to estimate onset probabilities, a multivariate survival analysis using Cox-Regression was performed [101].

3.4. Experiments and Results

TDA is applied to the genomic cancer data, which consists of a merged dataset of lung, pancreatic and renal cancer patients, further described in chapter 4. The outcomes from this initial analysis will be explored to see the impacts of adjusting the parameters has on the topology. The results from TDA and PTS for the MOSAIC data is presented followed by a clinical evaluation of those results.

3.4.1. Initial Topological Data Analysis on Genomic Cancer Data

The patients in the genomic cancer dataset were categorised into classes of 0, 1, 2 and 3 as sample patients, lung cancer patients, pancreatic tumour patients and renal tumour respectively. TDA is applied using the entire dataset and each node is resized to represent the density followed by using colour enrichment to allow the results to be visualised more effectively. Figure 10 shows an initial plot with 30% overlap and 12 number of bins.

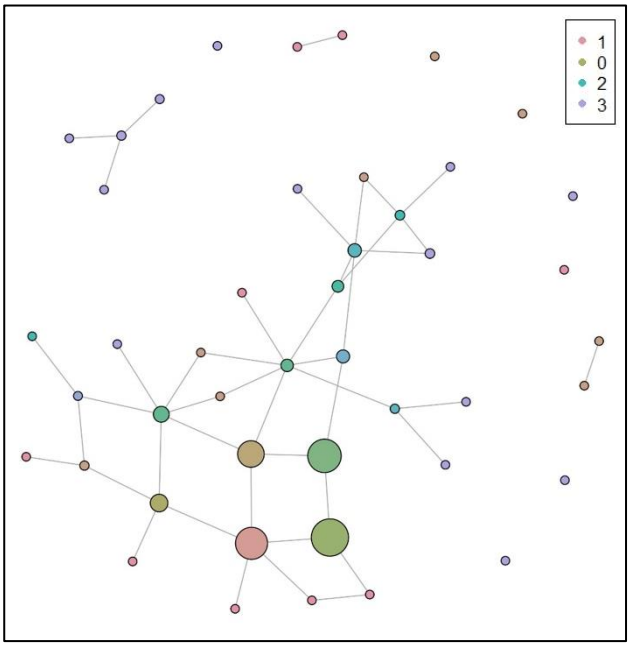


Figure 10: Topological Data Analysis plot with a 30% overlap and 12 number of bins

The clusters show the density within them, and a legend displays the colour allocation to each of the classes. The plot presents the data clustered very densely towards the bottom left with 4 main clusters that are linked via single trajectories. The largest cluster, bottom left of the 4 main clusters, seems to be allocated to the control group patients, which is an excellent starting point to investigate the disease progression to the 3 types of cancer patients. However, there are clusters on the top and right sections where they do not connect to the main cluster trajectories. This brings about 2 assumptions, either the clusters do not link with the main group of clusters, which means the data may not be as

meaningful as we had hoped, or it could mean that the parameters are too restrained, so it does not link all the true trajectories. For the first assumption to be justified, the clusters which are not linked through trajectories should be of one class type. Looking at these separate clusters further, it can be seen that they consist of a variety of the patient classes, which implies that that the first assumption is not valid. The next step that is taken would be to alter the topological parameters in order to find optimal values for analysis. Below is an example of the different TDA plots that were conducted.

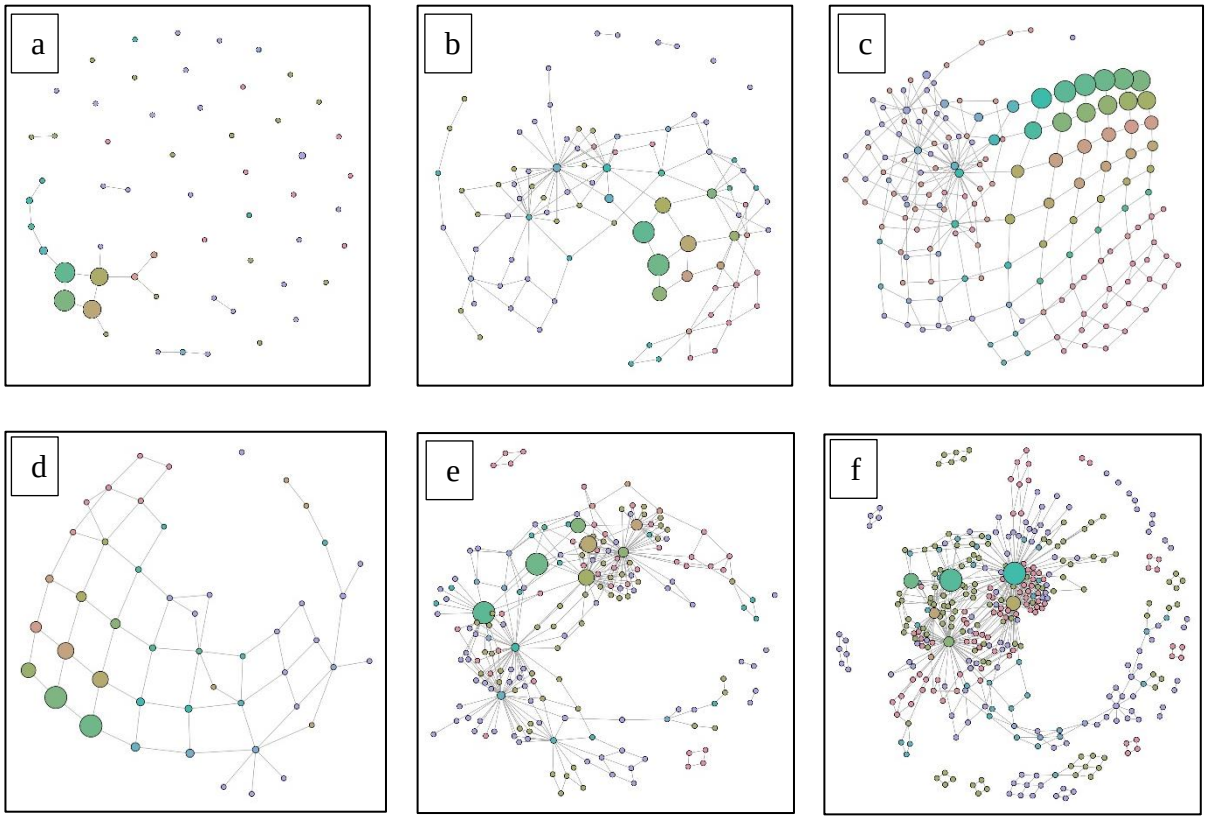


Figure 11: TDA plot with a) 10% overlap/12bins, b) 60% overlap/12 bins, c) 90% overlap/12 bins, d) 60% overlap/4 bins, e) 60% overlap/20 bins, f) 60% overlap/40 bins

It is evident from the sample of plots above, the effects of varying the parameters have on the trajectories and the clusters. Figure 11a is very meaningless as there are so few trajectories whereas Figure 11c creates too many clusters with equal densities, and it is

overly uniform in the top right section, which raises suspicions. Figure 11d is also displaying very uniformly arranged clusters and trajectories whereas Figure 11e and Figure 11f create an excessive number of forced clusters and trajectories. From initial observations, Figure 11b seems to be close to the optimal set of parameters and some minor adjustments and tweaks may clear the best TDA plot for the dataset. These initial results shows that highly dimensional data can be modelled through TDA and PTS and with the correct refinement, the progression pattern of the disease can be extracted.

3.4.2. Topological Data Analysis on MOSAIC Data

Topological Data Mapper was implemented for the analysis and the mapper2D function was used to conduct TDA [64]. The cosine distance along with single-value decomposition (SVD) was used for the parameterisation of the TDA. Consequently, L1-infinity centrality, which assigns the distance to the point farthest away to each point, was used as a filter function to build the topology [111]. The effect of adjusting the resolution parameter, such as the number of bins and their percentage overlap was investigated, as well as the geometric scale, which is the number of bins, while employing a grid search. It is essential to tune the parameters and adjust the scale in order to ensure that the shape details is sufficiently fine to detect temporal behaviours. This means the individual patients' repeated observations over time are not constrained to the same node. Trajectory discovery could be hampered by state changes within nodes if the granularity was too coarse [101]. Network analysis can be conducted on the topological graph resulting from the TDA algorithm. The cluster optimal function [112], which determines the optimal community structure of a graph by maximising the modularity measure across all possible partitions, was used to identify unique topology sections that enable us to retrieve subgroups of observations. A minimum spanning tree filter is used to find the shortest paths in a network topology, allowing for the identification of individual paths within the larger network. Temporal features were not employed to recover the initial topology, but they did play a role in shaping the minimum spanning tree used to depict the development of a disease over time.

The TDA algorithm's output is depicted graphically in Figure 12, Figure 13 and Figure 14. There is a temporal relationship between each node, which represents a collection of data points as observations. Each encounter's time since the first visit is represented by a different colour across the nodes. The continuous ordering of the colours can be seen in Figure 14a, which goes from blue, 0 days since first visit, to red, over 4000 days since first visit.

As mentioned before, the effect of changing the geometric scale, such as altering the number of clusters within each bin, has on the topology of the results, which can be seen in Figure 12. Higher values tend to make the network extremely sparse or loosely connected, while lower values typically produce very small clusters. Edges based on common samples are inaccessible in both cases, and the resulting shapes lack meaningful topological properties. Figure 12 shows that selecting between 8 to 12 clusters per bin maintains a relatively stable topology and hence, a value of 10 clusters per bin is selected for the analysis.

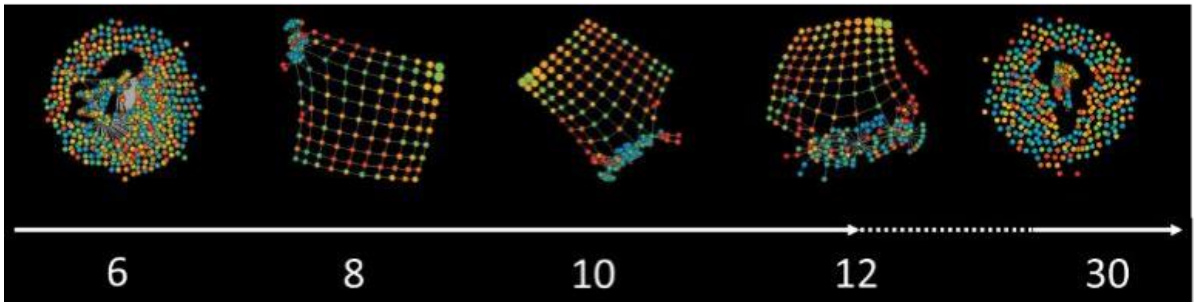


Figure 12: Plots showing the effects of altering the geometric scale

Additionally, the impact of the altering the resolution scale has on the topology is investigated by varying the percentage overlap or gain of the clusters. More edges can be expected when the gain is increased. By increasing the resolution of a graph, more bins can be created to display data better. The x-axis of Figure 13 displays the total number of overlapping intervals, and y-axis shows the percentage of overlap. It is important to note that the shape can be maintained for interval sizes between 6 to 14 regardless of the percentage, which has a negligible effect on the overall shape. At higher values, the

network loses its stability, making it harder to discern its basic structure and individual paths.

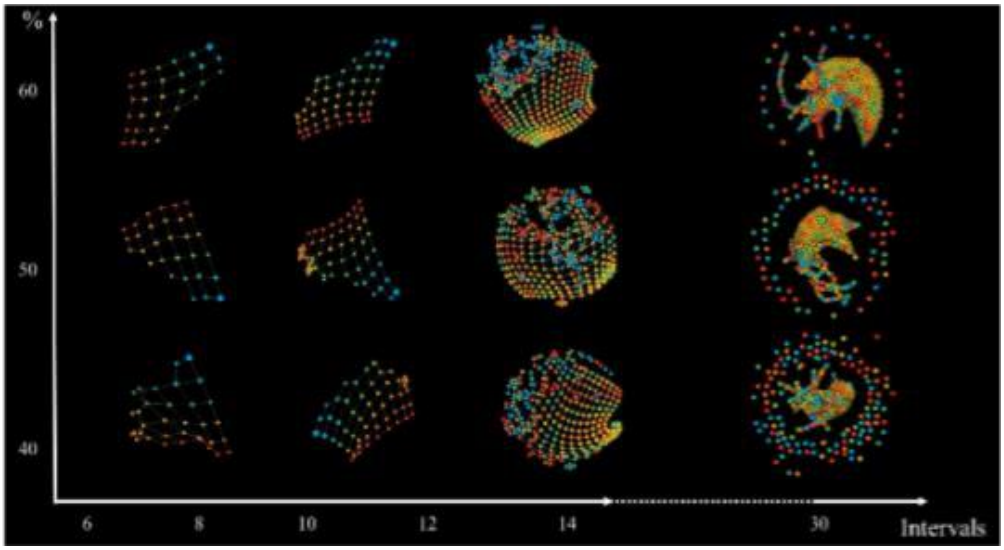


Figure 13: Plots showing the effects of altering the resolution scale and the percentage cluster overlap.

Figure 14 depicts a stable topology that was generated with 7 bins, 60% overlap, and an 8 point geometric scale, hence this particular topology will be utilised in the subsequent analysis steps. The distribution of the enriched topology with respect to the number of visits is shown in the bottom panel of Figure 14a. There is a discernible temporal direction from the blue bottom node towards the red nodes, suggesting that the temporal progression can be reconstructed by TDA. Moreover, the five clusters obtained by applying the optimal community structure cluster to the TDA results are reported in Figure 14b, which shows an enhanced topology.

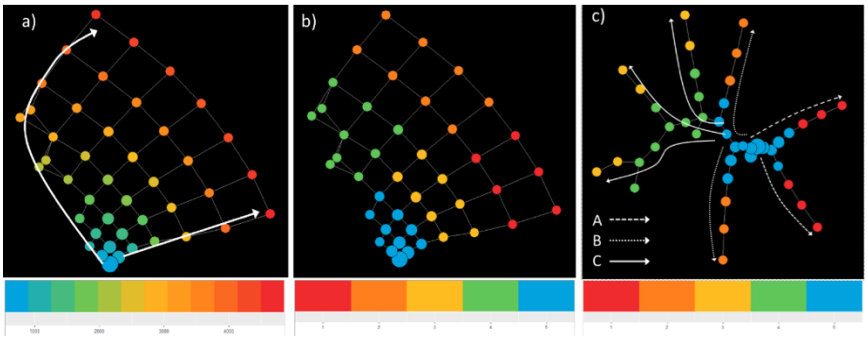


Figure 14: The network retrieved via TDA and displayed with igraph. In a) nodes are coloured by time from the first visit, in b) with the cluster membership. In c) The Minimum Spanning Tree identifies trajectories of patients. The node colouring is based upon the clustering membership.

Seven separate trajectories were found using the minimum spanning tree in Figure 14c. It shows all of them originating from the central blue cluster, which represents the earliest observations. The gathered trajectories are then classified by hand according to their final destination. Trajectory A shows two paths leading to the red clusters, trajectory B shows two paths leading to the orange clusters, and trajectory C shows three paths leading to the yellow clusters past the green clusters. These three categories are examples of temporal phenotypes, which are characterised by the progression of a disease over time.

3.4.3. Pseudo-Time Trajectories on MOSAIC Data

The graph that was utilised in the construction of the minimum spanning free in Figure 15b is displayed in Figure 15a, which illustrates the graph that was built on the basis of a cosine distance. One data point that was categorised as not having any microvascular complications was chosen at random to serve as the starting point, and one data point that was categorised as having at least one microvascular complication was chosen at random to serve as the end point. A single pseudo-time series was generated by finding the shortest path between the start and end nodes in the minimal spanning tree, shown in Figure 15c. For the purpose of creating numerous pseudo-time series, the resampling

process was performed a thousand times. Figure 16 shows a cosine distance map that has been enhanced with data about whether or not micro vascular complications occurred over the period of observation. A graph depicting the relationship between disease and complications has been constructed using the entire 10 samples of pseudo-time series. After generating 1000 pseudo-time series, a model was built using an Autoregressive Hidden Markov Model (ARHMM) with 5 discrete hidden states to capture the dynamics of the data's various trajectories. Experiments on multiple clinical datasets using PTS techniques led to the selection of the 5 underlying classes [113].

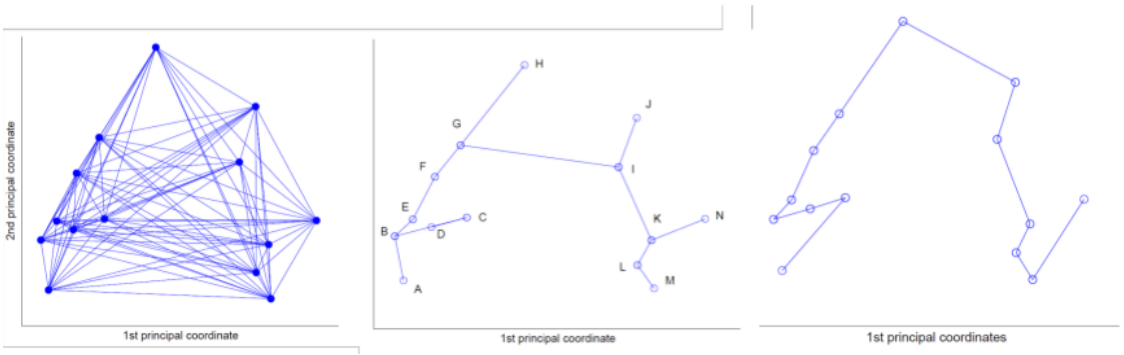


Figure 15: Plots showing the building pseudo-time series, a) the weighted graph of a sample of data (b) the mini- mum spanning tree of the weighted graph and (c) the Pseudo Time-Series

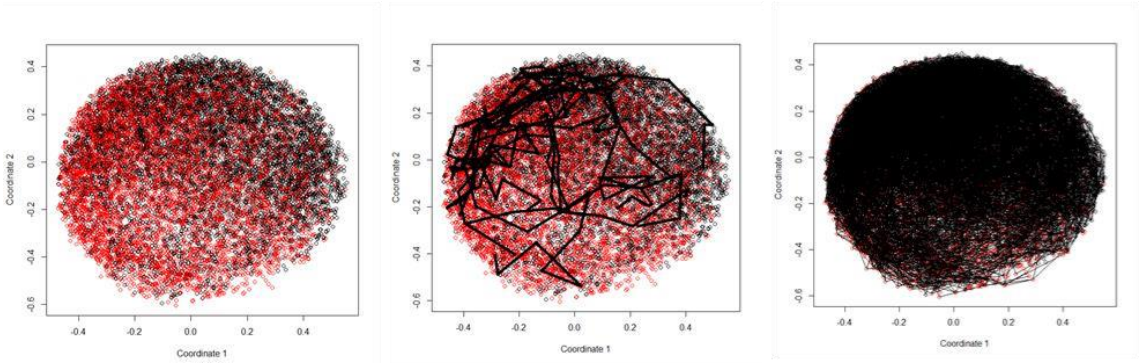


Figure 16: a) Multidimensional Scale plot of Cosine Distance where red represents patients with at least one microvascular complication, and black represents none, b) Cosine Plot with 10 sample Pseudo-time Series trajectories plotted, c) Full 1000 Pseudo-time Series Generated

3.4.4. Clinical Assessment

Patients with T2DM had their data used to construct a topological data network, with the most stable topology being chosen and then enriched with information about how long it had been since the patient's first visit. By doing this, we were able to see sub-groups of observations in the topology clustering shown in Figure 14b and possible disease progression trajectories shown in Figure 14a. After settling on the optimal topology, a minimum spanning tree was constructed from the graph in order to locate pseudo-time-based trajectories presented in Figure 14c. This method enabled us to pinpoint seven distinct trajectories. Three temporal phenotypes of T2DM (A, B, and C) have been identified, and their respective trajectories toward the rapid deterioration of the disease and their respective outcomes are shown in Figure 14c. Microvascular complications are more common in patients with the C phenotype (61.0%, n = 159) compared to those with the B phenotypes (43.0%, n = 574) and phenotypes A (23.0%, n = 191). Consequently, minimum spanning tree paths can detect groups of patients who are less exposed to the development of T2DM-related complications over time such as phenotype A or more exposed such as phenotypes B and C. In order to classify these phenotypes, we use values for important clinical features as they emerge over time, which is shown in Figure 17.

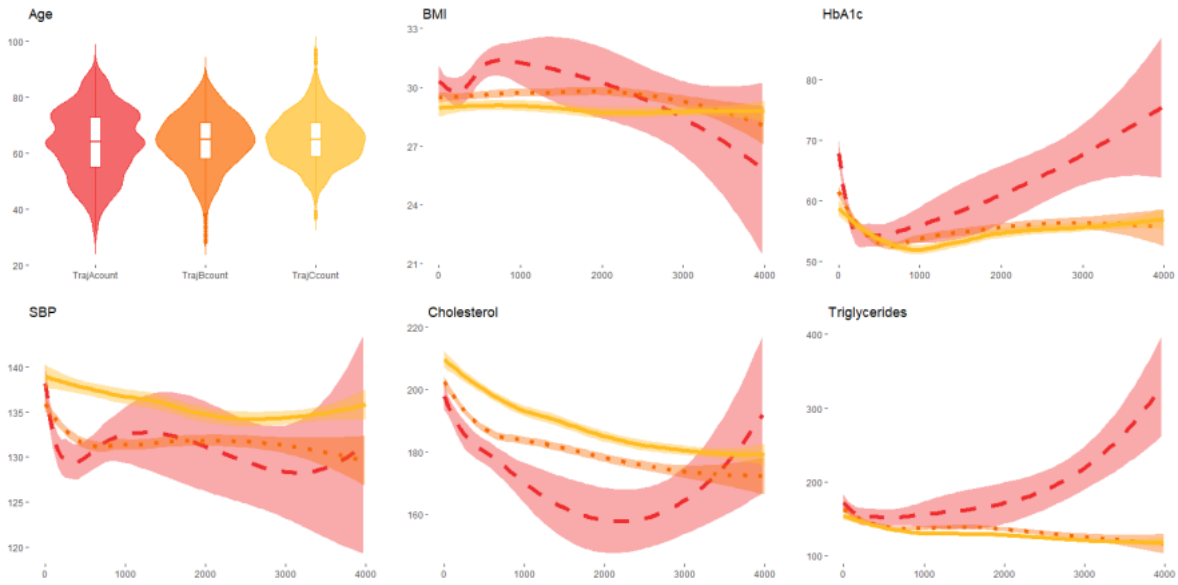


Figure 17: Clinical characteristics over time of subjects in the A (red-dashed), B (orange-dotted) and C trajectories (yellow-continuous)

Patients with the phenotype C had a higher baseline cholesterol and systolic blood pressure, and these trends persisted over time. HbA1c levels are higher and rising in people with the phenotype A, while cholesterol levels are falling and then rising, and triglyceride levels are going up.

The 1000 pseudo time-series we generated from the original data were then used to infer a five-state Auto Regressive Hidden Markov Model. The expected values for the primary characteristics of the data for each of the five secret states are shown in Table 2.

Table 2: Expected values for the 5 hidden states (t2d is time-since-first-visit, TotChol is total cholesterol and Trigl is triglycerides)

State	1	2	3	4	5
% Female	0	0	0	100	51
% Male	50	100	56	100	35
Age	59.16	**69.41**	63.7	**67.78**	56
t2d	3.77	9.76	**13.4**	**11.86**	5.42
HbA1c	47.66	50.3	**62.6**	53.54	60.7
BMI	28.1	27.58	30.07	30.31	**31.02**
SBP	129.59	129.5	**136.08**	134.8	132.73
TotChol	187.51	167.28	183.7	188.86	**207.62**
Trigl	126.98	108.13	136.71	124.38	**232.46**

The patients in State 1 denotes younger subjects that have the shortest time since their first visit, State 2 shows the oldest patients. State 3 displays subjects with the highest SBP and Hba1C values along with being patients that have been visiting for the longest of time since their first visit. State 4 represents older patients who have been visiting for a long period and finally, State 5 illustrates the youngest patients with the highest BMI.

Table 3: State Transition Matrix

	State 1	State 2	State 3	State 4	State 5
State 1	0.733	**0.145**	0.023	**0.084**	0.015
State 2	0.036	0.866	0.049	**0.049**	0
State 3	0.055	**0.127**	0.678	**0.132**	0.007
State 4	**0.159**	0.113	**0.157**	0.542	0.029
State 5	**0.134**	0	0.107	**0.140**	0.618

The probabilities of changing to and from these states are displayed in Table 3. Table 3's transition probabilities show that most states are quite stable, with higher odds of staying the same compared to transitioning to a new state. The two most likely changes in state are highlighted in bold. Figure 18 depicts this scenario graphically, showing the logical progression from State 5 to its two possible terminal states, States 3 and 4.

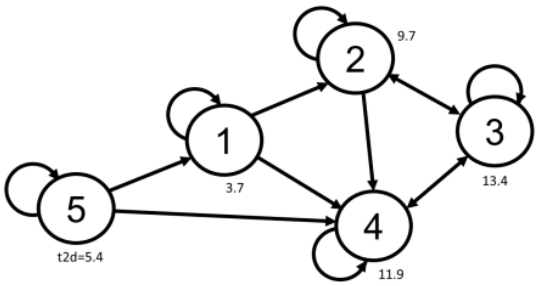


Figure 18: Transition Diagram with expected time since first visit

Both the length of time that has passed since the patient's initial visit and the patient's age contribute to this trend. Patients in end State 3 have extremely high HbA1c and low cholesterol, while those in end State 4 have higher cholesterol but lower HbA1c, very low triglycerides and are older. The HMM model provides two possible trajectories, depicted in Figure 19 as state transitions, which are 5-1-4 and 5-1-2-3, for triglycerides on the left and cholesterol on the right. It is intriguing that, like the TDA results in Figure 17, the lipid profiles were found to be a distinguishing feature of the two trajectories.

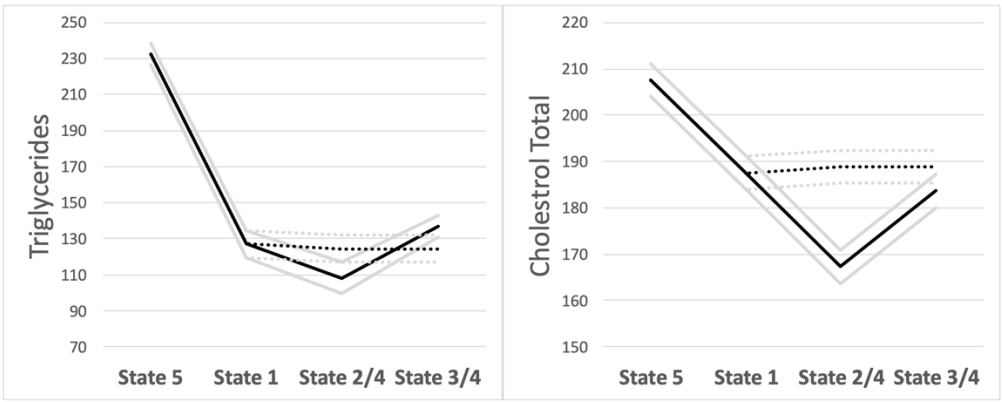


Figure 19: Triglycerides and Cholesterol mean statistics for two trajectories 5-1-4 (dashed) and 5-1-2-3 (solid).

3.5. Summary

In this chapter, we compared two methods for constructing temporal phenotypes automatically from medical records. The topology of the data has been captured using

TDA, and distinct trajectories have been singled out using a minimum spanning tree filter. It was found that one sub-cohort of people with T2DM has different cholesterol and initial HbA1c levels compared to the general population, and these differences were highlighted by this method. PTS techniques were also investigated; these allow for multiple trajectories to be bootstrapped from the data, and a state-space model with five hidden states to be learned. Despite the limited number of trajectories identified, the two discovered trajectories are clinically important and corroborate TDA's findings. Neither TDA nor PTS utilised temporal characteristics of the underlying data in the medical records during model construction. If the appropriate disease staging information is included, both methods could be used to construct temporal phenotypes from cross-sectional data. Throughout this chapter, we used micro-vascular comorbidity data, but any dataset that aids in disease staging would do. The comparison of temporal trajectories across patients and their use in predicting disease deviations or adverse outcomes is another crucial part of the analysis process [114]–[117]. One of the drawbacks of this PTS is that we did not investigate any ARHMMs with more than five hidden states. It is possible that larger datasets contain a greater number of hidden states, which could lead to the discovery of more complex trajectories as it will contain more data. However, this approach gives an incentive to combine the TDA and PTS methods to create a novel algorithm described in the next chapter.

Chapter 4 Defining the TDA-PTS Algorithm and Identifying Key Disease Progression Stages from Cross-Sectional Data

4.1. Chapter Outline

This objective of this chapter is to explore the combination of pseudo-time and topological data analysis to build realistic trajectories over disease topologies by defining the novel TDA-PTS algorithm. Initially, the novel algorithm will be described before applying it to three very different datasets: one simulated dataset, which consists of time series generated by an autoregressive HMM, one diabetes dataset and a combined genomic dataset from three cancer studies. We will explore how the combined method can highlight distinct temporal phenotypes in each disease based on the possible trajectories through the disease process. Additionally, the results (which has been published in publication number 3) will be evaluated to see if any insights can be obtained to assess the model. This chapter is organised as follows: Section 4.2 provides an introduction by briefly highlighting the motivation. Section 4.3 introduces and explains the novel TDA-PTS approach. Section 4.4 presents the 3 datasets that will be used for constructing and testing the novel algorithm as well as the how the experiments will be conducted. Section 4.5 illustrates the results from the analysis whilst explaining what effects these results may have on the study as well as the implications it could have on the actual disease prognosis, progression and treatment. Section 4.6 provides a summary.

4.2. Introduction

The subject of understanding the fundamental structure of clinical data is becoming significant. Topological data analysis facilitates the exploration of data by extracting the inherent topological structure, hence enabling the identification of unique regions. For instance, certain regions may be linked to the presence of disease in its first stages, whilst other places may be indicative of distinct subtypes of advanced disease. The discovery of

these regions can assist clinicians in gaining a deeper comprehension of the symptoms exhibited by individual patients, as these symptoms are influenced by their location within the disease topology. Consequently, doctors can implement more precise treatments. Nevertheless, these aforementioned topologies fail to encompass any sequential or temporal data. Pseudo-time series analysis is a methodology that utilises graph theory and expert knowledge, such as disease staging information, to build realistic trajectories from non-time-series data. This research aims to investigate the utilisation of pseudo time and topological data analysis in order to construct accurate trajectories across disease topologies. In this study, we investigate the identification of discrete temporal phenotypes in three diverse datasets: simulated data, diabetic data, and genomic data. Our approach involves utilising a combination strategy that examines the potential trajectories within each disease process.

4.3. TDA-PTS algorithm

The dataset, D, can be defined as a real valued matrix of m by n, where m (columns) is the number of samples that are the patients and n (rows) is the number of variables that are the clinical features in the data. These clinical features could be patient attributes such as age, gender, blood pressure or it can be certain genes used as a biomarker. D_i can be defined as the ith column of matrix D. Furthermore, $C = [c_1, c_2, \ldots, c_m]$ is used as a vector that represents class labels of the dataset, where $c_i \in \{0,1\}$ corresponds to the sample i. Subsequently, $c_i = 1$ and $c_i = 0$ represents the patients in the sample i that are the diseased cases and the healthy cases within the dataset respectively. The classes of the sample i have been determined based upon the diagnosis made by experts or clinicians. This is where the analysis can adopt a semi-supervised approach to enable a more robust investigation resulting in more meaningful results that could have clinical use.

We define a filter function, $F: D \subseteq X \rightarrow Y$, where X is the underlying space of the point cloud data (R_n for some $n \in N$) and Y is the parameter space ($Y \in \mathbb{R}$). This is to homogenise the data. We then find the range I of the filter function F restricted to D so that the filter function is proportional to the dataset. The range I is partitioned into

subintervals S so that it creates a cover of D which overlap, so that common data points can be mapped to link the subintervals. This step produces two parameters which can be used to control the resolution, specifically the length of the smaller intervals L, and the overlap percentage between successive intervals O. For every subinterval $S_i \in S$, the following set is created which forms its domain $X_i = \{x \mid F(x) \in S_i\}$. The set $U = \{X_i\}$ forms a cover of D and $D \subseteq U_i X_i$. A suitable metric is used to get the set of all interpoint distances $B_i = \{d(x_a, x_b) \mid x_a, x_b \in X_i\}$. Using a suitable metric for calculating interpoint distances is essential for accurately quantifying dissimilarity, evaluating model performance, and ensuring algorithm convergence, as the choice of metric influences the interpretation of results and aligns with the specific characteristics and goals of the data analysis. The shortest path or distances is used known as the Euclidean distances between the clusters but other distances such as Minkowski or Manhattan distances can be used based on the needs of the study. Clusters X_{ij} are found for each X_i with the set of distances B_i. Finally, each cluster X_{ij} represents a node or a vertex in the complex and an edge is created between nodes or vertices X_{ij} and X_{rs} if $X_{ij} \cap X_{rs} \neq 0$, which means that the two clusters share a common point. A topology has now been created for the dataset, ready to be enriched. The enrichment is conducted so that PTS can be applied to plot trajectories through the cross-sectional data (generated from the TDA) based upon distances between each node or vertex using the prior knowledge of healthy and disease patients (classes, C).

The topology graph $G = (V, E)$, where V is a set of elements known as vertices and E is a set of two-sets of vertices, known as edges, produces an adjacency matrix based upon the vertices. The adjacency matrix is a square matrix that is employed to symbolise a graph. In this representation, each row and column of the matrix corresponds to a vertex in the graph. The entries within the matrix indicate the presence or absence of an edge between the corresponding vertices. Typically, a value of 1 is used to denote the presence of an edge, while a value of 0 signifies the absence of an edge. In the context of a weighted graph, the values assigned to the entries often correspond to the weights associated with the edges. This adjacency matrix is defined by a h by h matrix of binary values in which

location $(i, j) = 1$ *if* $(i, j) \in E$ and *0 otherwise*. Thus, for an undirected graph the matrix is symmetric and 0 along the diagonal. The vertices are linked via edges based on a certain level of overlap, meaning that 2 vertices will have common data points. We replace the $(i, j) = 1$ values of the adjacency matrix with the amount of overlap between each corresponding vertex, which defines a weighted distance matrix, W, of the topology, which is represented by a t by t matrix, where t_{ij} is the level of overlap between t_i and t_j. Subsequently, each vertex consists of datapoints from different classes, we determine the majority class within each vertex and calculate the amount of majority. The process of identifying the dominant class within each vertex in a graph, as well as quantifying the proportion of the dominant class, offers a more streamlined approach to representing clusters or groupings of data points. The process of aggregating data in this manner serves to improve the clarity of vertex labelling, streamline graph-based classification, and boost interpretability and decision-making, particularly in situations involving imbalanced datasets or vertex-centric methodologies. This defines a new vector $M = [m_1, m_2, \ldots, m_t)$ that represents the weighted majority classes on the vertices in the topology of the dataset. The TDA of the dataset has now been enriched to produce a weighted distance matrix, W, and a weighted class vector, M. The matrix and corresponding vector can be used as the input for building a PTS to plot the trajectories on the topology.

We define a set of pseudo time-series indices as $P = \{p_1, p_2, \ldots, p_k\}$ and every p_i is a t length vector. Subsequently, p_{ij} is defined as the jth element of p_i and each $p_{ij} \in (0, \ldots, m)$. Now the function $F(pi) = [p_{i1}, \ldots, p_{it}]$ is defined, where $F(p_i) = W(p_{ij})$ and finally, a PTS can be built by using this operator from each p_i. Also, the corresponding weighted class vector of each PTS produced by the function $F(p_i)$ is given by $G(p_i) = [M(p_{i1}), \ldots, M(p_{it})]$. We define a set of k PTS with their associated vector M sampled from the matrix W indexed by the elements of p_i. The ordering of the p_i elements is defined based upon randomly indicating a start and end p_i so that the m_{start} is a starting vertex and the m_{end} is an ending vertex in the associated topology graph. This procedure should map trajectories from a starting node, which could be a healthy state to an ending node, which should be a diseased state. However, the trajectories might pick up insightful starting and ending nodes between the different classes of disease states. The ordering can now be determined

by the shortest path, which is computed from the Floyd–Warshall algorithm [118] and then applied to the minimum spanning tree of the Euclidean distance matrix Z_i between samples in $F(p_i)$ [113]. The application of the Floyd-Warshall method for the computation of shortest pathways, when utilised on the minimal spanning tree derived from the Euclidean distance matrix across samples, guarantees an effective, comprehensive, and geometrically significant arrangement. This methodology takes into account the comprehensive interconnectivity, spatial associations, and fundamental linkages within the dataset, rendering it resilient against noise and adaptable to diverse data formats.

Finally, we can create a PTS from the TDA to show how the trajectories progress from a starting vertex (e.g., with a healthy state) to an ending vertex (disease state). The benefit of this approach is that we have discovered trajectories not only from healthy to disease states but also revealed multiple potential routes to multiple end states. Using the output of TDA as the input of PTS creates more meaningful trajectories.

Algorithm The Pseudo code of TDA-PTS Algorithm.

Input: cross section data D; class labels C, sample size m, gene expression variables n, interval length L, percentage overlap O, PTS number k.

1. Select top differential genes of dataset D.
2. Create PCA on the top differential genes of dataset D.
3. **for** $i = 1$ to m;
   3.1. Generate a filter function, F: $D \subseteq X \rightarrow Y$, where X is the underlying space of the point cloud data (R^n for some $n \in N$) and Y is the parameter space ($Y \in \mathbb{R}$).
   3.2. Find the range I of the filter function F restricted to D.
   3.3. Split I into subintervals S with length L which covers D and overlaps one another by O.
   3.4. Produce the following set $X_i = \{x \mid F(x) \in S_i\}$ for every subinterval $S_i \in S$, the set $U = \{X_i\}$ forms a cover of D and $D \subseteq U_i X_i$.
   3.5. Use the Euclidean distances to get the set of all interpoint distances $B_i = \{d(x_a, x_b) \mid x_a, x_b \in X_i\}$.
   3.6. Cluster every element X_i of U with the set of distances B_i to create a set of clusters X_{ij}, every cluster represents a vertex or a node in the graph.
   3.7. Draw an edge between nodes X_{ij} and X_{rs} if $X_{ij} \cap X_{rs} \neq 0$, which means they share a common point. This creates the topological graph $G = (V, E)$, where V is a set of elements known as vertices and E is a set of two-sets of vertices known as edges.
4. **end for**
5. Export the $h \times h$ adjacency matrix, where $(i, j) = 1$ if $(i, j) \in E$ and 0 otherwise.
6. Construct a weighted distance matrix $W = t \times t$, by replacing the $(i, j) = 1$ values of the adjacency matrix with the amount of overlap between each corresponding vertex, where $t_{ij} = O_t(t_i, t_j)$.
7. Create new vector $M = [m_1, m_2, \ldots, m_t]$, that represents the weighted majority class for each $V(X_{ij})$.
8. **for** $i = 1$ to k;
   8.1. Uniformly randomly sample t row indices from W to create w_i.
   8.2. Uniformly randomly select a row index from w_i, start, from where $1 \leq start \leq t$ and an endpoint, end, where $1 \leq end \leq t$ where $M(w_{i, start})$ represents a starting vertex and $M(w_{i, end})$ represents an ending vertex.
   8.3. Construct a $t^* \times t^*$ matrix, Z_i, of Euclidean distances between each $W(w_{ia})$ and $W(w_{ib})$ for all combinations of indices in w_i.
   8.4. Order w_i to create w_i^* based upon the shortest path between $W(w_{i, start})$ and $W(w_{i, end})$ given the weighted graph Z_i using the Floyd–Warshall algorithm constrained so that every index in w_i is included in the path.
   8.5. Add the ordered w_i^* to the set of pseudo time-series P.
9. **end for**
10. Use the set P of k PTS to train the PTS model.

Output: Pseudo Time-Series Model based upon the Topological Data Analysis graph of the dataset.

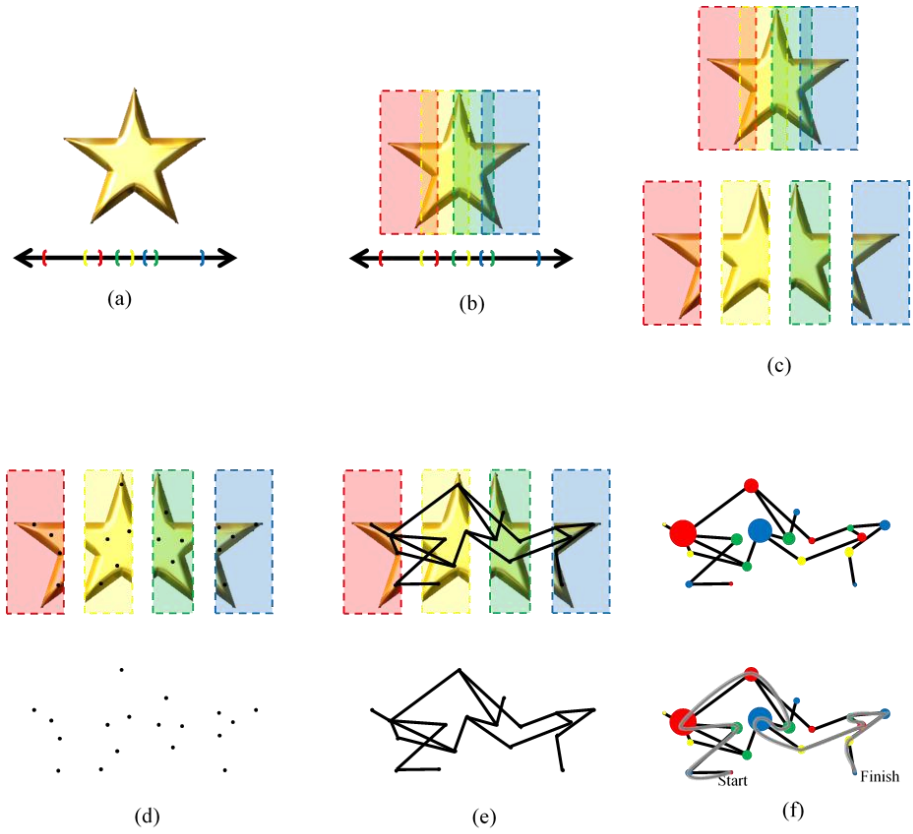


Figure 20: TDA and PTS on the star shape data cloud: a) First we project the whole data cloud to embedded space (here x-axis). b) Then we partition the embedded space into overlapping bins (here showed as coloured intervals). c) Then we put data into overlapping bins. d) Next, we use any clustering algorithm to cluster the points in the cloud data. e) Each cluster of points in every bin represents a vertex of the graph and we draw an edge between two vertices if they share a common data point. f) Then, we enrich the graph so that the vertex sizes represent the density of the cluster, and the colours represent the class majority. Finally, we create a PTS model which maps trajectories from one predefined starting vertex to another predefined ending vertex

4.4. Datasets

4.4.1. Simulated HMM Data

Simulated data was used, generated from an autoregressive Hidden Markov Model with 2 variables. This enabled the manipulation of underlying states to direct the topology

of the data. We hand-crafted the model to simulate multiple patient time-series with 5 underlying states. These represented a branching structure from a single healthy state to two possible advanced disease states via two intermediate disease states. From 100 time-series that were generated from this process, we sample a single point to mimic a cross-sectional study. Figure 21 shows some sample data generated where each data point represents a single sample from a time-series.

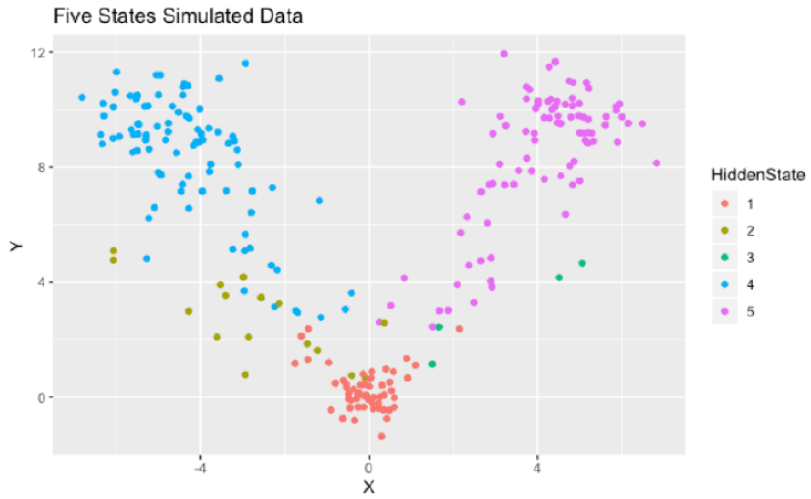


Figure 21: Sampled data from the ARHMM with 5 underlying hidden states, one representing healthy patients (red), two representing early-stage disease (brown and green) and two representing advanced disease states (blue and purple)

4.4.2. Diabetes Patient Data

The dataset used was from the MOSAIC project co-funded by the Seventh Framework Programme of the European Union, which aims to improve the way Type 2 Diabetes Mellitus (T2DM) and its related complications are diagnosed and managed [109]. The dataset is described in the previous chapter.

4.4.3. Genomic Cancer Data

The genomic cancer dataset combines 3 datasets of lung cancer, pancreatic tumour, and renal tumour patients along with their respective control patients.

The 91 human lung tissue samples in the lung cancer dataset (healthy and diseases) were analysed using the Human Genome U133 Plus 2.0 chip from Affymetrix in Hospital Universitario San Cecilio. The aim of the research for this dataset was to determine any correlation between the phenotypic heterogeneity and genetic diversity of lung cancer [119]. However, the data was processed through microarray analysis to generate expressed gene sequence, which is ideal for building trajectories through the topology of the data.

The pancreatic tumour data was collected and processed at the Mayo Clinic in the United States. 52 samples were collected and similar to the lung cancer data, microarrays were used to identify the expression differences of FKBP5 gene between the pancreatic tumour and normal samples. It was discovered that on average normal samples had more FKBP5 expression compared to tumour samples [120]. The data was processed into gene expression data, which can be used along with the lung cancer data.

For the renal tumour data set, which was collected in Erasmus Medical Centre Rotterdam, the Affymetrix microarray was used to establish the gene expression signatures of normal kidneys and different types of renal tumours. This investigation was conducted to identify and evaluate specific molecular markers with the aim of reliable diagnostics and outcome prediction of renal neoplasms [121]. The data was recorded as gene expression data.

The 3 datasets are combined, and the batch effect is removed. Additionally, the class vector is adapted to define c_i = 1, 2, 3 representing the lung cancer, pancreatic tumour, and renal tumour patients respectively and c_i = 0 represents all healthy cases. The next stage of the data pre-processing is to select the top 100 differentially expressed genes as using the entire dataset is computationally expensive and inefficient. The merged dataset, with all control samples allocated to the same class and the different cancer types assigned unique class labels, will allow the data to be analysed to see if there are links between the different cancer types based on their gene expression and whether trajectories can be built through the discovered topology.

4.4.4. Experiments

First, we apply the TDA-PTS algorithm on the simulated data. This data can be used to explore the topology and trajectories in some detail as we know the underlying states and temporal process. The positions of each sample are used within their original time-series to validate if the topology and trajectories are realistic. We should see an ordering of increasing timepoints as a trajectory is traversed. The underlying hidden states can be used to label the topology and trajectories to confirm that the start, intermediate and end states are appropriately located.

For the diabetes data we use a similar approach to the simulated data because these have a true time-series that the data was sampled from in the form of time-since diagnosis. We can validate the topologies and trajectories by using the distributions of this time information, as well as plot how real patients' time-series move over the topology.

Finally, for the genomic cancer data we have no temporal information, so we aim to explore the behaviour of the gene expression for several genes to see how they vary from trajectories as they move from control datasets to the three types of cancer.

4.5. Results

Here is where the results from the three different datasets are explored to see how combined TDA-PTS algorithm performs. As stated before, the data is pre-processed before TDA is applied and subsequently, PTS is applied to map out the trajectories through the topology. The outcome of the TDA-PTS approach is accompanied by the distribution of the datapoints through the trajectory that PTS takes.

4.5.1. Simulated HMM Data

We first investigate the results of applying the combined TDA-PTS algorithm on the simulated data. It is visible in Figure 22 how the V-shaped topology of the simulated data is preserved within the topology. We enrich the graphs using the hidden state where the initial state is in red and the two end states are in blue and purple with intermediate as green and yellow. The earlier stages of the temporal process are located to the left of the

topology and dominated by the first (healthy disease state) with nodes coloured red or orange, whilst the two end states are characterised in the top and bottom right of the topology (in blue and purple).

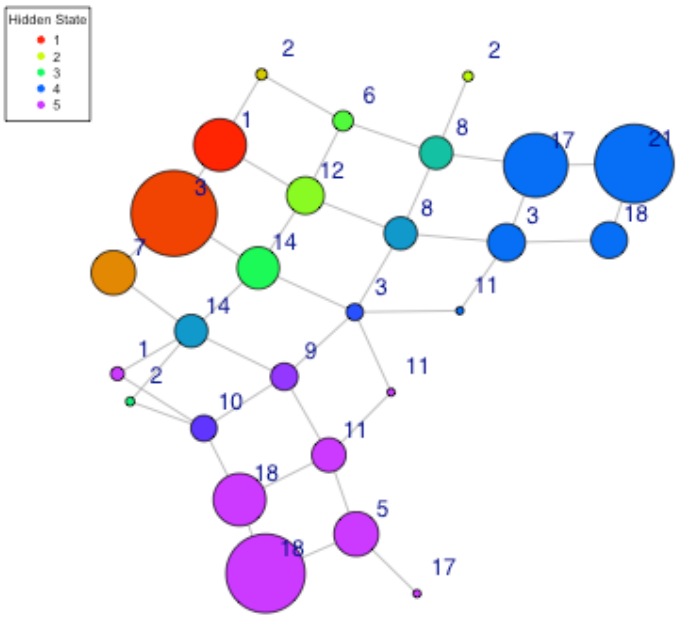


Figure 22: TDA plot learnt from the ARHMM data with colour indicating majority hidden state for data allocated to each vertex, size of vertex represents number of datapoints assigned and labels indicate the position in the original generating ARHMM

As these datapoints are sampled from a real temporal process (generated by the ARHMM) we can also label each vertex with the position in the original ARHMM time series. We would expect to see lower (earlier) positions in the time-series in the healthy red state and higher (later) positions in the disease states (blue and purple). This is exactly what is observed with the large red vertex (healthy stages) having a mean position in the time-series of 3, and the largest blue (disease stages) having a position of 21 and the purple (disease stages) having 18.

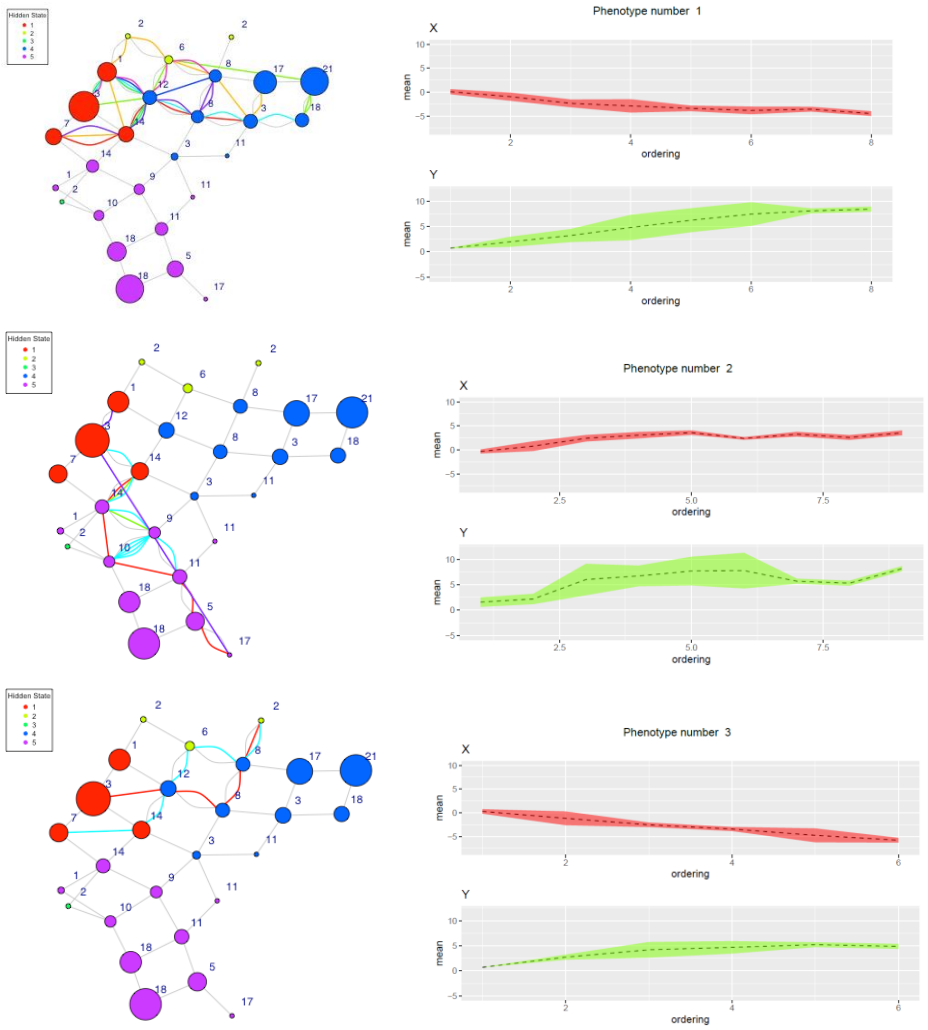


Figure 23: Sample pseudo time-series over three types of trajectories along with the distributions of the simulated variables for each vertex in the topology ordered along each PTS

Figure 23 shows three types of PTS trajectory (which we refer to as *temporal phenotype*). Notice that the trajectories over the topology capture the two main paths from healthy in red to either disease state in blue or disease state in purple (though sometimes the PTS ends early at an intermediate state as in the final phenotype). The distributions of data over the vertices on the right show that the appropriate trajectory behaviour is

captured. For example, X values decrease and Y increase for the first example phenotype (red to blue), whilst both X and Y values increase in the second phenotype (red to purple).

We now explore how the method works on two real datasets.

4.5.2. Diabetes Patient Data

We now explore the diabetes dataset. Recall for this set that we also have the real underlying time-series in terms of time since diagnosis. This allows us to explore the discovered trajectories to see if they are realistic. Figure 24 on the left shows three types of trajectories (or temporal phenotype discovered by grouping PTS trajectories that move from and to similar regions). Here, a blue vertex represents data with multiple comorbidities and red represents none. On the right of Figure 24 are the associated distributions of the key features for that phenotype: glycated haemoglobin (HbA1c), body mass index (BMI), systolic blood pressure (SBP), total cholesterol, and triglycerides. In the same way as with the simulated data we enrich the discovered topology with temporal information, here we use an index associated with time since diagnosis. The picture is quite complex with different regions capturing different stages in disease progression. For example, the far right and left has vertices associated with later stage disease (with generally higher values), whilst the centre and bottom regions of the topology capture earlier stages (with lower values). The three phenotypes capture quite distinct behaviours. For example, the first phenotype is characterised by most features decreasing in value over time, whilst the second phenotype shows an increasing trend for haemoglobin and triglycerides but more stable over other features. The third phenotype is characterised by much more variation in the middle stage of the trajectory. In summary the combination of PTS and TDA has enabled complex temporal phenotypes to be characterised over a topology.

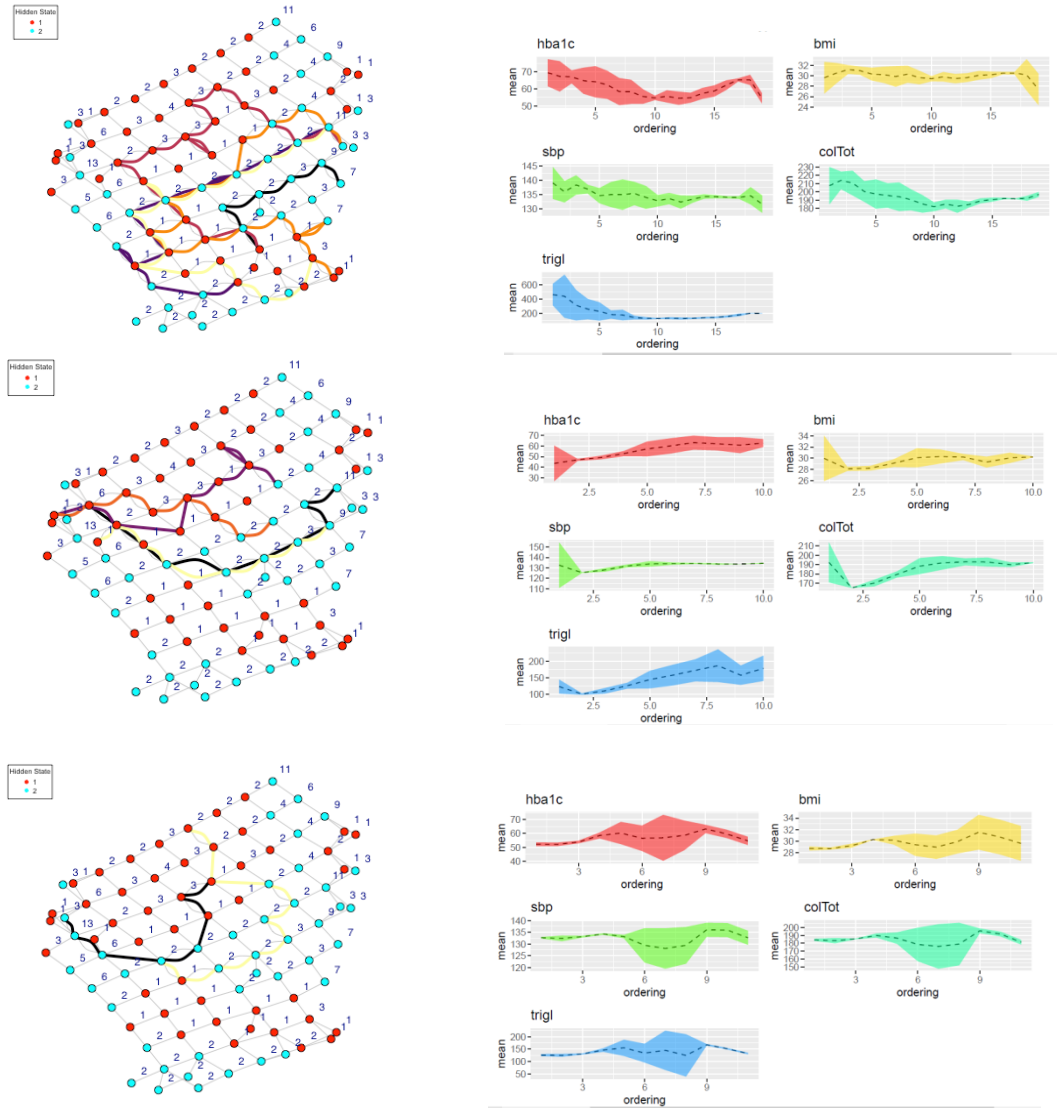


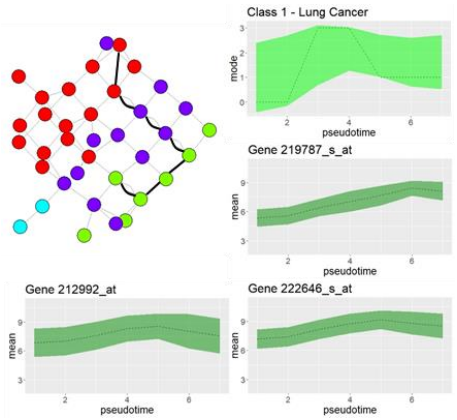
Figure 24: Three temporal phenotypes moving from absence of comorbidities (red vertices) to the presence of comorbidities (blue). On the right side, the distributions of the main clinical characteristics over the topology ordered along each PTS Trajectory

4.5.3. Genomic Cancer Data

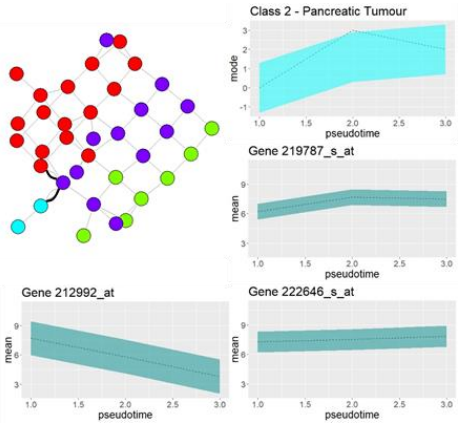
For the genomic cancer data, we do not have any temporal validation, but we have three types of labels based on cancer type: Lung, Pancreatic and Renal, which can be explored in terms of how the gene expression behaviour differs. Figure 25 shows the topology and three sample trajectories with their distribution of classes and gene values

for 3 key genes. As well as the topology on the left, we have plotted the majority class with the mean distribution of each vertex as the trajectory progresses (in a lighter colour on the right). Furthermore, we have also plotted the mean distribution of three top-ranked differentially expressed genes: 219787_s_at, 212992_at and 222646_s_at within each vertex along their corresponding trajectories. Notice that the healthy class is neatly located in the top of the topology (in red) and the three cancer types are at the bottom left and right (green, blue, and purple). Also notice in the class distribution plots how the trajectories move from vertices that are dominated by the healthy class (state 0) to ones dominated by one cancer type or another (class states 1, 2 and 3). It is interesting that the class distribution starts off with a vertex that are far more uniform in class membership but shortly after we see a mixture of class states indicating an early warning signal along the trajectory that a particular cancer type is likely further downstream.

(a)



(b)



(c)

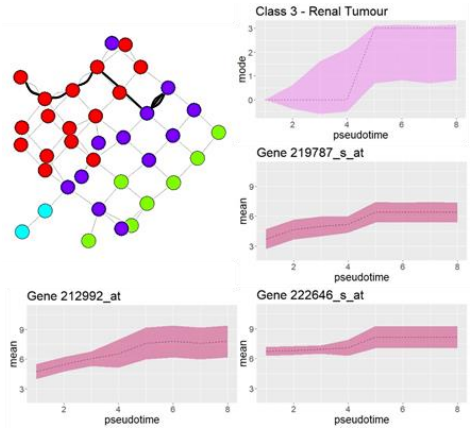


Figure 25: (Left) TDA plots with sample PTS where colours indicate the majority class at each vertex. (Right) Distribution of the 3 top-ranked genes in each vertex as they travel along their associated PTS

From the sample trajectories in Figure 25a-Figure 25c it seems that class 3 (renal tumour patients) acts like a transition point for progression to the other two cancer types. This implies that there could be similar genetic mechanisms at play between the cancers in the early stages. The 3 top-ranked gene distribution plots give us an effective visual representation of how each gene progresses through the different trajectories. Figure 25a shows that all three genes generally increase as the trajectory progresses. Gene 219787_s_at starts very low at the beginning of the trajectory, which is an indication that for healthy patients, this gene has a low distribution but the gene increases and reaches a peak at vertex 6. Subsequently, for pancreatic cancer in Figure 25b, gene 219787_s_at shows a similar increasing trend though to a lesser degree, whilst 212992_at shows a decreasing trend. For the renal tumour plots in Figure 25c, the gene distribution for all 3 gene peaks at the same point where the class distribution of the vertex become dominantly class 3 (renal tumour).

Gene 219787_s_at has been identified as the epithelial cell transforming sequence 2 oncogene (ECT2) [122]. ECT2 has an oncogenic role in lung adenocarcinoma cells, and it is stated to be commonly upregulated at the early-stage lung adenocarcinoma [123]. From Figure 25 we can see that ETC2 has the largest increase in distribution when the

trajectory goes towards a vertex which is predominantly occupied by lung cancer patients. This is a strong indication that this gene may be a potential target for the treatment of lung cancer and can work as a signal for early-stage intervention. Gene 212992_at, also known as AHNAK nucleoprotein 2 (AHNAK2) and gene 222646_s_at, also known as endoplasmic reticulum oxidoreductase 1 alpha (ERO1A) are proteins within humans. These proteins have been identified by the TDA-PTS approach as possibly being an indication to disease progression. This is further insight on how effectively the gene distribution for the 3 top-ranked genes can help in discovering key stages in disease progression. It helps to demonstrate the power of combining topology with trajectory analysis in identifying key points in a topology/trajectory that can enable earlier intervention. Being able to identify the genes that seem to contribute or drive the disease progression will contribute in a clinical setting. Further testing and observing these highly expressed genes can reenforce the significance of the gene on the disease. This will allow clinicians to trial drugs that can affect the genes of interest to see if the disease progression can be slowed down or even reversed. This can also contribute to implementing personalised prognosis and treatment for patients.

4.6. Summary

In this chapter, a novel method, TDA-PTS, has been described with the aim to create realistic disease trajectories over meaningful topologies based upon graph theory, distance metrics and expert knowledge on staging within a disease.

It has been demonstrated how the resulting TDA plots and the associated PTS trajectories capture realistic temporal processes in three different case-studies. Firstly, we assessed the ability to learn realistic temporal processes by using data that has underlying temporal information enabling us to validate the resulting trajectories: Time-indexed simulated data and time-since-diagnosis information from real diabetes data was used to enrich the topologies and trajectories showing realistic processes form early stages to late stages. The simulated data experiments showed that the approach can successfully extract the underlying temporal phenotypes from starting states to multiple end states. The

diabetes data showed that complex temporal phenotypes could be extracted with distinct characteristics that moved from early stages to late stages. We also explored the use of genomic cancer data which does not have any temporal validation but allowed us to explore three different cancer types. From the sample topologies and PTS trajectories we found realistic temporal phenotypes transitioning from uniform healthy states, through diverse intermediate states and ending in uniform but distinct cancer states.

The approach discussed in this chapter has provided proof-of-concept results, which gives an incentive to further explore how combining both shape and sequential analysis can assist in explaining complex disease processes. These results will be used as a steppingstone to add a further layer to this novel approach, which will be discussed in the next chapter.

Chapter 5 Implementing CBPTS to Improve Disease Progression Trajectory

5.1. Chapter Outline

The objective of this chapter is to implement the novel constraint-based pseudo time series (CBPTS) to simulated and real-life data to test if creating constraints leads to improved disease progression trajectories. This chapter is organised as follows: Section 5.2 provides an introduction, by setting the scene of how the novel approach will be constructed and implemented. Section 5.3 will introduce and explain how the CBPTS approach will be applied, and progression trajectories mapped. Section 5.4 presents and describes the two datasets that will be used for CBPTS and the testing methodology. Section 5.5 presents the results of CBPTS on the two datasets. Section 5.6 provides a summary.

5.2. Introduction

In the previous chapter, a novel combined machine learning approach known as TDA-PTS was introduced, which aims to map the topology of the data and then build multiple trajectories through that topology. This approach can be used to construct temporal models for prediction, which can be further explored to cluster and identify insightful intermediate stages in disease progression. Yet, it must be noted that this approach will build temporal models using cross-sectional data but without genuine time stamps, it may cause some limitations to the model.

This chapter will use prior knowledge in the form of disease labelling and staging to improve PTS analysis. Start and end points are predefined, and staging information in the form of different labels is exploited to produce more reliable pseudo time trajectories of disease that can inform clinicians about how disease progresses. Specifically, we use constraints [124] [125] to direct the pseudo time algorithm based upon pre-existing knowledge of disease encoded as staging labels. This approach will allow the implementation of certain rules to prevent impossible switches in disease state such as

trajectories switching from a degenerative disease state to a healthy state or from an older population to a younger population. This improved approach will be used to explore the simulated dataset used in the previous chapter and a freely available dataset on breast cancer [126].

5.3. Constraint-Based Pseudo Time Series (CBPTS)

The dataset, *D*, which is used, can be defined as a real valued matrix of *m* by *n*, where *m* (columns) is the number of samples that are the patients and *n* (rows) is the number of variables that are the clinical features in the data. D_i can be defined as the *i*th column of matrix *D*. Furthermore, $C = [c_1, c_2, \ldots, c_m]$ is used as a vector that represents class labels of the dataset, where $c_i \in \{1,2\}$ corresponding to the sample *i*. Subsequently, $c_i = 2$ and $c_i = 1$ represents the patients in the sample *i* that are the diseased cases and the healthy cases within the dataset respectively. The classes of the sample *i* have been determined based upon the diagnosis made by experts or clinicians. Constraints are defined by banning certain trajectories based upon prior knowledge or clinicians' knowledge (e.g., banning trajectories from $c_i = 2$ to $c_i = 1$). Furthermore, other data features can be used as a staging proxy by creating a staging vector, $S = [s_1, s_2, \ldots, s_n]$, where $s_i \in \{1,2,3,\ldots, k\}$ corresponding to the sample *i* that are the progressive diseased cases and the healthy cases within the dataset respectively. The staging proxy of the sample *i* have been extracted from the features of the collected dataset. Subsequently, constraints can be applied to restrict receding trajectories in the staging vector (e.g., banning trajectories from $s_i = 2$ to $s_i = 1$, $s_i = 3$ to $s_i = 2$, etc.).

Now, a full distance matrix, *W*, can be built by using the Euclidean distances between the datapoints but other distances can be used based on the needs of the study. To create the constraints and reduce the probability of certain trajectories, the distances between impossible transitions are increased dramatically. This means biasing the distance from $c_i = 2$ to $c_i = 1$ so that a PTS will not construct an unrealistic trajectory from a disease state to the health state. Also, distances can be biased from $s_i = 2$ to $s_i = 1$, $s_i = 3$ to $s_i =$

2, etc. to prevent trajectories from moving back through the stages. For this study we simply set distances for these "impossible" transitions to 999.

A weighted matrix has now been created for the dataset, which is ready for constructing PTS. PTS can be applied to plot trajectories through the cross-sectional data (generated from the distance matrix) based upon distances between each node or vertex using the prior knowledge of healthy and disease patients along with the defined constraints.

A set of pseudo time-series indices can be defined as $P = \{p_1, p_2, \ldots, p_k\}$ and every p_i is a t length vector. Subsequently, p_{ij} is defined as the jth element of p_i and each $p_{ij} \in (0, \ldots, m)$. Now the function $F(p_i) = [p_{i1}, \ldots, p_{it}]$ is defined, where $F(p_i) = W(p_{ij})$ and finally, a PTS can be built by using this operator from each p_i. Also, the corresponding class vector of each PTS produced by the function $F(p_i)$ is given by $G(p_i) = [C(p_{i1}), \ldots, C(p_{it})]$. A set of k PTS can be defined with their associated vector C sampled from the matrix W indexed by the elements of p_i. The ordering of the p_i elements is defined based upon randomly indicating a start and end p_i so that the m_{start} is a starting vertex and the m_{end} is an ending vertex in the associated weighted graph. This procedure should map trajectories from a starting node, which could be a healthy state to an ending node, which should be a diseased state. The ordering can now be determined by the shortest path, which is computed from the Floyd–Warshall algorithm [118] and then applied to the minimum spanning tree of the distance matrix [113]. Having the constraints implemented into the distance matrix results in the shortest path algorithm avoiding the impossible trajectories due to the high distances.

Consequently, a PTS is created from the topology to show how the trajectories progress from a starting vertex (healthy state) to an ending vertex (disease state) through transition vertices (different stages of disease state). The benefit of this approach is that the discovered trajectories not only from healthy to disease states but also revealed multiple potential routes to multiple end states. Using the output of TDA as the input of PTS creates more meaningful trajectories [127].

Finally, the PTS points are extracted from the model to setup a Hidden Markov Model (HMM). Subsequently, the HMM is fitted to PTS data and the probability matrix of

transitions between each staging within the data is extracted. This entire process is repeated 100 times to obtain a more robust mean transition matrix for the according constraint. Furthermore, the constraints applied to the trajectories are increased to see the effects upon the PTS and subsequent transition matrix. This approach will be trained and tested on the simulated data to see how incrementally adding constraints affects the transition matrix compared to the original transition matrix where the data has been simulated from. Afterwards, the approach is applied to the Wisconsin Breast Cancer dataset and the PTS with its associated HMM are inspected.

The Wisconsin Breast Cancer dataset does not have an actual transition matrix to use as a comparison for the output transition matrix obtained from the PTS. However, the no constraint transition matrix can be used as a starting to point to visualise how the trajectories move throughout the dataset. Subsequently, prior knowledge regarding the attributes can be used to aid the implementation of constraints such as associating increase in Uniformity of Cell Size to the progression of the cancer. This will allow a more robust CBPTS to be constructed and state transition diagrams to be drawn.

5.4. Datasets

The simulated HMM data that is used in this investigation has been described in the previous chapter and is shown in Figure 21. The other dataset used is the Wisconsin Breast Cancer dataset. Using real-life data is susceptible to missing or noisy data, which is not present within the simulated HMM data. Nevertheless, a method to deal with missing data must be implemented. In this research, any missing data will be omitted from the testing as the aim is to assess the effectiveness of the novel approach without external factors affecting the execution of the method.

This dataset consists of nine cytological characteristics of breast fine-needle aspirates (FNAs), collected by Wolberg and Mangasarian at the University of Wisconsin Hospitals in 1990 [126]. From the original 699 patients, 683 were used due to missing attribute values. The aim of the research was to use the nine attributes to determine if the samples (patients) were benign or malignant [128]. All nine attributes have an impact in the

diagnosis and therefore, the attributes can be exploited to find disease progression trajectories. The features and class are listed below along with their possible values / states:

1. Clump Thickness (1 – 10)
2. Uniformity of Cell Shape (1 – 10)
3. Marginal Adhesion (1 – 10)
4. Single Epithelial Cell Size (1 – 10)
5. Bare Nuclei (1 – 10)
6. Bland Chromatin (1 – 10)
7. Normal Nucleoli (1 – 10)
8. Mitoses (1 – 10)
9. Uniformity of Cell Size (1 – 10) – used as a staging proxy

- Class: (Benign / Malignant)

Understanding the attributes will enable us to use one of the attributes as a staging proxy. We make the strong assumption that as the cancer progresses, the uniformity of cell size increases within the patient. By constraining the trajectories from going back in cell size, we aim to construct more robust CBPTS to model how the disease progresses. Figure 26 shows a principal component analysis (PCA) plot of the first two principal components, with each point coloured based on the 2-class dataset (benign and malignant tumours) on the left, and the 10 staging states based on uniformity of cell size on the right.

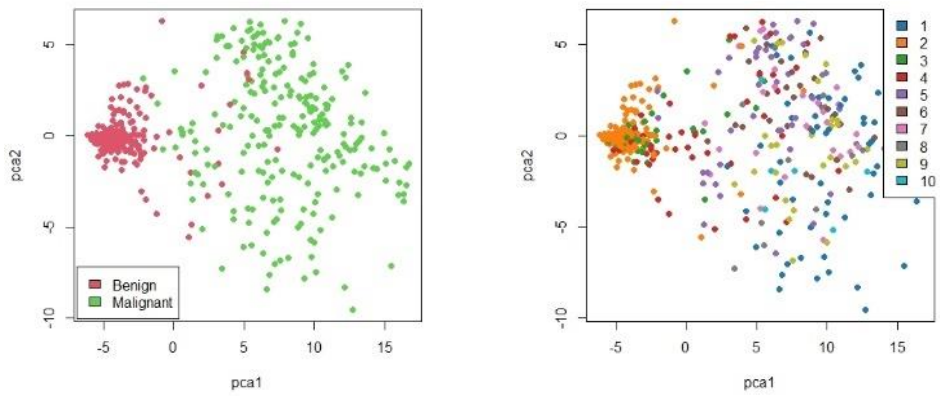


Figure 26: Principal Component Analysis plot of the Wisconsin Breast Cancer data showing 2 class dataset (left) and 10 staging states based on uniformity of cell size (right)

5.4.1. Experiments

Initially, the standard PTS algorithm is applied on the simulated data to see how the natural trajectories form and to explore error trajectories that are present. After understanding how the trajectory travels, the underlying hidden states are used to label the topology and trajectories to confirm that the start, intermediate and end states are appropriately located. Next, incremental constraints are applied to the trajectories to form CBPTS and the mean transition matrices are extracted from each set of constraints. Lastly, the constraints are compared to the actual transition matrices on which the data was simulated. The difference and variances of each constraint parameter is compared to see if adding constraints reduces the error in each parameter.

Finally, for the Wisconsin Breast Cancer data, actual transition matrices do not exist, so the standard PTS algorithm is applied to see the natural trajectories and transition probabilities once again. Using prior knowledge of the data attributes such as uniformity of cell size and how it shows progression of the disease, constraints can be applied to ban trajectories receding in cell size. The Wisconsin dataset has nine other attributes apart from the uniformity of cell size, which will be used as a class vector for progression, such as clump thickness, bare nuclei, etc. To utilise as much of this information as possible for the analysis, the attributes have been normalised and PCA applied to reduce the

dimensionality. The first two principal components are used as it accounts for 87% of the overall variability of the data.

5.5. Results

Here is where the results from the two different datasets are investigated to see how constraint-based pseudo-time series performs. As stated before, the data is pre-processed before the distance matrix is constructed, constraints are implemented and subsequently, PTS is applied to map out the trajectories through the topology. The outcome of the CBPTS approach is shown accompanied by the effects of the constraints on the trajectory that PTS takes.

5.5.1. Simulated HMM Data

We first investigate the results of applying the combined CBPTS algorithm on the simulated data. The V-shaped topology of the simulated data is preserved within the topology (Fig 3a). We enrich the graphs using the staging state where the initial state is in blue and the two end states are in red and purple with intermediate as green and yellow. Initially, no constraints were implemented to see the natural trajectories (3a left) and variances of the data parameters (4a left). Constraints were then incrementally applied to restrict backward trajectories from intermediate to start, end to intermediate and end to start points. As the constraints were increased, the difference between the extracted transition matrix and the actual matrix were calculated and shown on a graph with their respective variances. Fig. 4b shows a sample of this graph for a specific cell in the matrix, where it can be seen how the transition matrix improves with the increase of constraints until it collapses due to over constraining the trajectories. The output trajectories of the constrained CBPTS are shown in Fig 3a (right) and the variances of the parameters can be seen in 4a (right). Fig. 3b shows distributions of the staging labels over the pseudo time for the constrained CBPTS (left) and the standard unconstrained PTS (right).

(a)

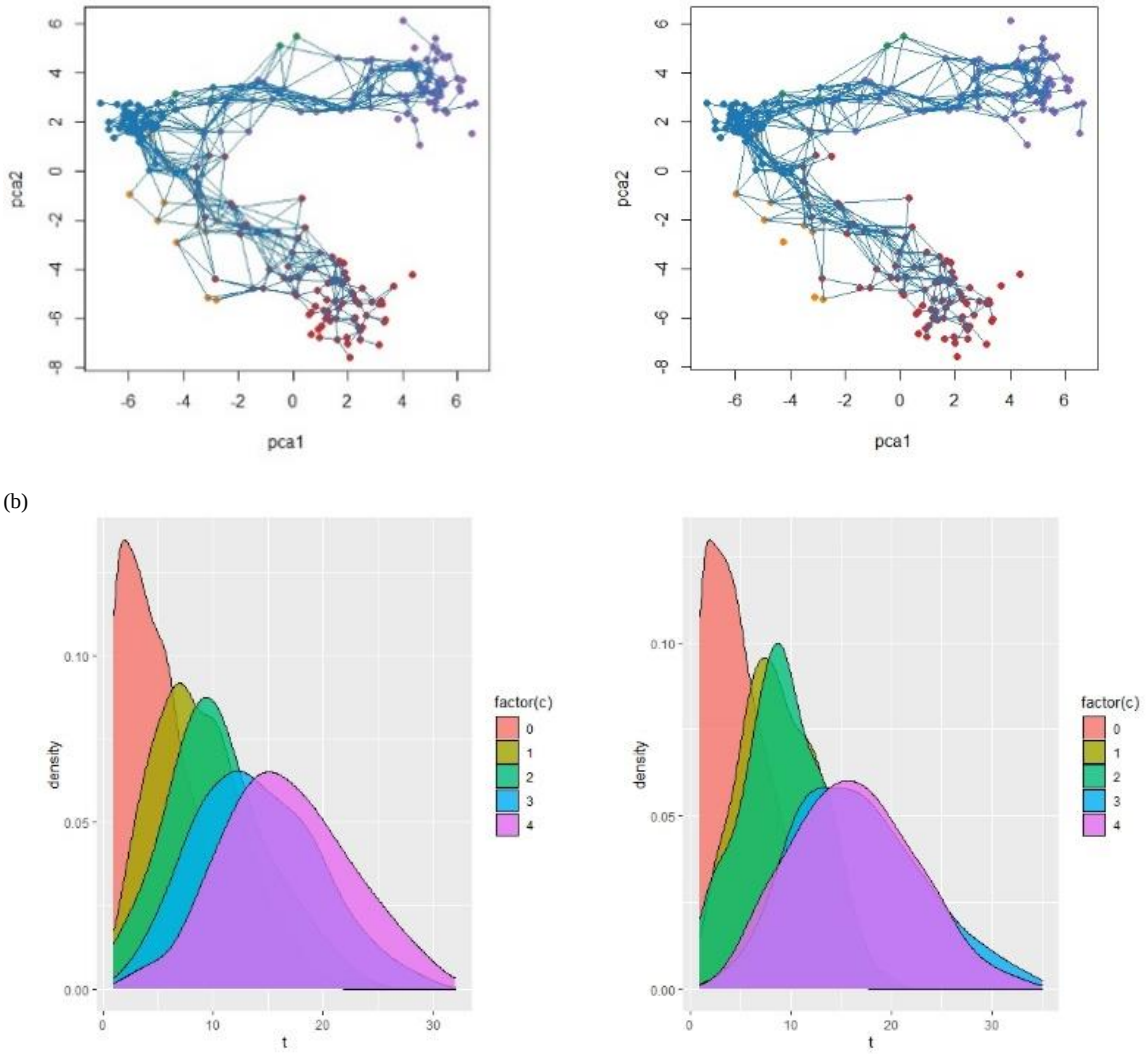
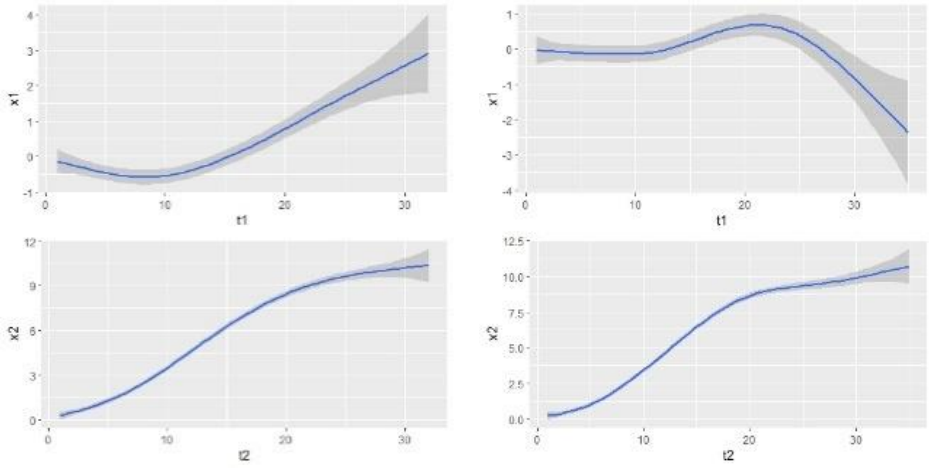


Figure 27: Simulated HMM data analysis, a) PCA plot of staging with no constraints (left) and with constraints (right), b) trajectory density with no constraints (left) and with constraints (right)

From Figure 27a, it is evident that both methods create trajectories over the V-shaped topology, but the standard PTS generates slightly more impossible crossover links nearer the junction (with switching between the early-stage states). However, CBPTS restricts such trajectories and allows the paths to be constructed to show the progression, which produces a more robust model for further analysis. From the staging distributions in Figure 27b we can see that the trajectories generally should travel from a healthy state to one of the two intermediate states and finally end up in the appropriate one of two final

advanced disease states. The no constraint densities (left) show this trend but there appears to be a bias to ordering stages as follows 1→2→3→4→5 whilst the constraint-based distributions (right) correctly illustrate two more distinct paths 1→2→4 and 1→3→5 representing the underlying V structure. This is further evidence of how constraint-based analysis can produce a most robust model.

(a)



(b)

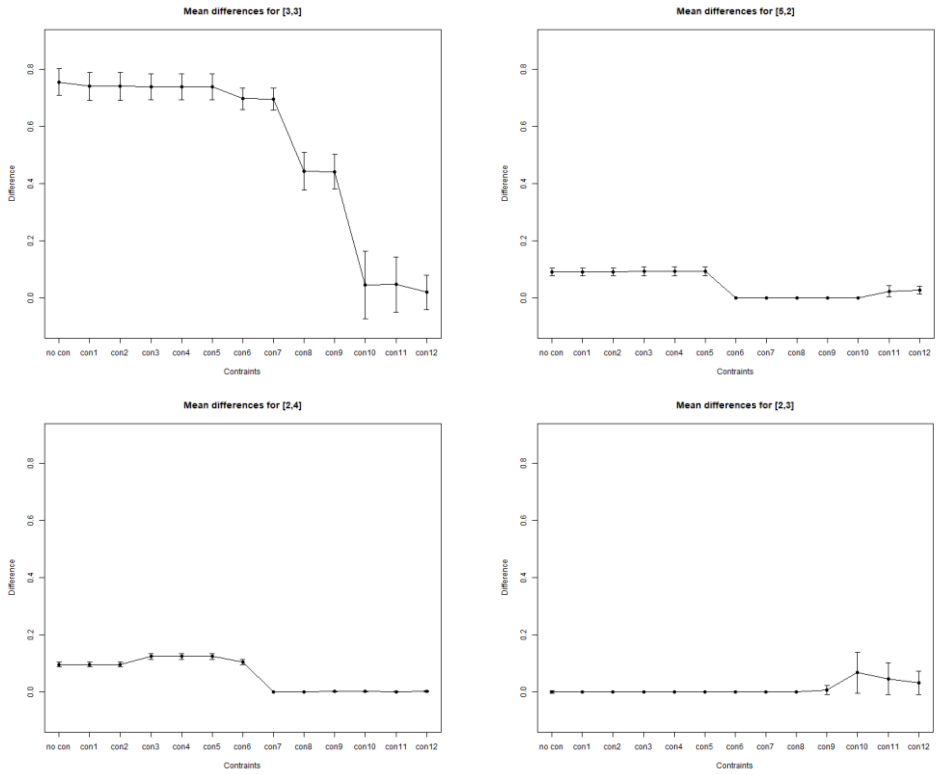


Figure 28: a) Trajectory behaviour of features over pseudo-time with no constraints (left) and with constraints (right), b) Error in estimated transition parameters for increasing number of constraints on 4 sample transition parameters [3,3],[5,2],[2,4],[2,3]

The distributions of data over the trajectories are shown in Figure 28a for each feature. The trends displayed in X2 are very similar between with and without constraints, but the main difference is evident in X1. In Figure 28b we explore how close model parameters learnt from the pseudo time data matches the original underlying transition probabilities. Four such transition parameters are shown. It can be seen how the difference in parameters and variances of the extracted transition matrix to the actual transition matrix improves as the constraints are increased. In some cases the use of many constraints can lead to greater difference to the original parameters (e.g., transition [2,3] in the bottom right of Figure 28b when more than 10 constraints are imposed). This over constraining may have led to unrealistic trajectories being forced due to limited samples. Table 4 shows how the error in the estimated transition compared to the original underlying transition probability improves when constraints are introduced. There are negligible differences for many. However, for all significant changes (in bold) we see improvements - for transitions [1,2], [2,2], [3,3], [4,4] and [5,3] - supporting the effectiveness of implementing the constraint-based approach to PTS construction.

Table 4: Error in estimated transition and difference

State Transition	Probability from No Constraint Model	Probability from Constrained Model	Difference
[1,1]	0.020	0.025	-0.005
[1,2]	0.337	0.028	**0.309**
[1,3]	0.022	0.003	0.019
[1,4]	0.036	0.038	-0.002
[1,5]	0.038	0.038	0.000
[2,1]	0.096	0.044	0.052
[2,2]	0.592	0.064	**0.528**
[2,3]	0.001	0.033	-0.032
[2,4]	0.097	0.003	0.094
[2,5]	0.036	0.003	0.033
[3,1]	0.047	0.049	-0.001
[3,2]	0.002	0.007	-0.005
[3,3]	0.756	0.021	**0.735**
[3,4]	0.000	0.001	-0.001
[3,5]	0.031	0.001	0.030
[4,1]	0.019	0.017	0.002
[4,2]	0.162	0.070	0.092
[4,3]	0.000	0.036	-0.036
[4,4]	0.166	0.056	**0.110**

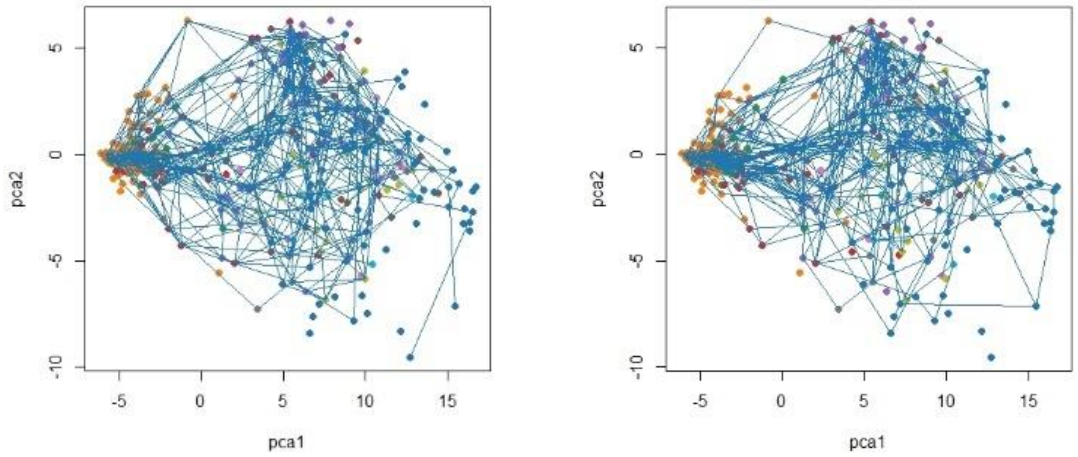
[4,5]	0.001	0.015	-0.014
[5,1]	0.048	0.015	0.032
[5,2]	0.091	0.027	0.064
[5,3]	0.778	0.051	**0.727**
[5,4]	0.033	0.015	0.018
[5,5]	0.105	0.057	0.048

5.5.2. Wisconsin Data

We now explore the Wisconsin breast cancer dataset. Recall for this dataset that we do not have an actual transition matrix as we did with the simulated data. Hence, we build a standard PTS with no constraints to see how the trajectories naturally form based on uniformity of cell size. Subsequently, we extract the transition matrix and apply constraints to restrict any trajectories from receding in size.

Looking closely at the distribution of stages over pseudo time in Figure 29b, we can see that for both methods there is a general trend for earlier stages to be correctly identified earlier in pseudo time and later stages to be identified with later points. However, when constraints are placed on the trajectories (right) the earlier stages are better identified with the earlier points in pseudo time, particular stages 1 and 2. Comparing the feature variance over pseudo-time in Figure 30a and Figure 30b, we can clearly see the effects of constraining the trajectories. Figure 30a shows how the feature progression over pseudo-time is generally increasing but it can fluctuate as seen in var 5 and 7, as well as having a general poor variance especially in var 8. However, Figure 30b illustrates how the constraints improve the feature progression over pseudo-time by showing a smoother and monotonically increasing values for all features along with a slowly increasing variance until the end of the pseudo-time as is expected. Finally, the effectiveness of the constraints is further shown in Figure 30c, where the state transition diagram confirms how the trajectories do not allow a backwards transition in the disease staging (all states only transition to the same or a later state). CBPTS has enabled more robust and meaningful trajectories to be generated.

(a)



(b)

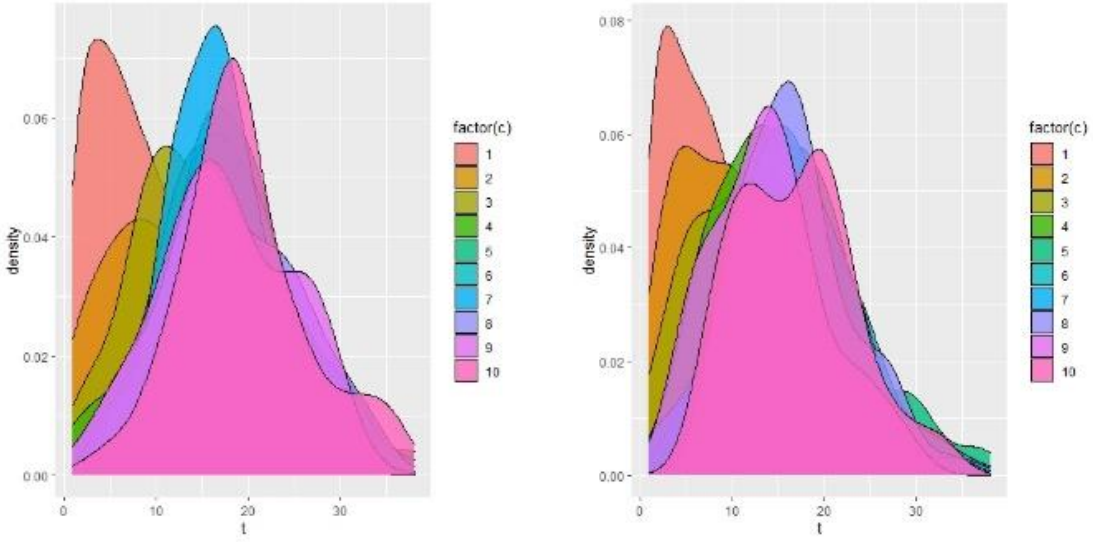
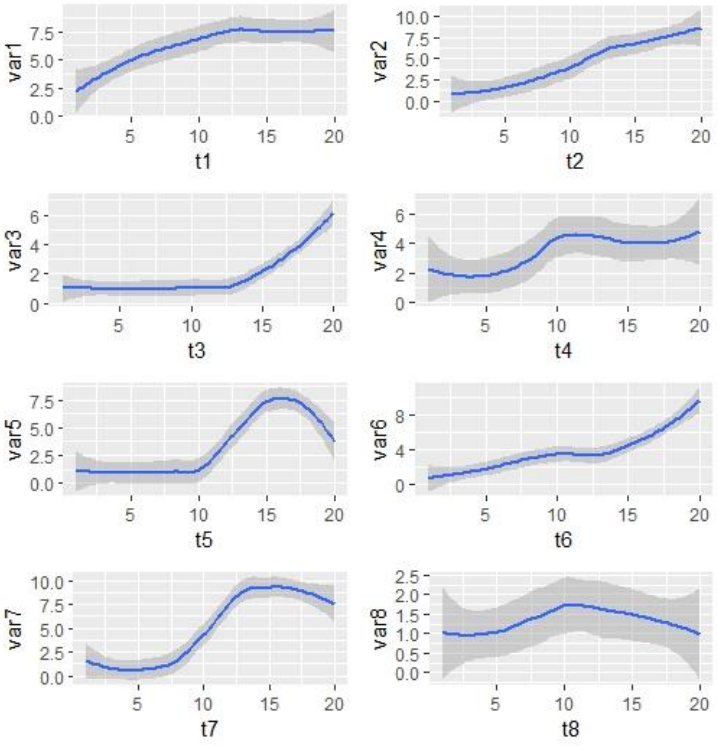
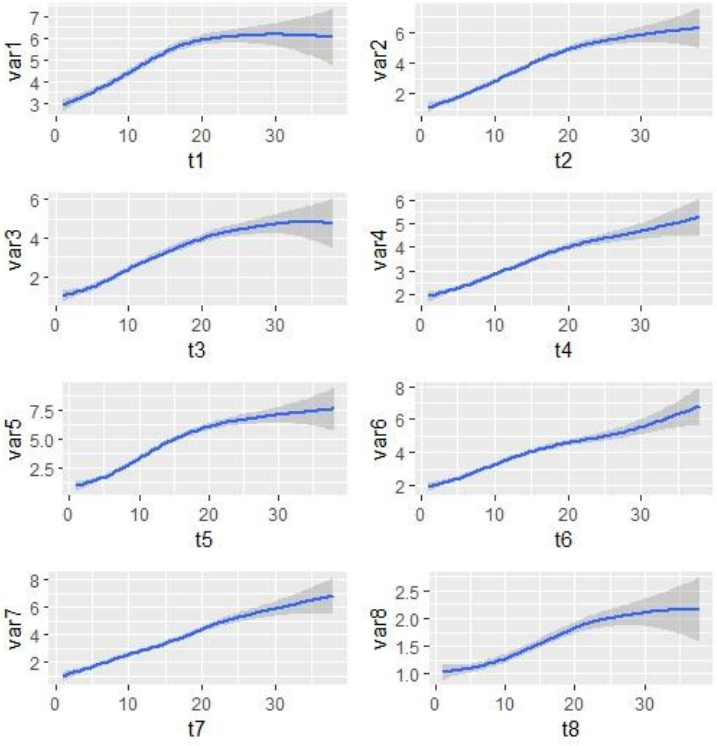


Figure 29: Wisconsin breast cancer data analysis, a) PCA plot of uniformity of cell size as staging with no constraints (left) and with constraints (right), b) trajectory density with no constraints (left) and with constraints (right)

(a)


var1
t1
var2
t2
var3
t3
var4
t4
var5
t5
var6
t6
var7
t7
var8
t8

(b)


var1
t1
var2
t2
var3
t3
var4
t4
var5
t5
var6
t6
var7
t7
var8
t8

(c)

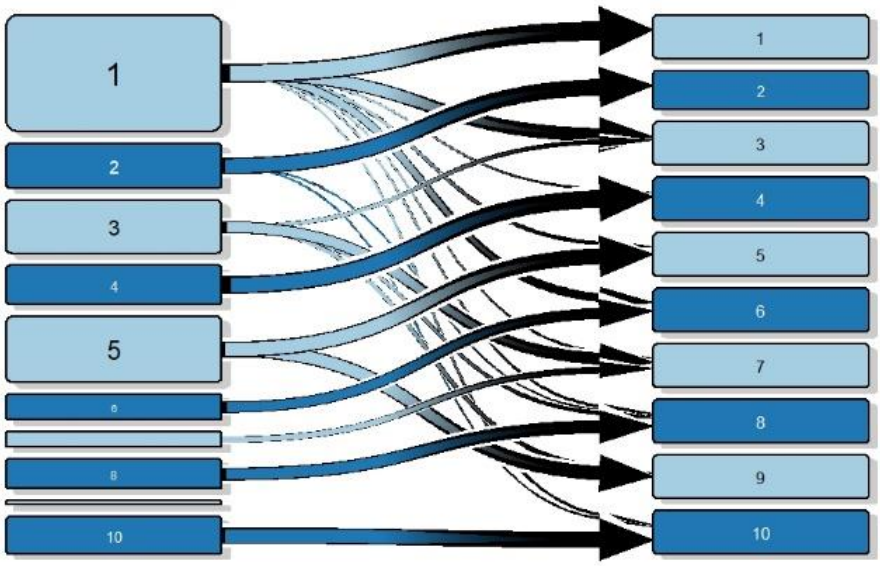


Figure 30: Trajectory behaviour of features over pseudo-time with no constraints b) Trajectory behaviour of features over pseudo-time with constraints, c) State transition diagram for disease stages for constraint-based trajectories

5.6. Summary

In this chapter, a novel approach (CBPTS) is described for building pseudo-time series with constraints based upon pre-existing knowledge of disease encoded as staging labels. We have demonstrated how the implementation of these constraints has generated CBPTS trajectories that prevent impossible switches in disease state such as trajectories switching from a degenerative disease state to a healthy state. Exploring this approach on the simulated data has shown how effectively these constraints affect the trajectory formation. Also, studying the extracted transition matrices from the incrementally constrained pseudo-time, we can clearly see that these sets of rules have narrowed the gap between the transition probabilities of the trajectory and the true underlying transition probabilities that were used to generate the simulated data. However, the risk of over constraining must be considered as the pseudo-time may force certain trajectories depending on the sample size used to construct the trajectory.

The Wisconsin data gave us the opportunity to evaluate the same approach used on the simulated data but on real breast cancer data. We constrained trajectories from reducing in uniformity of cell size (essentially using the feature as a proxy for disease staging). CBPTS has effectively built dense areas of trajectories, which is an effective insight for further analysis and clinical intervention. Furthermore, we can see that eliminating impossible staging transitions results in significantly improving how the features of the data progress over pseudo-time with smoother progression of disease. This is a clear indication of how using pre-existing knowledge of diseases to direct the pseudo-time algorithm will result in more meaningful trajectories to aid the clinicians in better diagnosis of the disease at an earlier stage and for more personalised treatment.

Chapter 6 A Case Study in Ophthalmology

6.1. Chapter Outline

The objective of this chapter is to implement the novel methods and algorithms described in the previous chapters to three different glaucoma datasets. This chapter is organised as follows: Section 6.2 provides an introduction. Section 6.3 describes glaucoma and touches on some medical background before addressing the datasets and Visual Fields. Section 6.4 briefly illustrates how the experiment will be carried out on the glaucoma data. Section 6.5 shows the results from the experiments and what kind of implications can be applied. Section 6.6 provides a summary.

6.2. Introduction

When constructing sample trajectories for glaucoma, multi-dimensional scaling is applied to the first two datasets to plot the first two components. Trajectory comparison in a healthcare setting is demonstrated. The probability of transitioning between states are then represented by their corresponding final state diagrams. Normative values of clinical data are supplied for both healthy and diseased individuals to facilitate comparisons between the two groups. We also present the HMM learned from the unlabelled pseudo-time-series, which we use to predict the values of the data associated with each state in these diagrams. Subsequently, the TDA-PTS and CBPTS methods are applied to the final multi-dimensional glaucoma visual field dataset to model and illustrate the disease progression with the use of a clinical scoring system to guide the trajectories.

6.3. Real-World Cross-Sectional Data

Globally, glaucoma is the primary cause of blindness. Sometimes diagnosis is delayed because symptoms may not appear until a reasonably advanced stage. Primary care physicians can better suggest high-risk patients for full ophthalmologic examination and actively participate in the care of patients with this illness if they have a basic awareness of the disease's pathophysiology, diagnosis, and therapy [129]. Glaucoma is a neuropathic

disease of the eye and is the main cause of permanent blindness around the globe, as it affects more than 70 million individuals all over the world, of which about 10% are affected to the point that they are blind in both eyes [130]. Glaucoma can go unnoticed until it has reached a severe stage, which means that the number of people who are afflicted by it is likely to be far larger than the number of people who are now recognised to have it. Moreover, population surveys have shown that only about one-fifth to half of those who have glaucoma know they have it. [131], [132]. Although glaucoma currently has no known cure, it has been established through clinical practise that early therapy can reduce the progression of the illness [133], [134]. On the other hand, early diagnosis is a task that is extremely difficult to do due to the variable nature of the pathology and its overlap with the physiology of the individual.

There are two main kinds of glaucoma, open-angle glaucoma, which is the most common type and features an open angle of the anterior chamber, alterations to the optic nerve head, a loss of peripheral vision followed by a loss of central vision, and is a chronic, progressive, and irreversible form of multifactorial optic neuropathy [135]. High intraocular pressure, whether from primary or secondary sources, is a major risk factor for open-angle glaucoma. Knowing more about open-angle glaucoma will help stop the severe blindness that results from the disease if it is left untreated. A patient's eyesight will deteriorate gradually and permanently due to this ailment, but they won't experience any symptoms until it's too late. The iris and cornea's drainage angle are still open. However, some areas of the drainage system are not functioning effectively, which is why this will lead to a gradual rise in eye pressure. The other main type of glaucoma is angle-closure glaucoma which occurs when the iris bulges. The drainage angle is partially or entirely blocked by the bulging iris. As a result, the eye's pressure rises, and fluid cannot flow through it. Angle-closure glaucoma can develop gradually or suddenly.

Glaucoma is typically detected during a routine eye exam, frequently before it manifests any obvious symptoms. Afterward, more tests are frequently required to identify and track the problem. If an optometrist suspects you have glaucoma following a standard eye exam, they can perform a variety of tests. The pressure of the eye can be tested using a tonometer in a process called tonometry. This procedure will be repeated

during prognosis in order to determine if there are signs of glaucoma. Gonioscopy can be performed on the patient, whereby the front part of the eye is examined to see if the fluid drains out of the eye. This procedure can aid in determining if the area or angle is open or blocked, which can affect how fluid drains out of the eye. Subsequently, this information will inform the optometrist what type of glaucoma is present. The health of the optic nerve can be assessed through optical coherence tomography (OCT), which is a non-invasive imaging test that will take cross-sectional images of the retina with the use of light waves. Finally, a visual field (VF) test, also known as perimetry, can be conducted to check for areas of vision loss, especially in the peripheral vision. This is the most common test for glaucoma, which will be explained in the next sub-section.

Glaucoma can occur for several reasons, but most cases are caused by a build-up of pressure in the eye when fluid is unable to drain effectively, which can damage the optic nerve. It is often unclear why this happens but there are certain risk factors that will prompt a referral from a medical practitioner such as, older age, certain ethnicities (people of African, Caribbean or Asian origin are at a higher risk), family history of glaucoma, other medical conditions (such as short-sightedness, long-sightedness and diabetes) and high intraocular pressure. The basic objectives of glaucoma treatment are to slow disease progression and maintain quality of life. It may be earlier than previously believed that glaucoma causes a decline in quality of life, highlighting the significance of early detection and treatment. The only treatment for glaucoma that has been shown effective is lowering intraocular pressure [129], commonly with the aid of eyedrops and laser treatment if eyedrops are ineffective, where a high-energy light beam is targeted towards part of your eye to stop fluid building up inside it. Consequently, in rare cases surgery can be recommended, which involves removing part of the eye-drainage tubes to increase the ease of fluid drainage.

6.3.1. Heidelberg Retina Tomography and Visual Field Data

The Visual Field (VF) test evaluates how sensitive the retina is to different types of light. In most cases, automated perimetry is used to measure it. This is a technique in

which the patient observes a dark background while brighter spots of light are shone onto the background at various positions in a regular grid pattern. The degree of retinal sensitivity a patient possesses determines the level of brightness at which that subject can see spots of light [136]. There are a variety of disorders and ailments that can impact the VF, the most prevalent of which being glaucoma and neurological problems [137].

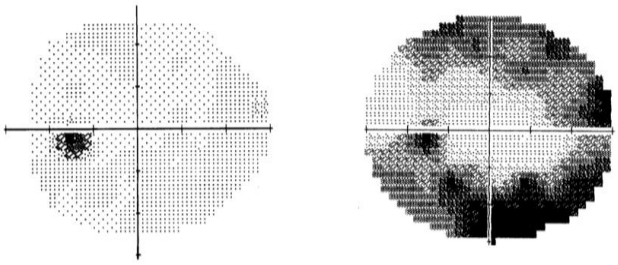


Figure 31: Visual Field test results, left shows a healthy eye without vision loss and right shows a glaucomatous eye with darker grey and black areas representing loss in vision (*the optic disc appears black in both fields since there is no vision there, which is normal*)

For the purposes of this chapter, the data have been compiled into average values according to their relationship with one of six nerve fibre bundles using the mappings shown in [138]. We also investigate data from Heidelberg Retinal Tomography (HRT) [139], which entails creating photographs of the retina to determine specific metrics (such as the neuro-retinal rim area) connected with the three-dimensional form of the optic nerve head. There are a total of six regions of the retina included in the calculations: the nasal (n), nasal inferior (ni), and nasal superior (ns) regions, as well as the temporal (t), temporal inferior (ti), and temporal superior (ts) regions.

6.4. Experiments

Initially, a high-level exploratory analysis is conducted on the HRT and VF dataset to visualise if the datasets show any distinctions between healthy and glaucoma patients. Once preliminary trends have been identified, the HRT and VF datasets will be combined to test if the identified trajectories capture the interplay between the two forms of data

during the glaucoma progression. Datasets of HRT and VF was collected from a study of about 162 participants [139], and a pre-defined procedure was used to categorise each patient as either healthy or glaucomatous through their AGIS score. The threshold programme single-field test STATPAC-2 analysis total deviation printout is used to calculate the AGIS visual field defect score, which is based on the number and depth of clusters of adjacent depressed test sites in the upper and lower hemifields as well as in the nasal area [140]. While this could potentially bias towards the VF data when constructing the pseudo time series, it is not expected to affect the final diseased stages because the relabelling algorithm learns the states from the beginning [141]. Finally, the novel TDA-PTS and CBPTS algorithms, which were explored in chapters 4 and 5 will be applied to a different large VF glaucoma dataset with 54698 entries treated as separate patients to allow for a more robust learning. The AGIS scores were categorised into 4 classes with 0 representing no severity of field defect, 1-5 being mild, 6-11 being moderate and 12-20 being severe. This categorisation of the AGIS scores will allow for progression trajectories to be constructed on the data topology.

6.5. Results

6.5.1. Exploratory Box Plots

Initially 2 sets of 6 box plots were generated for the HRT dataset showing the distribution of healthy, and glaucomatous patient data split into the 6 regions of the retina. The values have been standardised to have a standard deviation and mean of 1 and 0 respectively so that it eliminates one of the large data skewing the results.

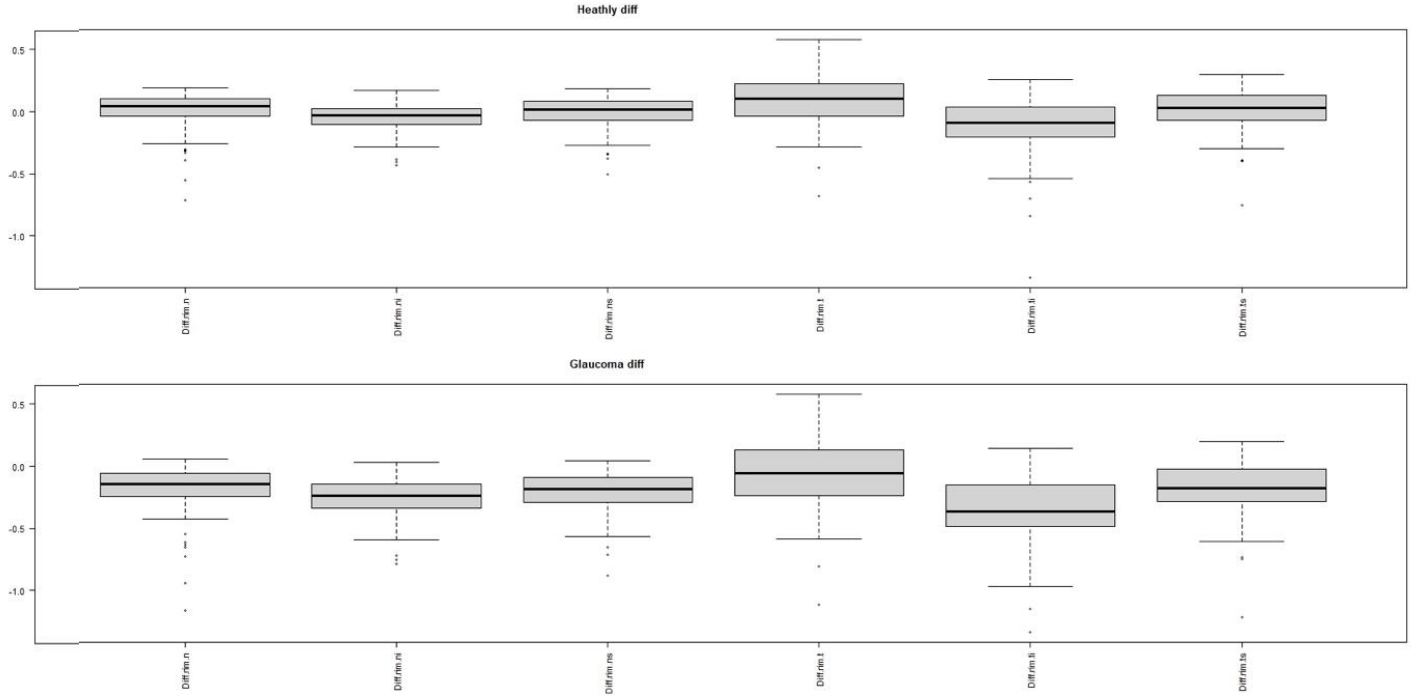


Figure 32: Boxplots showing the variations in the rim narrowing of the 6 regions of the retina for healthy (top) and glaucomayous (bottom) patients

Figure 32 illustrates the diff_rim of the 6 regions of retina, which represents the rim narrowing regions. It shows that the glaucomatous patient data has a wider interquartile range and a lower median compared to the healthy patient data. This reinforces that the narrowing of the rim contributes to the prognosis and progression of glaucoma. Subsequently the mean VF data points are used to construct further exploratory box plots, shown in Figure 33. It is evident from the VF points box plots that glaucomatous patients have a lower sensitivity, which again reinforces the existing knowledge.

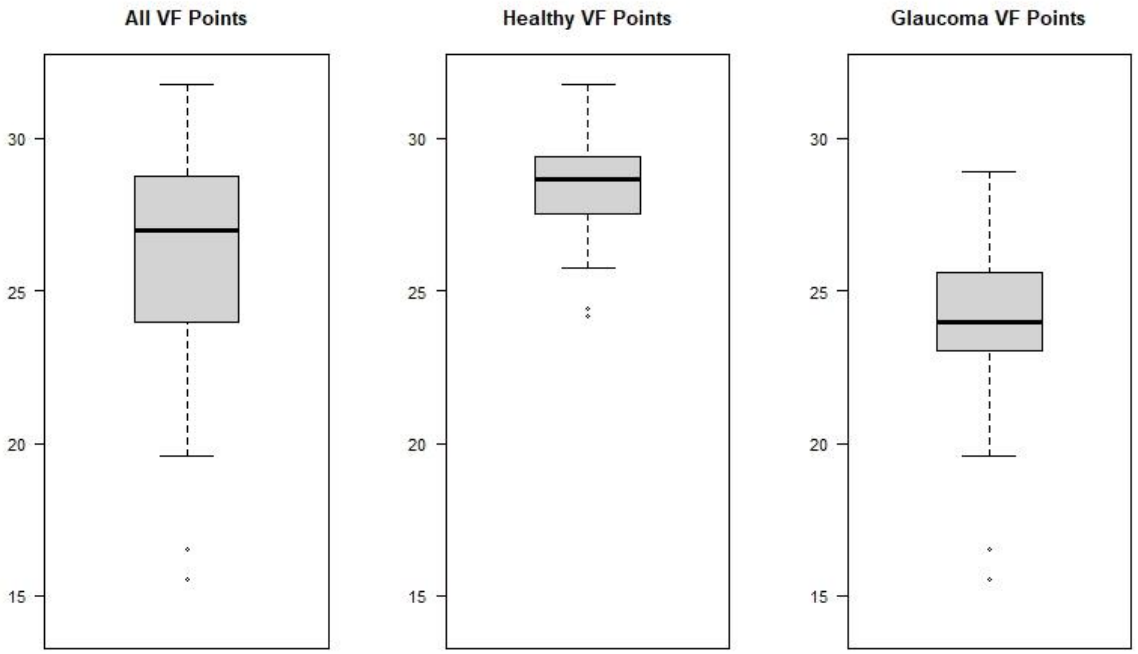


Figure 33: Boxplots showing the variation in Visual Field (VF) points of all patients within the study (left), healthy (middle) and glaucomatous (right) patients.

6.5.2. Glaucoma PTS Trajectories and Transition Analysis

Data points from the glaucoma dataset has been plotted against the first two components so that exploratory sample PTS trajectories can be constructed from a healthy region to a diseased region.

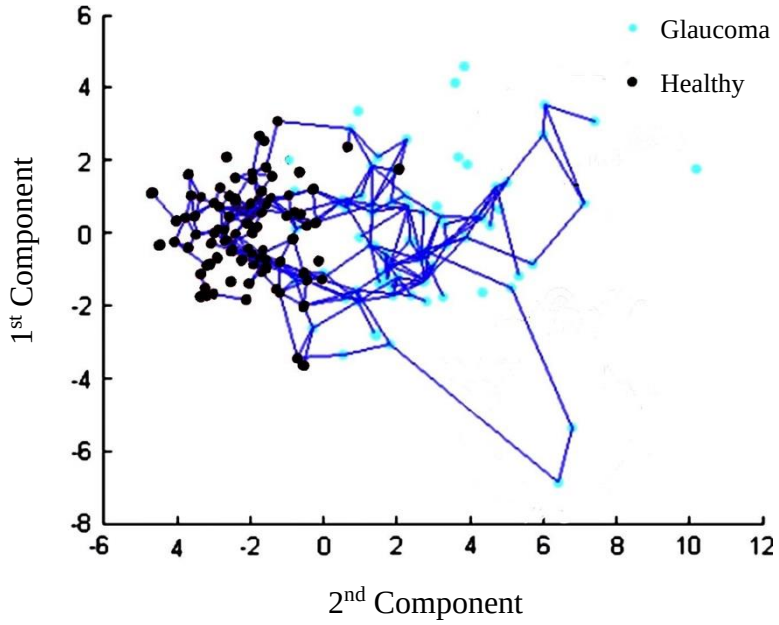


Figure 34: Trajectories showing progression of healthy to diseased glaucoma states the combined HRT and VF data.

Figure 34 identifies two clear regions of diseased end states, which are in the top right and left areas of the graph. Using a relabelling approach to determine sequence transitions in a medical setting, this part of the investigation seeks to validate this, even though it is not immediately apparent.

Now a glaucoma state transition diagram is constructed along with a transition matrix, which is obtained from the HMM.

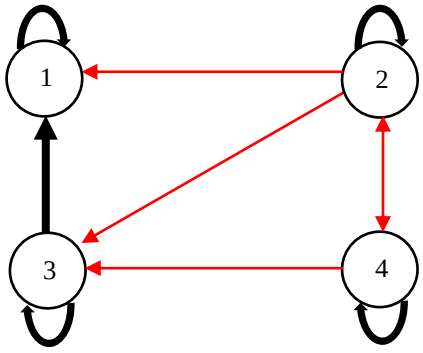


Figure 35: Glaucoma Data State Transition Diagram (black line p>0.15) (red line p<0.15)

Table 5: Visual Field (VF) State Transition Matrix

	State 1	State 2	State 3	State 4
State 1	0.97	0.03	0	0
State 2	0.52	0.74	0.08	0.11
State 3	0.27	0	0.69	0.43
State 4	0	0.07	0.10	0.84

Looking at both the transition diagram and matrix it can been seen that state 4 shows a healthy state, states 1 and 2 display the two diseased end states and finally, state 3 seems to be an intermediate region. We can learn more about these states by comparing the mean values of normal and glaucomatous data to the predicted values of the variables associated with each condition, as well as the clustering values of the variables obtained by k-means clustering shown in Figure 38. The NFB represents the sensitivity of a specifier Nerve Fibre Bundle with the VF. Similar to the diff_rim, these values are also standardised in the same way. Figure 37 shows that state 4 has normal rim width and VF sensitivity, with high NFB sensitivity and low rim-associated variables, like the control in Figure 36, while state 1 has moderate loss of retinal sensitivity (low NFB sensitivities) and marked diffuse rim narrowing (high rim-associated variables), like glaucomatous state shown in the state transition diagram. In accordance with established anatomical connections, this is to be expected.

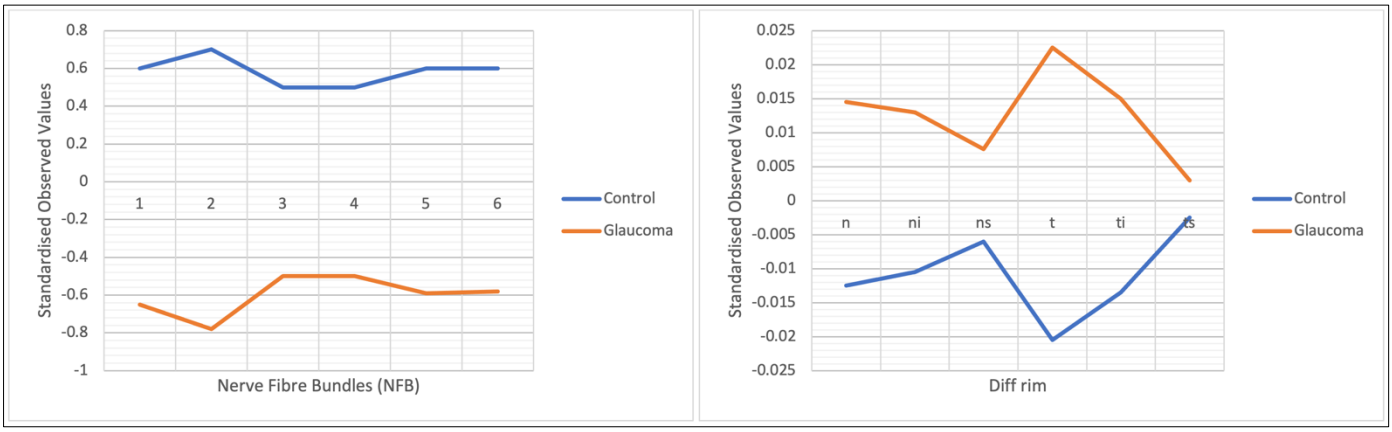


Figure 36: Mean value data for VF (left) and HRT (right) for normal and glaucomatous patient data

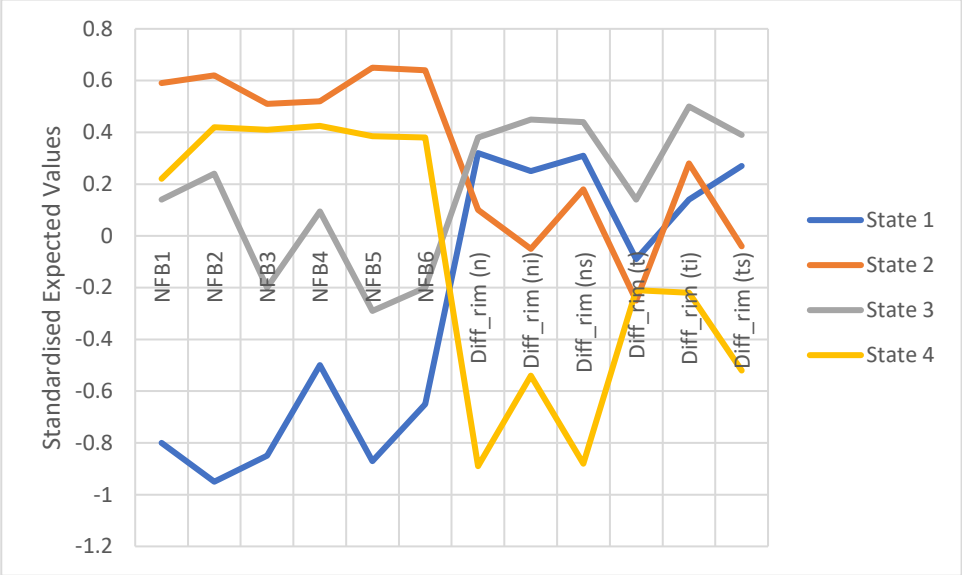


Figure 37: Standardised expected data for VF & HRT from temporal bootstrap for PTS

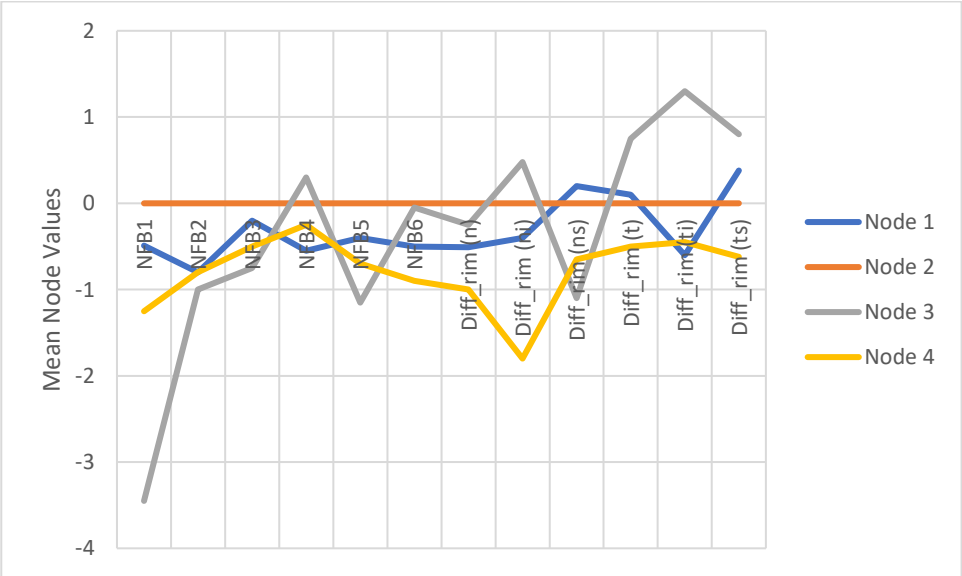


Figure 38: Mean node profiles for VF and HRT from k-means clustering

There is rim narrowing and no loss of retinal sensitivity in state 2, which is a reasonably steady state, but there is rim narrowing and some loss of retinal sensitivity in state 3, which reinforces that this is an intermediate state. It is intriguing that the algorithm

has recognised these variations as indicators of future complete disease progression; these variations are known to occur, and they are presented in the HRT rim data as progression in the field without advancement in the optic disc. These pseudo-temporal models allow for a more effective understanding of glaucoma progression.

6.5.3. TDA-PTS and CBPTS

Initially, the TDA-PTS algorithm is applied to the multi-dimensional VF data to create a topology of the dataset and using the AGIS scores as classes to build trajectories through the data structure. As mentioned before, the AGIS scores were sppit into 4 classes with 0 representing no severity of field defect, 1-5 being mild, 6-11 being moderate and 12-20 being severe.

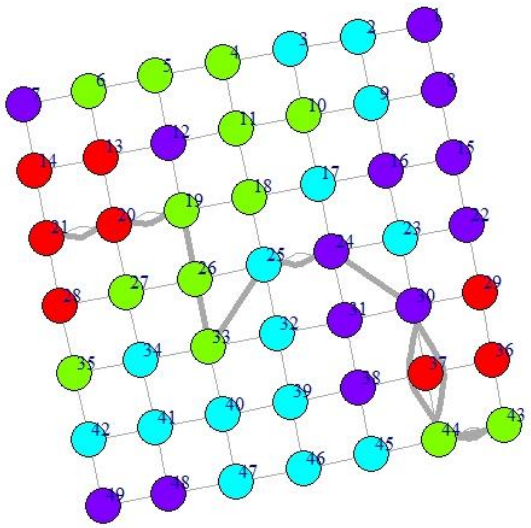


Figure 39: TDA-PTS analysis on glaucoma data to show trajectory path modelling disease progression.

The aim of Figure 39 was to visualise how the trajectories travels through the data nodes to get an understanding of the transitions of this multi-dimensional dataset. This sample trajectory plot initially travels very effectively by starting from a healthy red state

to the next level of severity represented by the green nodes 19, 26 and 33 and then showing progression by going to the next level represented by blue and then purple. However, the trajectory continues to go from what should be an end state purple node to 44 and 43, which are green mild severity field defect node. VF data is not always reliable, there is a learning effect with the test; often the elderly patients doing the VF tests are fatigued and perform variably each visit. This leads to VF data inconsistencies as can be seen in this plot showing improvements in the VF whereas from prior clinical knowledge, we know that glaucoma is a progressive disease of the optic nerve head, and any established reliable and repeatable visual field loss is permanent.

These impossible trajectories prompt the use of the CBPTS to introduce constraints for developing more robust and meaningful outcomes. Firstly, CBPTS is applied without any constraints, followed by constraining to eliminate one backward step such as 2→1, 3→2 and 4→3 and then fully constraining all impossible transitions. Finally, the distribution of the stages of progression is constructed for each level of constraints to see the effectiveness of the transitions.

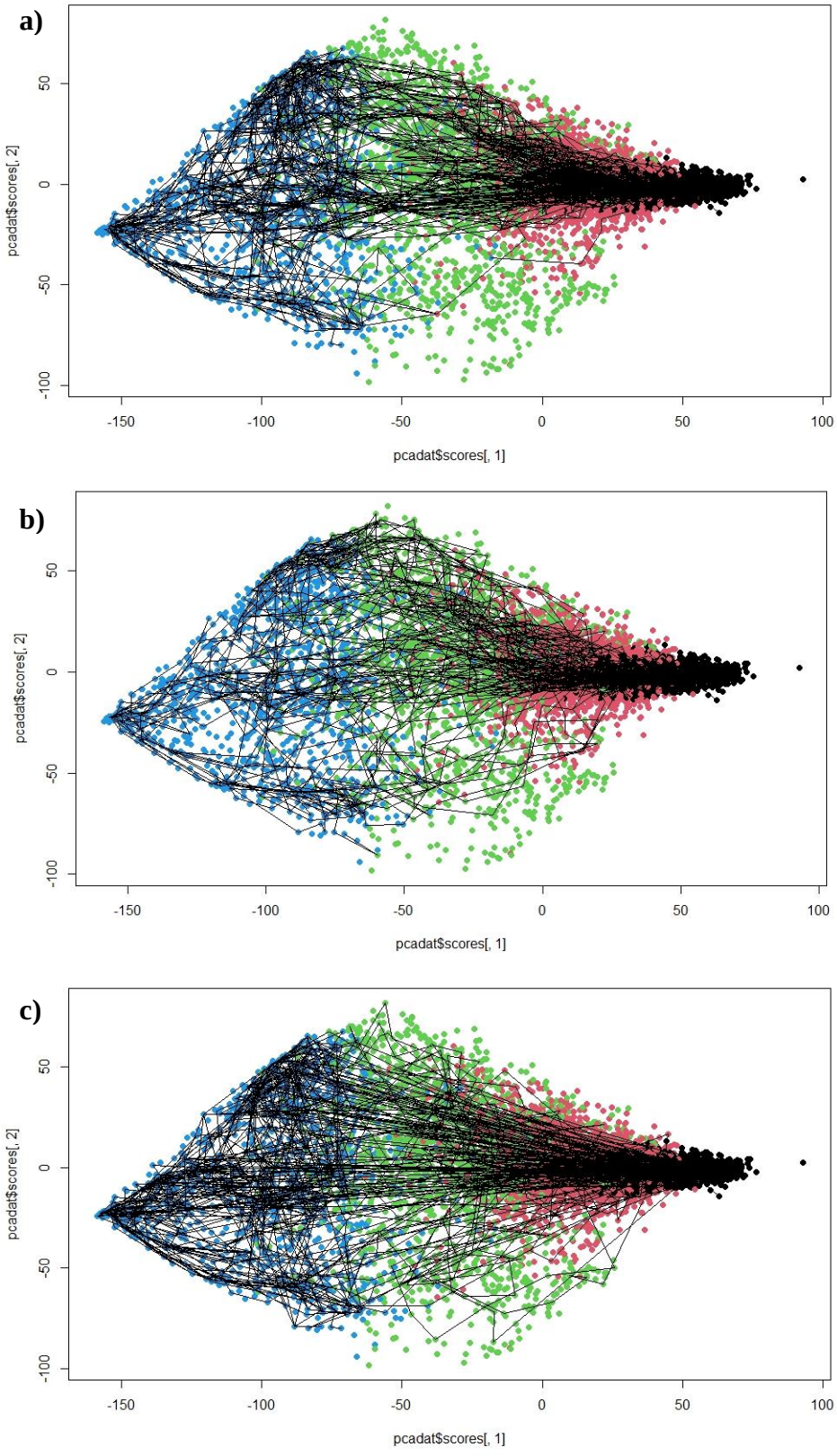


Figure 40: CBPTS trajectory plots modelling disease progression from a healthy black state to a severe diseased blue state going through intermediate stages of red and green. a) no constraints, b) single backward constraint, c) fully backward constraint

Initially, we build a standard PTS with no constraints to see how the trajectories naturally form based on the AGIS scores. Figure 40a shows lots of noisy trajectories, but we can start to see three key directions where the trajectories travel. Subsequently, we have extracted the transition probability matrix to see the transitions statistically.

Table 6: State transition probability matrix extracted from TDA-PTS output

	State 1	State 2	State 3	State 4
State 1	0.855	0.137	0.008	0.000
State 2	0.138	0.615	0.242	0.005
State 3	0.123	0.109	0.548	0.220
State 4	0.031	0.085	0.113	0.771

The transition probability matrix reinforces what we discovered in the TDA-PTS plots where there are numerous impossible trajectories present, which explains the noisy PTS plot. As previously mentioned, two levels of constraints are applied to the initial PTS plot to eliminate these “illegal” trajectories. The CBPTS plot shown in Figure 40b starts to display the three key paths more clearly but the fully constraint CBPTS plot in Figure 40c starts to force the trajectories, which results in an extremely noise over constraint plot. From the staging distributions in Figure 41, we can see that the trajectories generally travel from a healthy state to a more advanced glaucomatous state. The no constraint densities show this trend but also indicates that each of the advanced stages could be its own final diseased stage. In real world scenarios this can be true as when a patient is diagnosed with glaucoma, interventions are taken place to slow or even stop the progression, which can result in mild or moderate visual field loss being the final diseased stage. However, looking even closer at the densities, we can see that stage 4 can precede stage 3 where it shows disease regression. This could be due to VF data inconsistencies. The constraint-based distribution shows the effective progression of the disease from healthy to mild, moderate, and ending in severe diseased state. This is another proof that a strong model can be constructed via constraint analysis.

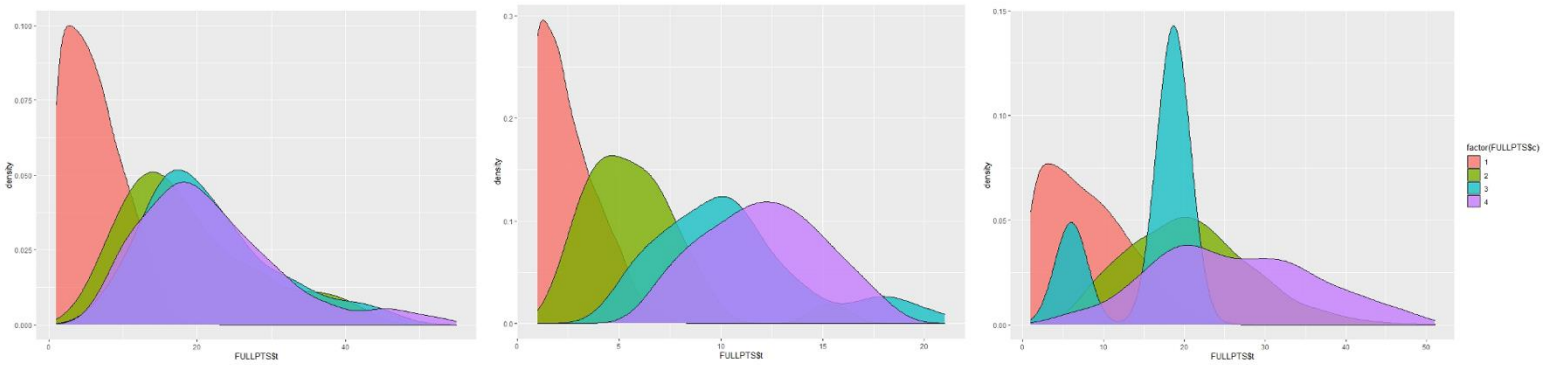


Figure 41: Trajectory density plots extracted from CBPTS, no constraints (left), single backward constraint (middle), fully backward constraint (right) – (red-healthy, green-mild, blue-moderate, purple-severe)

6.6. Summary

There is a lack of knowledge about the true underlying state transitions along a disease process in real-world clinical data. On the other hand, the observed changes can be investigated in a clinical setting. This chapter has provided empirical evidence for the benefits of using relabelling on pseudo time series. It has been examined with real glaucoma data, and its emphasis on identifying important intermediate stages of disease is highlighted. In this chapter we initially looked at two issues: first, how to generate time-series models from cross-sectional data, and second, how to automatically detect distinct disease states and transitions along these trajectories. This analysis of the glaucoma data has allowed us to distinguish between stable states characterised by abnormal VF sensitivity, substantial rim narrowing, and transitory states characterised by moderate rim narrowing and mild loss of retinal sensitivity in the central macula. This is consistent with the current understanding of glaucoma development, in which symptoms may first manifest in the periphery but not the central vision.

VF testing is fully dependent on patient engagement as it measures how far the eye sees in any direction without moving and how sensitive the vision is in different parts of the visual field. If the patient's eye is turned so that it faces the object or the light, only the central part of his or her field of vision will be evaluated. The examiner will explain to the patient exactly where to look so that the test is correct. There is a learning effect

with VF data, and senior patients who do the tests are often tired and have inconsistent results from visit to visit, all of which can lead to inaccurate readings. Hence, the AGIS scores are given ranges within each class of visual field loss severity so that it can accommodate for these inaccurate readings that may be present. It is clear that a constraint approach to using AGIS scores to determine glaucoma progression is effective.

Chapter 7 Conclusions

The findings of this thesis's research are summarised in this chapter. After a brief overview of the major findings, the study's limitations will be discussed. Finally, potential future research directions are presented with the aim of expanding the scope of the methods described in this thesis.

7.1. Conclusion

The study of artificial intelligence, in particular as it pertains to the field of machine learning, is revealing itself to be a path that is radically redefining how we comprehend the complexities of degenerative diseases. Deciphering the complex course of symptoms presented by these illnesses, which are characterised by the gradual but unrelenting degradation of organs or tissues over time, presents a tremendous challenge. The steady worsening of the condition is frequently accompanied with changing rates of deterioration, periods of stability, and fascinating instances of improvement as a result of intervention. In spite of developments in therapies like medicine and surgery that improve quality of life and halt progression of disease, a fundamental transition from health to early onset to severe stages continues to be a distinguishing feature of these conditions.

This thesis, which is founded on its overall aims and objectives, presents an innovative methodology that offers major contributions to the process of deciphering the intricate dynamics of the evolution of disease. The investigation starts off with a thorough examination of Topological Data Analysis (TDA) and Pseudo-Time Series (PTS), which catapults the process of building temporal phenotypes and modelling the course of disease into new dimensions.

The invention of the unique combined TDA-PTS algorithm, which represents a huge leap in the techniques of data analysis, is at the core of this academic endeavour. A paradigm shift in the way we approach the study of disease dynamics is reflected in the

meticulously defined formal structure and pseudo code, which provide a solid platform for novel algorithmic applications.

This research provides a visual pathway into data shapes, the identification of intermediate disease phases, and a sophisticated comprehension of the dynamic landscape of disease progression. A cornerstone of this research is the pioneering building of topological models from cross-sectional data. This methodological innovation extends to the use of Pseudo-Time Series to model and analyse real-world biomedical datasets, which ultimately results in the building of detailed disease progression trajectories.

These trajectories, which depict transitions through the topology of the dataset, provide crucial insights that are evaluated against the actual state of the dataset. This helps to ensure that a high degree of model reliability is achieved. The integration of constraints into Constrained Pseudo-Time Series (CBPTS), which draws upon clinician expertise, signals a paradigm shift in the model refining process. This improvement is clearly demonstrated on both simulated and real-world breast cancer datasets, which bolsters the resilience and reliability of the model.

The application of the created approaches to three separate glaucoma datasets marked the completion of this innovative journey. This application underscored the model's ability to understand and properly depict clinically verified progressive and irreversible diseases. The goals and ambitions laid out in earlier chapters are effectively accomplished by means of this exhaustive investigation, which encompasses everything from a review of the relevant literature to the creation of a methodology and the careful implementation of that technique to a variety of datasets.

The study of artificial intelligence, more specifically in the area of machine learning, offers itself as a paradigm-shifting undertaking with the goal of understanding the intricacies of degenerative diseases. Understanding the complex progression of symptoms presented by these diseases, which are marked by the gradual deterioration of organs or tissues over time, is a substantial challenge due to the complexity of the disease's presentation. The constant increase in intensity frequently manifests itself as different patterns of decline, periods of stability, and intriguing occurrences of augmentation with intervention. These can all be observed during the course of the progression. Even though

there have been significant advances in medical interventions, such as medicine and surgery, that improve the overall well-being of individuals and slow down the progression of diseases, it is essential to recognise that a fundamental transition from a state of good health to the early stages and then eventually to the advanced stages of these diseases continues to be a significant characteristic. This is the case even though there have been notable breakthroughs in these types of medical interventions.

The methodology presented in this thesis is one of a kind, and it makes a substantial contribution to understanding the subtle dynamics of disease progression. This helps the thesis fit with its overarching goals and objectives. An in-depth analysis of Topological Data Analysis (TDA) and Pseudo-Time Series (PTS), which enables the generation of temporal phenotypes and the expansion of disease progression modelling into other areas, is the first step in this investigation.

The presentation of the novel combined TDA-PTS algorithm, which represents an important step forward in methods for data analysis, is at the heart of this academic endeavour. It is also the most important contribution that this work makes. The formation of a solid and dependable foundation for the development of unique algorithmic applications is achieved through the detailed outlining of a formal structure and the implementation of pseudo code. This is a change in the approach that we take in order to comprehend the dynamics of diseases.

Building topological models from cross-sectional data is an essential part of this research because it enables a clearer understanding of the dynamic landscape of disease progression, gives a visual representation of the data shapes, and makes it easier to identify disease phases in between the extremes of the disease spectrum. This innovation in methodology involves employing Pseudo-Time Series in the study of both simulated and real-world biological datasets, which ultimately results in the creation of complex disease progression trajectories.

The trajectories that have been described in this study offer valuable insights into the transitions that have been seen in the topology of the dataset. These insights have been validated by comparing them to the original state of the dataset, which helps to ensure that the model is as accurate as possible. Utilising the knowledge and experience of

clinicians, the incorporation of constraints in Constrained Pseudo-Time Series (CBPTS) constitutes a significant paradigm shift in the approach taken to model refinement. This improvement has been convincingly demonstrated on both simulated and real-world breast cancer datasets, hence confirming the resilience and reliability of the model.

The completion of this transformative process is illustrated by the application of the devised methodologies on three independent glaucoma datasets, which demonstrates the efficacy of the model in grasping and faithfully portraying clinically validated progressive and irreversible diseases. By conducting an exhaustive inquiry that includes a literature review, the development of a methodology, and the exact application of this technique on a variety of datasets, the current study was able to effectively complete the objectives that were outlined in this thesis.

This thesis not only advances scientific grasp of the evolution of disease by combining the discoveries and contributions, but it also propels the field of healthcare towards methods that are more educated and accurate. This research makes a substantial contribution to the ongoing discourse on degenerative diseases by detecting intermediate disease stages and showing trends. This contribution is made possible through the application of sophisticated trajectory analysis, which was utilised in this study. The information that was gained by participating in this forward-thinking endeavour reveals the possibility for future progress, which will ultimately enhance our ability to comprehend, diagnose, and intervene in the course of these complex medical issues.

7.2. Limitations and Further Work

In the first stages of this investigation, the researchers struggled with the difficulty of constructing trajectories using topology without having any prior insights. As a result, the procedure was prone to producing trajectories that were not true to reality. Because of this limitation, a focused analysis was conducted, which is described in Chapter 5. During this inquiry, constraints were applied in order to improve the robustness of trajectories. On the other hand, it is of the utmost importance to acknowledge the possibility of over-constraining, which is a situation in which the pseudo-time may be pushed to follow

particular trajectories determined by the sample size that is utilised in the creation of the trajectory. Taking this into mind highlights the importance of taking a nuanced approach to the application of constraints.

The utilisation of the temporal behaviour that might be extracted from cross-sectional data was the primary emphasis of the research. It is a disadvantage of this approach since the model has not been properly evaluated on longitudinal data, which necessarily possesses temporal characteristics. Despite the fact that this approach produced useful discoveries, it is a restriction. The combination of longitudinal and cross-sectional data could be beneficial, since it would allow for the modelling of a diverse population that includes samples from all stages of the disease. By including this integration, the true temporal features of disease processes might be properly encoded, hence making the technique more applicable to situations that occur in the real world.

Another area of investigation that could be pursued for the purpose of further validating and expanding the scope of this technique is the application of the methodology to other clinical datasets that contain more precise staging data. With this enlarged analysis, we want to gain a better understanding of how different patient characteristics change as the disease advances. The utilisation of multi-class data, which enables the investigation of the influence that a wide variety of variables have on the analysis, is a method that has proven to be effective during this process. Not only does this increase the applicability of the approach in a wider variety of clinical settings, but it also broadens the breadth of the methodology.

Furthermore, one potential direction for future research is to investigate the performance of the approach in situations where external factors, such as comorbidities or changes in lifestyle, can have an impact on the advancement of the disease. If such real-world complexities were incorporated into the analysis, it could be possible to obtain a more nuanced understanding of the ways in which various variables interact with one another and influence the progression of diseases. Subsequently, introducing more data will result in an increased amount of missing data and information. The absence of data, known as missing data, significantly influences the reliability of TDA and PTS. In TDA, missing values disrupt the continuity of the dataset, potentially leading to distorted

topological structures and inaccurate representations. In PTS, where temporal trajectories are constructed, missing values can distort the temporal ordering and result in biased or incomplete representations of disease progression. Omitting missing values for analysis is a common practice. In this research, the base test for the novel approach is initially based on simulated data which has no missing values. Also, the real-world datasets used are pre-processed to deal with any missing datapoints, which is usually by omitting the entries. Omitting missing values has drawbacks as it could lead to information loss, introducing bias, reducing the sample size, and can impact model performance. Best practices include employing imputation techniques, transparently reporting missing data handling methods, conducting sensitivity analyses, and consulting domain experts for valuable insights. Handling missing data is crucial for preserving the integrity of analyses and ensuring the accuracy and applicability of results in complex datasets.

In conclusion, despite the fact that this research has made substantial progress in resolving its initial limitations and demonstrating the application of constraints for robust trajectory building, there are still paths that need to be further refined and explored. In the larger context of disease progression modelling, the incorporation of longitudinal data, the investigation of a wide range of clinical datasets, and the inclusion of external factors are all aspects that have the potential to enhance the validity, application, and potential impact of the methodology.

Bibliography

[1] D. M. Goodman, C. Lynm, and E. H. Livingston, ‘Genomic Medicine’, *JAMA*, vol. 309, no. 14, pp. 1544–1544, Apr. 2013, doi: 10.1001/JAMA.2013.1927.

[2] G. Alterovitz, C. Tuthill, I. Rios, K. Modelska, and S. Sonis, ‘Personalized medicine for mucositis: Bayesian networks identify unique gene clusters which predict the response to gamma-D-glutamyl-L-tryptophan (SCV-07) for the attenuation of chemoradiation-induced oral mucositis’, 2011, doi: 10.1016/j.oraloncology.2011.07.006.

[3] M. López-Lázaro, ‘The stem cell division theory of cancer’, *Crit Rev Oncol Hematol*, vol. 123, pp. 95–113, Mar. 2018, doi: 10.1016/J.CRITREVONC.2018.01.010.

[4] B. N. Bimber, R. Ramakrishnan, R. Cervera-Juanes, R. Madhira, S. M. Peterson, R. B. Norgren Jr and B. Ferguson, ‘Whole genome sequencing predicts novel human disease models in rhesus macaques’, *Genomics*, vol. 109, no. 3–4, pp. 214–220, Jul. 2017, doi: 10.1016/J.YGENO.2017.04.001.

[5] A. C. Tan and D. Gilbert, ‘An empirical comparison of supervised machine learning techniques in bioinformatics’, 2003, doi: 10.5555/820189.820218.

[6] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, 2nd ed. MIT Press, 2001.

[7] J. Shavlik, L. Hunter, and D. Searls, ‘Introduction’, *Machine Learning 1995 21:1*, vol. 21, no. 1, pp. 5–9, 1995, doi: 10.1023/A:1022661429642.

[8] E. Alpaydin, *Introduction to Machine Learning*, 3rd ed. MIT Press, 2014.

[9] T. Mitchell, *Machine learning*. McGraw-Hill Education, 1997. Accessed: Sep. 07, 2022. [Online]. Available: https://profs.info.uaic.ro/~ciortuz/SLIDES/2017s/ml0.pdf

[10] X. Zhou, S. Liu, E. S. Kim, R. S. Herbst, and J. J. J. Lee, ‘Bayesian adaptive design for targeted therapy development in lung cancer - A step toward personalized medicine’, *Clinical Trials*, vol. 5, no. 3, pp. 181–193, Jun. 2008, doi: 10.1177/1740774508091815.

[11] J. Jack Lee and C. T. Chu, ‘Bayesian clinical trials in action’, *Stat Med*, vol. 31, no. 25, pp. 2955–2972, Nov. 2012, doi: 10.1002/SIM.5404.

[12] A. Dagliati, L. Sacchi, A. Zambelli, V. Tibollo, L. Pavesi, J. H. Holmes and R. Bellazzi, ‘Temporal electronic phenotyping by mining careflows of breast cancer patients’, *J Biomed Inform*, vol. 66, pp. 136–147, Feb. 2017, doi: 10.1016/J.JBI.2016.12.012.

[13] G. Hripcsak and D. J. Albers, ‘Next-generation phenotyping of electronic health records’, *J Am Med Inform Assoc*, vol. 20, no. 1, pp. 117–121, 2013, doi: 10.1136/AMIAJNL-2012-001145.

[14] M. J. Feio, C. Viana-Ferreira, and C. Costa, ‘Combining multiple machine learning algorithms to predict taxa under reference conditions for streams bioassessment’, *River Res Appl*, vol. 30, no. 9, pp. 1157–1165, Nov. 2014, doi: 10.1002/RRA.2707.

[15] B. Lantz, *Machine Learning with R*, 2nd ed. Packt Publishing Ltd, 2015.

[16] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafe, A. Perez and V. Robles, ‘Machine learning in bioinformatics’, *Brief Bioinform*, vol. 7, no. 1, pp. 86–112, Mar. 2006, doi: 10.1093/BIB/BBK007.

[17] R. Castelo and A. Siebes, ‘Priors on network structures. Biasing the search for Bayesian networks’, *International Journal of Approximate Reasoning*, vol. 24, no. 1, pp. 39–57, Apr. 2000, doi: 10.1016/S0888-613X(99)00041-9.

[18] D. K. Slonim, ‘From patterns to pathways: gene expression data analysis comes of age’, *Nat Genet*, vol. 32 Suppl, no. 4S, pp. 502–508, 2002, doi: 10.1038/NG1033.

[19] K. P. Murphy, ‘Dynamic Bayesian networks: Representation, inference and learning’, 2002.

[20] E. W. Watt and A. A. T. Bui, ‘Evaluation of a Dynamic Bayesian Belief Network to Predict Osteoarthritic Knee Pain Using Data from the Osteoarthritis Initiative’, *AMIA Annual Symposium Proceedings*, vol. 2008, p. 788, 2008, Accessed: Sep. 07, 2022. [Online]. Available: /pmc/articles/PMC2656041/

[21] M. A. J. van Gerven, B. G. Taal, and P. J. F. Lucas, ‘Dynamic Bayesian networks as prognostic models for clinical patient management’, *J Biomed Inform*, vol. 41, no. 4, pp. 515–529, Aug. 2008, doi: 10.1016/J.JBI.2008.01.006.

[22] C. Rose, C. Smaili, and F. Charpillet, ‘A dynamic Bayesian network for handling uncertainty in a decision support system adapted to the monitoring of patients treated by hemodialysis’, *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, vol. 2005, pp. 594–598, 2005, doi: 10.1109/ICTAI.2005.7.

[23] S. Ceccon, D. Garway-Heath, D. Crabb, and A. Tucker, ‘Non-stationary Clustering Bayesian Networks for glaucoma’.

[24] L. Zhang, D. Samaras, N. Alia-klein, N. Volkow, and R. Goldstein, ‘Modeling Neuronal Interactivity using Dynamic Bayesian Networks’, *Adv Neural Inf Process Syst*, vol. 18, 2005.

[25] T. Charitos, L. C. van der Gaag, S. Visscher, K. A. M. Schurink, and P. J. F. Lucas, ‘A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients’, *Expert Syst Appl*, vol. 36, no. 2, pp. 1249–1258, Mar. 2009, doi: 10.1016/J.ESWA.2007.11.065.

[26] M. J. Zvelebil and J. O. Baum, *Understanding Bioinformatics*. Garland Science, 2008. Accessed: Sep. 07, 2022. [Online]. Available: https://books.google.com/books/about/Understanding_Bioinformatics.html?id=Li0WBAAAQBAJ

[27] H. He and E. A. Garcia, ‘Learning from imbalanced data’, *IEEE Trans Knowl Data Eng*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[28] A. Mozaffari, M. Emami, and A. Fathi, ‘A comprehensive investigation into the performance, robustness, scalability and convergence of chaos-enhanced evolutionary algorithms with boundary constraints’, *Artif Intell Rev*, vol. 52, no. 4, pp. 2319–2380, Dec. 2019, doi: 10.1007/S10462-018-9616-4/FIGURES/9.

[29] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. E. Mohamed, and H. Arshad, ‘State-of-the-art in artificial neural network applications: A survey’, *Heliyon*, vol. 4, no. 11, p. e00938, Nov. 2018, doi: 10.1016/J.HELIYON.2018.E00938.

[30] S. C. Odewahn, E. B. Stockwell, R. L. Pennington, R. M. Humphreys and W. A. Zumach, ‘Automated Star/Galaxy Discrimination With Neural Networks’, *AJ*, vol. 103, no. 1, p. 318, Jan. 1992, doi: 10.1086/116063.

[31] S. B. Nagl, ‘Can correlated mutations in protein domain families be used for protein design?’, *Brief Bioinform*, vol. 2, no. 3, pp. 279–288, Sep. 2001, doi: 10.1093/BIB/2.3.279.

[32] P. J. Lisboa and A. F. G. Taktak, ‘The use of artificial neural networks in decision support in cancer: a systematic review’, *Neural Netw*, vol. 19, no. 4, pp. 408–415, May 2006, doi: 10.1016/J.NEUNET.2005.10.007.

[33] L. Budagyan and R. Abagyan, ‘Weighted quality estimates in machine learning’, *Bioinformatics*, vol. 22, no. 21, pp. 2597–2603, Nov. 2006, doi: 10.1093/BIOINFORMATICS/BTL458.

[34] M. Frize, C. M. Ennett, M. Stevenson, and H. C. E. Trigg, ‘Clinical decision support systems for intensive care units: using artificial neural networks’, *Med Eng Phys*, vol. 23, no. 3, pp. 217–225, 2001, doi: 10.1016/S1350-4533(01)00041-8.

[35] L. Satish and B. I. Gururaj, ‘Partial discharge pattern classification using multilayer neural networks’, *IEE Proceedings A Science, Measurement and Technology*, vol. 140, no. 4, p. 323, 1993, doi: 10.1049/IP-A-3.1993.0049.

[36] J. M. Jerez-Aragonés, J. A. Gómez-Ruiz, G. Ramos-Jiménez, J. Muñoz-Pérez, and E. Alba-Conejo, ‘A combined neural network and decision trees model for prognosis of breast cancer relapse’, *Artif Intell Med*, vol. 27, no. 1, pp. 45–63, 2003, doi: 10.1016/S0933-3657(02)00086-6.

[37] D. B. Fogel, E. C. Wasson, and E. M. Boughton, ‘Evolving neural networks for detecting breast cancer’, *Cancer Lett*, vol. 96, no. 1, pp. 49–53, Sep. 1995, doi: 10.1016/0304-3835(95)03916-K.

[38] Tuba. Kiyan and T. Yıldırım, ‘Breast Cancer Diagnosis using Statistical Neural Networks’, *IU-Journal of Electrical & Electronics Engineering*, vol. 4, pp. 1149–1153, 2004.

[39] C. te Chen, W.-L. Lin, T.-S. Kuo, and C.-Y. Wang, ‘Adaptive control of arterial blood pressure with a learning controller based on multilayer neural networks’, *IEEE Trans Biomed Eng*, vol. 44, no. 7, pp. 601–609, Jul. 1997, doi: 10.1109/10.594901.

[40] H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, ‘A multilayer perceptron-based medical decision support system for heart disease diagnosis’, *Expert Syst Appl*, vol. 30, no. 2, pp. 272–281, Feb. 2006, doi: 10.1016/J.ESWA.2005.07.022.

[41] Y. C. Li, L. Liu, W. T. Chiu, and W. S. Jian, ‘Neural network modeling for surgical decisions on traumatic brain injury patients’, *Int J Med Inform*, vol. 57, no. 1, pp. 1–9, Jan. 2000, doi: 10.1016/S1386-5056(99)00054-4.

[42] M. L. Vaughn, S. J. Cavill, S. J. Taylor, M. A. Foy, and A. J. B. Fogg, ‘Direct explanations and knowledge extraction from a multilayer perceptron network that performs low back pain classification’, *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, vol. 1778, pp. 270–285, 2000, doi: 10.1007/10719871_19/COVER.

[43] R. Bellazzi and B. Zupan, ‘Predictive data mining in clinical medicine: current issues and guidelines’, *Int J Med Inform*, vol. 77, no. 2, pp. 81–97, Feb. 2008, doi: 10.1016/J.IJMEDINF.2006.11.006.

[44] L. Breiman, ‘Random Forests’, *Machine Learning 2001 45:1*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[45] W. Koehrsen, ‘An Implementation and Explanation of the Random Forest in Python’, *Towards Data Science*, 2018. https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76 (accessed Sep. 08, 2022).

[46] J. A. Hartigan, *Clustering Algorithms*, 99th ed. USA: John Wiley & Sons, Inc., 1975.

[47] B. Mirkin, *Clustering for Data Mining : A Data Recovery Approach*, 1st ed. Chapman and Hall/CRC, 2005. doi: 10.1201/9781420034912.

[48] 'Gaussian mixture models', *scikit-learn*. https://scikit-learn.org/stable/modules/mixture.html (accessed Sep. 08, 2022).

[49] M. Ouyang, W. J. Welsh, and P. Georgopoulos, 'Gaussian mixture clustering and imputation of microarray data', *Bioinformatics*, vol. 20, no. 6, pp. 917–923, Apr. 2004, doi: 10.1093/bioinformatics/bth007.

[50] X. Y. Wang, J. Garibaldi, and T. Ozen, 'Application of The Fuzzy C-Means Clustering Method on the Analysis of non Pre-processed FTIR Data for Cancer Diagnosis', in *in the Proceedings of the 8th Australian and New Zealand Conference on Intelligent Information Systems*, 2003, p. 238.

[51] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, 'A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data', *IEEE Trans Med Imaging*, vol. 21, no. 3, pp. 193–199, Mar. 2002, doi: 10.1109/42.996338.

[52] J. C. Dunn, 'A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters', *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, Jan. 1973, doi: 10.1080/01969727308546046.

[53] E. Martinez-Zeron, M. A. Aceves-Fernandez, E. Gorrostieta-Hurtado, A. Sotomayor-Olmedo and J. M. Ramos-Arreguin, 'Method to Improve Airborne Pollution Forecasting by Using Ant Colony Optimization and Neuro-Fuzzy Algorithms', *Int J Intell Sci*, vol. 4, no. 4, pp. 81–90, Sep. 2014, doi: 10.4236/IJIS.2014.44010.

[54] J. Wang, D. K. Schreiber, N. Bailey, P. Hosemann, and M. B. Toloczko, 'The Application of the OPTICS Algorithm to Cluster Analysis in Atom Probe Tomography Data', *Microscopy and Microanalysis*, vol. 25, no. 2, pp. 338–348, Apr. 2019, doi: 10.1017/S1431927618015386.

[55] H. P. Kriegel, P. Kröger, J. Sander, and A. Zimek, 'Density-based clustering', *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 1, no. 3, pp. 231–240, May 2011, doi: 10.1002/WIDM.30.

[56] S. B. Kotsiantis, 'Supervised Machine Learning: A Review of Classification Techniques', *Informatica*, vol. 31, pp. 249–268, 2007.

[57] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, 'A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science', in *Unsupervised and Semi-Supervised Learning*, Springer, Cham, 2020, pp. 3–21. doi: 10.1007/978-3-030-22475-2_1.

[58] M. W. Libbrecht and W. S. Noble, ‘Machine learning applications in genetics and genomics’, *Nat Rev Genet*, vol. 16, no. 6, pp. 321–332, May 2015, doi: 10.1038/NRG3920.

[59] A. Ng, ‘Supervised learning’, *Mach Learn*, pp. 1–30, 2012, Accessed: Sep. 20, 2022. [Online]. Available: https://see.stanford.edu/materials/aimlcs229/cs229-notes1.pdf

[60] T. Hofmann, ‘Unsupervised Learning by Probabilistic Latent Semantic Analysis’, *Machine Learning 2001 42:1*, vol. 42, no. 1, pp. 177–196, 2001, doi: 10.1023/A:1007617005950.

[61] L. Li, W. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger and J. T. Dudley, ‘Identification of type 2 diabetes subgroups through topological analysis of patient similarity’, *Sci Transl Med*, vol. 7, no. 311, Oct. 2015, doi: 10.1126/SCITRANSLMED.AAA9364.

[62] J. L. Nielson, J. Paquette, A. W. Liu, C. F. Guandique, C. A. Tovar, T. Inoue, K. Irvine, J. C. Gensel, J. Kloke, T. C. Petrossian, P. Y. Lum, G. E. Carlsson, G. T. Manley, W. Young, M. S. Beattie, J. C. Bresnahan and A. R. Ferguson, ‘Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury’, *Nature Communications 2015 6:1*, vol. 6, no. 1, pp. 1–12, Oct. 2015, doi: 10.1038/ncomms9581.

[63] B. Y. Torres, J. H. M. Oliveira, A. Thomas Tate, P. Rath, K. Cumnock, and D. S. Schneider, ‘Tracking Resilience to Infections by Mapping Disease Space’, *PLoS Biol*, vol. 14, no. 4, p. e1002436, Apr. 2016, doi: 10.1371/JOURNAL.PBIO.1002436.

[64] G. Singh, F. Mémoli, and G. Carlsson, ‘Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition’, in *Eurographics Symposium on Point-Based Graphics*, 2007, pp. 91–100.

[65] M. Nicolau, A. J. Levine, and G. Carlsson, ‘Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival’, *Proc Natl Acad Sci U S A*, vol. 108, no. 17, pp. 7265–7270, Apr. 2011, doi: 10.1073/PNAS.1102826108/SUPPL_FILE/PNAS.201102826SI.PDF.

[66] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson and G. Carlsson, ‘Extracting insights from the shape of complex data using topology’, *Scientific Reports 2013 3:1*, vol. 3, no. 1, pp. 1–8, Feb. 2013, doi: 10.1038/srep01236.

[67] P. Torres-Tramón, H. Hromic, and B. R. Heravi, ‘Topic detection in twitter using topology data analysis’, in *ICWE 2015: Current Trends in Web Engineering. 9396*, 2015, vol. 9396, pp. 186–197. doi: 10.1007/978-3-319-24800-4_16/COVER.

[68] S. Gholizadeh, A. Seyeditabari, and W. Zadrozny, ‘Topological Signature of 19th Century Novelists: Persistent Homology in Text Mining’, *Big Data and Cognitive Computing*, vol. 2, no. 4, p. 33, Oct. 2018, doi: 10.3390/BDCC2040033.

[69] D. Nilsson and A. Ekgren, ‘Topology and Word Spaces’, KTH Computer Science and Communication, Stockholm, 2013.

[70] X. Zhu, ‘Persistent Homology: An Introduction and a New Text Representation for Natural Language Processing’, in *IJCAI International Joint Conference on Artificial Intelligence*, 2013, pp. 1953–1959.

[71] M. E. Sardiu, J. M. Gilmore, B. Groppe, L. Florens, and M. P. Washburn, ‘Identification of Topological Network Modules in Perturbed Protein Interaction Networks’, *Sci Rep*, vol. 7, no. 1, pp. 1–13, Mar. 2017, doi: 10.1038/srep43845.

[72] A. H. Rizvi, P. G. Camara, E. K. Kandror, T. J. Roberts, I. Schieren, T. Maniatis and R. Rabadan, ‘Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development’, *Nat Biotechnol*, vol. 35, no. 6, pp. 551–560, Jun. 2017, doi: 10.1038/nbt.3854.

[73] D. Romano, M. Nicolau, E. Quintin, P. K. Mazaika, A. A. Lightbody, H. C. Hazlett, J. Piven, G. Carlsson and A. L. Reiss, ‘Topological methods reveal high and low functioning neuro-phenotypes within fragile X syndrome’, *Hum Brain Mapp*, vol. 35, no. 9, pp. 4904–4915, 2014, doi: 10.1002/HBM.22521.

[74] A. Dagliati, N. Geifman, N. Peek, J. H. Holmes, L. Sacchi, S. E. Sajjadi and A. Tucker, ‘Inferring temporal phenotypes with topological data analysis and pseudo time-series’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Jun. 2019, vol. 11526 LNAI, pp. 399–409. doi: 10.1007/978-3-030-21642-9_50.

[75] I. T. Jollife and J. Cadima, ‘Principal component analysis: a review and recent developments’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, Apr. 2016, doi: 10.1098/RSTA.2015.0202.

[76] J. Kruskal and M. Wish, *Multidimensional Scaling*. SAGE Publications, Inc., 1978. doi: 10.4135/9781412985130.

[77] A. Buja, D. F. Swayne, M. L. Littman, N. Dean, H. Hofmann, and L. Chen, ‘Data visualization with multidimensional scaling’, *Journal of Computational and Graphical Statistics*, vol. 17, no. 2, pp. 444–472, Jun. 2008, doi: 10.1198/106186008X318440.

[78] L. Wasserman, ‘Topological Data Analysis’, *Annu Rev Stat Appl*, vol. 5, pp. 501–532, Mar. 2018, doi: 10.1146/ANNUREV-STATISTICS-031017-100045.

[79] Z. Liu, Y. Lin, and M. Sun, ‘Network Representation’, *Representation Learning for Natural Language Processing*, pp. 217–283, 2020, doi: 10.1007/978-981-15-5573-2_8.

[80] A. Mohan and K. v. Pramod, ‘Network representation learning: models, methods and applications’, *SN Appl Sci*, vol. 1, no. 9, pp. 1–23, Sep. 2019, doi: 10.1007/S42452-019-1044-9/FIGURES/10.

[81] G. Leban, B. Zupan, G. Vidmar, and I. Bratko, ‘VizRank: Data visualization guided by machine learning’, *Data Min Knowl Discov*, vol. 13, no. 2, pp. 119–136, Sep. 2006, doi: 10.1007/S10618-005-0031-5/FIGURES/6.

[82] W. S. Cleveland and R. McGill, ‘The many faces of a scatterplot’, *J Am Stat Assoc*, vol. 79, no. 388, pp. 807–822, 1984, doi: 10.1080/01621459.1984.10477098.

[83] D. A. Keim and H. P. Kriegel, ‘Visualization techniques for mining large databases: A comparison’, *IEEE Trans Knowl Data Eng*, vol. 8, no. 6, pp. 923–938, 1996, doi: 10.1109/69.553159.

[84] M. Teliti, G. Cogni, L. Sacchi, A. Dagliati, S. Marini, V. Tibollo, P. De Cata, R. Bellazzi and L. Chiovato, ‘Risk factors for the development of micro-vascular complications of type 2 diabetes in a single-centre cohort of patients’, *Diab Vasc Dis Res*, vol. 15, no. 5, pp. 424–432, Sep. 2018, doi: 10.1177/1479164118780808/ASSET/IMAGES/LARGE/10.1177_1479164118780808-FIG1.JPEG.

[85] K. R. Campbell and C. Yau, ‘Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data’, *Nat Commun*, vol. 9, no. 1, pp. 1–12, Dec. 2018, doi: 10.1038/s41467-018-04696-6.

[86] A. Tucker and D. Garway-Heath, ‘The pseudotemporal bootstrap for predicting glaucoma from cross-sectional visual field data’, in *IEEE Transactions on Information Technology in Biomedicine*, Jan. 2010, vol. 14, no. 1, pp. 79–85. doi: 10.1109/TITB.2009.2023319.

[87] M. S. Setia, ‘Methodology Series Module 3: Cross-sectional Studies’, *Indian J Dermatol*, vol. 61, no. 3, p. 261, May 2016, doi: 10.4103/0019-5154.182410.

[88] At Work, ‘Cross-sectional vs. longitudinal studies’, 2015. https://www.iwh.on.ca/what-researchers-mean-by/cross-sectional-vs-longitudinal-studies (accessed Sep. 21, 2022).

[89] C. J. Mann, ‘Observational research methods. Research design II: cohort, cross sectional, and case-control studies’, *Emergency Medicine Journal*, vol. 20, no. 1, pp. 54–60, Jan. 2003, doi: 10.1136/EMJ.20.1.54.

[90] L. Janzon, S. E. Lindell, E. Trell, and P. Larme, ‘Smoking habits and carboxyhaemoglobin. A cross-sectional study of an urban population of middle-aged men’, *J Epidemiol Community Health (1978)*, vol. 35, no. 4, pp. 271–273, 1981, doi: 10.1136/JECH.35.4.271.

[91] P. Diggle, *Analysis of longitudinal data.*, 2nd ed., vol. 25. Oxford University Press, 2002.

[92] P. S. Albert, ‘Longitudinal Data Analysis (Repeated Measures) in Clinical Trials’, *Stat Med*, vol. 18, no. 13, pp. 1707–1732, 1999, doi: 10.1002/0470023678.CH3C.

[93] S. Lillioja, D. M. Mott, B. V. Howard, P. H. Bennett, H. Yki-Järvinen, D. Freymond, B. L. Nyomba, F. Zurlo, B. Swinburn and C. Bogardus , ‘Impaired glucose tolerance as a disorder of insulin action. Longitudinal and cross-sectional studies in Pima Indians’, *N Engl J Med*, vol. 318, no. 19, pp. 1217–1225, May 1988, doi: 10.1056/NEJM198805123181901.

[94] M. P. Amato, V. Zipoli, and E. Portaccio, ‘Multiple sclerosis-related cognitive changes: a review of cross-sectional and longitudinal studies’, *J Neurol Sci*, vol. 245, no. 1–2, pp. 41–46, Jun. 2006, doi: 10.1016/J.JNS.2005.08.019.

[95] S. M. Resnick, D. L. Pham, M. A. Kraut, A. B. Zonderman, and C. Davatzikos, ‘Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain’, *J Neurosci*, vol. 23, no. 8, pp. 3295–3301, Apr. 2003, doi: 10.1523/JNEUROSCI.23-08-03295.2003.

[96] K. v. Allen, B. M. Frier, and M. W. J. Strachan, ‘The relationship between type 2 diabetes and cognitive dysfunction: Longitudinal studies and their methodological limitations’, *Eur J Pharmacol*, vol. 490, no. 1–3, pp. 169–175, Apr. 2004, doi: 10.1016/j.ejphar.2004.02.054.

[97] J. H. Ware, D. W. Dockery, T. A. Louis, X. Xu, B. G. Ferris, and F. E. Speizer, ‘Longitudinal and cross-sectional estimates of pulmonary function decline in never-smoking adults’, *Am J Epidemiol*, vol. 132, no. 4, pp. 685–700, 1990, doi: 10.1093/OXFORDJOURNALS.AJE.A115710.

[98] J. Cohen, ‘A Coefficient of Agreement for Nominal Scales’, *Educ Psychol Meas*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.

[99] A. T. Broman, H. A. Quigley, S. K. West, J. Katz, B. Munoz, K. Bandeen-Roche, J. M. Tielsch, D. S. Friedman, J. Crowston, H. R. Taylor, R. Varma, M. C. Leske, B. Bengtsson, A. Heijl, M. He and P. J. Foster, ‘Estimating the Rate of Progressive Visual Field Damage in Those with Open-Angle Glaucoma, from Cross-Sectional Data’, *Invest Ophthalmol Vis Sci*, vol. 49, no. 1, p. 66, Jan. 2008, doi: 10.1167/IOVS.07-0866.

[100] M. Offroy and L. Duponchel, ‘Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry’, *Anal Chim Acta*, vol. 910, pp. 1–11, Mar. 2016, doi: 10.1016/J.ACA.2015.12.037.

[101] A. Dagliati, N. Geifman, N. Peek, J. H. Holmes, L. Sacchi, S. E. Sajjadi and A. Tucker, ‘Using topological data analysis and pseudo time series to infer temporal phenotypes from electronic health records’, *Artif Intell Med*, vol. 108, Aug. 2020, doi: 10.1016/J.ARTMED.2020.101930.

[102] E. H. Shortliffe and M. J. Sepúlveda, ‘Clinical Decision Support in the Era of Artificial Intelligence’, *JAMA*, vol. 320, no. 21, pp. 2199–2200, Dec. 2018, doi: 10.1001/JAMA.2018.17163.

[103] J. Tierny, *Topological Data Analysis for Scientific Visualization*, 1st ed. Springer Cham, 2017. doi: 10.1007/978-3-319-71507-0.

[104] S. Gholizadeh and W. Zadrozny, ‘A Short Survey of Topological Data Analysis in Time Series and Systems Analysis’, *arXiv preprint arXiv:1809.10745*, Sep. 2018, Accessed: Sep. 29, 2022. [Online]. Available: http://arxiv.org/abs/1809.10745

[105] K. R. Campbell and C. Yau, ‘Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data’, *Nature Communications 2018 9:1*, vol. 9, no. 1, pp. 1–12, Jun. 2018, doi: 10.1038/s41467-018-04696-6.

[106] P. M. Magwene, P. Lizardi, and J. Kim, ‘Reconstructing the temporal ordering of biological samples using microarray data’, *Bioinformatics*, vol. 19, no. 7, pp. 842–850, May 2003, doi: 10.1093/BIOINFORMATICS/BTG081.

[107] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, 1st ed. CRC Press, 1993.

[108] R. W. Floyd, ‘Algorithm 97: Shortest path’, *Commun ACM*, vol. 5, no. 6, p. 345, Jun. 1962, doi: 10.1145/367766.368168.

[109] ‘MOSAIC Project’, 2013. http://www.mosaicproject.eu/index.html (accessed Oct. 03, 2022).

[110] A. Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, V. Tibollo, R. De Cata, L. Chiovato and R. Bellazzi, ‘Machine Learning Methods to Predict Diabetes Complications’, *J Diabetes Sci Technol*, vol. 12, no. 2, pp. 295–302, Mar. 2018, doi: 10.1177/1932296817706375/ASSET/IMAGES/LARGE/10.1177_1932296817706375-FIG2.JPEG.

[111] L. Li, W. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger and J. T. Dudley, ‘Identification of type 2 diabetes subgroups through topological

analysis of patient similarity', *Sci Transl Med*, vol. 7, no. 311, p. 311ra174, Oct. 2015, doi: 10.1126/SCITRANSLMED.AAA9364.

[112] U. Brandes D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski and D. Wagner, 'On modularity clustering', *IEEE Trans Knowl Data Eng*, vol. 20, no. 2, pp. 172–188, Feb. 2008, doi: 10.1109/TKDE.2007.190689.

[113] Y. Li, S. Swift, and A. Tucker, 'Modelling and analysing the dynamics of disease progression from cross-sectional studies', *J Biomed Inform*, vol. 46, no. 2, pp. 266–274, Apr. 2013, doi: 10.1016/j.jbi.2012.11.003.

[114] I. Batal, H. Valizadegan, G. F. Cooper, and M. Hauskrecht, 'A Temporal Pattern Mining Approach for Classifying Electronic Health Record Data', *ACM Trans Intell Syst Technol*, vol. 4, no. 4, 2013, doi: 10.1145/2508037.2508044.

[115] I. Batal, D. Fradkin, J. Harrison, F. Moerchen, and M. Hauskrecht, 'Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data', *KDD*, vol. 2012, pp. 280–288, 2012, doi: 10.1145/2339530.2339578.

[116] R. Moskovitch and Y. Shahar, 'Classification of multivariate time series via temporal abstraction and time intervals mining', *Knowl Inf Syst*, vol. 45, no. 1, pp. 35–74, Oct. 2015, doi: 10.1007/S10115-014-0784-5/FIGURES/24.

[117] R. Moskovitch and Y. Shahar, 'Fast time intervals mining using the transitivity of temporal relations', *Knowl Inf Syst*, vol. 42, no. 1, pp. 21–48, Jan. 2015, doi: 10.1007/S10115-013-0707-X.

[118] R. W. Floyd, 'Algorithm 97: Shortest path', *Commun ACM*, vol. 5, no. 6, p. 345, Jun. 1962, doi: 10.1145/367766.368168.

[119] A. Sanchez-Palencia, M. Gomez-Morales, J. A. Gomez-Capilla, V. Pedraza, L. Boyero, R. Rosell and M. E. Fárez-Vidal , 'Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer', *Int J Cancer*, vol. 129, no. 2, pp. 355–364, Jul. 2011, doi: 10.1002/ijc.25704.

[120] H. Pei, L. Li, B. L. Fridley, G. D. Jenkins, K. R. Kalari, W. Lingle, G. Petersen, Z. Lou and L. Wang, 'FKBP51 Affects Cancer Cell Response to Chemotherapy by Negatively Regulating Akt', *Cancer Cell*, vol. 16, no. 3, pp. 259–266, Sep. 2009, doi: 10.1016/j.ccr.2009.07.016.

[121] '24. The National Center for Biotechnology Information: Gene Expression Omnibus (GEO) – Accession Display'. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11151 (accessed Mar. 27, 2020).

[122] C. Rosty, M. Sheffer, D. Tsafrir, N. Stransky, I. Tsafrir, M. Peter, P. de Crémoux, A. de La Rochefordière, R. Salmon, T. Dorval, J. P. Thiery, J. Couturier, F. Radvanyi, E. Domany and X. Sastre-Garau, 'Identification of a proliferation gene cluster associated with HPV E6/E7 expression level and viral DNA load in invasive cervical carcinoma', *Oncogene*, vol. 24, no. 47, pp. 7094–7104, Oct. 2005, doi: 10.1038/sj.onc.1208854.

[123] H. Tan, X. Wang, X. Yang, H. Li, B. Liu, and P. Pan, 'Oncogenic role of epithelial cell transforming sequence 2 in lung adenocarcinoma cells', *Exp Ther Med*, vol. 12, no. 4, pp. 2088–2094, Oct. 2016, doi: 10.3892/etm.2016.3584.

[124] A. Bordbar, J. M. Monk, Z. A. King, and B. O. Palsson, 'Constraint-based models predict metabolic and associated cellular functions', *Nat Rev Genet*, vol. 15, p. 107, 2014, doi: 10.1038/nrg3643.

[125] J. Han, L. V. S. Lakshmanan, and R. T. Ng, 'Constraint-based, multidimensional data mining', *Computer (Long Beach Calif)*, vol. 32, no. 8, pp. 46–50, Aug. 1999, doi: 10.1109/2.781634.

[126] W. H. Wolberg and O. L. Mangasarian, 'Multisurface method of pattern separation for medical diagnosis applied to breast cytology', *Proc Natl Acad Sci U S A*, vol. 87, no. 23, pp. 9193–9196, 1990, doi: 10.1073/pnas.87.23.9193.

[127] S. E. Sajjadi, B. Draghi, L. Sacchi, A. Dagliani, J. Holmes, and A. Tucker, 'Building Trajectories Over Topology with TDA-PTS: An Application in Modelling Temporal Phenotypes of Disease', Springer, Cham, 2020, pp. 48–61. doi: 10.1007/978-3-030-65965-3_4.

[128] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, 'Nuclear feature extraction for breast tumor diagnosis', in *Biomedical Image Processing and Biomedical Visualization*, Jul. 1993, vol. 1905, pp. 861–870. doi: 10.1117/12.148698.

[129] R. N. Weinreb, T. Aung, and F. A. Medeiros, 'The Pathophysiology and Treatment of Glaucoma: A Review', *JAMA*, vol. 311, no. 18, pp. 1901–1911, May 2014, doi: 10.1001/JAMA.2014.3192.

[130] H. Quigley and A. T. Broman, 'The number of people with glaucoma worldwide in 2010 and 2020', *Br J Ophthalmol*, vol. 90, no. 3, p. 262, Mar. 2006, doi: 10.1136/BJO.2005.081224.

[131] M. T. Leite, L. M. Sakata, and F. A. Medeiros, 'Managing glaucoma in developing countries', *Arq Bras Oftalmol*, vol. 74, no. 2, pp. 83–84, 2011, doi: 10.1590/S0004-27492011000200001.

[132] A. P. Rotchford, J. F. Kirwan, M. A. Muller, G. J. Johnson, and P. Roux, 'Temba glaucoma study: a population-based cross-sectional survey in urban South Africa',

Ophthalmology, vol. 110, no. 2, pp. 376–382, Feb. 2003, doi: 10.1016/S0161-6420(02)01568-3.

[133] A. Heijl, M. C. Leske, B. Bengtsson, L. Hyman, B. Bengtsson, and M. Hussein, 'Reduction of intraocular pressure and glaucoma progression: results from the Early Manifest Glaucoma Trial', *Arch Ophthalmol*, vol. 120, no. 10, pp. 1268–1279, Oct. 2002, doi: 10.1001/ARCHOPHT.120.10.1268.

[134] M. A. Kass, D. K. Heuer, E. J. Higginbotham, C. A. Johnson, J. L. Keltner, J. P. Miller, R. K. Parrish 2nd, M. R. Wilson and M. O Gordon, 'The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma', *Arch Ophthalmol*, vol. 120, no. 6, pp. 701–713, 2002, doi: 10.1001/ARCHOPHT.120.6.701.

[135] N. Mahabadi, L. A. Foris, and K. Tripathy, 'Open Angle Glaucoma', *Essence of Anesthesia Practice*, p. 161, Aug. 2022, doi: 10.1016/b978-1-4377-1720-4.00143-6.

[136] A. Tucker and D. Garway-Heath, 'The pseudotemporal bootstrap for predicting glaucoma from cross-sectional visual field data', *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 79–85, Jan. 2010, doi: 10.1109/TITB.2009.2023319.

[137] Y. Li, S. Swift, and A. Tucker, 'Modelling and analysing the dynamics of disease progression from cross-sectional studies', *J Biomed Inform*, vol. 46, no. 2, pp. 266–274, Apr. 2013, doi: 10.1016/J.JBI.2012.11.003.

[138] D. F. Garway-Heath, D. Poinoosawmy, F. W. Fitzke, and R. A. Hitchings, 'Mapping the visual field to the optic disc in normal tension glaucoma eyes', *Ophthalmology*, vol. 107, no. 10, pp. 1809–1815, 2000, doi: 10.1016/S0161-6420(00)00284-0.

[139] D. S. Kamal, D. F. Garway-Heath, R. A. Hitchings, and F. W. Fitzke, 'Use of sequential Heidelberg retina tomograph images to identify changes at the optic disc in ocular hypertensive patients at risk of developing glaucoma', *British Journal of Ophthalmology*, vol. 84, no. 9, pp. 993–998, Sep. 2000, doi: 10.1136/BJO.84.9.993.

[140] 'Advanced Glaucoma Intervention Study: 2. Visual Field Test Scoring and Reliability', *Ophthalmology*, vol. 101, no. 8, pp. 1445–1455, Aug. 1994, doi: 10.1016/S0161-6420(94)31171-7.

[141] Y. Li and A. Tucker, 'Uncovering disease regions using pseudo time-series trajectories on clinical trial data', in *Proceedings - 2010 3rd International Conference on Biomedical Engineering and Informatics, BMEI 2010*, 2010, vol. 6, pp. 2356–2362. doi: 10.1109/BMEI.2010.5639726.