Opening the Black Box:

Personalised Disease Prediction Using Hidden Variables and Dynamic Bayesian Networks

A thesis submitted for the degree of Doctor of Philosophy

by

Leila Yousefi

Doctor of Philosophy (Ph.D.)

Department of Computer Science

Brunel University London

September 2020

Abstract

The prediction of the onset of different complications of disease, in general, is challenging due to the existence of unmeasured risk factors, imbalanced data, time-varying data due to dynamics, and various interventions to the disease over time. Scholars share a common argument that many Artificial Intelligence techniques that successfully model disease are often in the form of a "black box" where the internal workings and complexities are extremely difficult to understand, both from practitioners' and patients' perspective. There is a need for appropriate Artificial Intelligence techniques to build predictive models that not only capture unmeasured effects to improve prediction, but are also transparent in how they model data so that knowledge about disease processes can be extracted and trust in the model can be maintained by clinicians. The proposed strategy builds probabilistic graphical models for prediction with the inclusion of informative hidden variables. These are added in a stepwise manner to improve predictive performance whilst maintaining as simple a model as possible, which is regarded as crucial for the interpretation of the prediction results. This thesis explores this key issue with a specific focus on diabetes data. According to the literature on disease modelling, especially on major diseases such as diabetes, a patient's mortality often occurs due to the associated complications caused by the disease over time and not the disease itself. This is often patient-specific and will depend on what type of cohort a patient belongs to. Another main focus of this thesis is patient personalisation via precision medicine by discovering meaningful subgroups of patients which are characterised as phenotypes. These phenotypes are explained further using Bayesian network analysis methods and temporal association rules. Promising results are documented on a real-world dataset of diabetes sufferers from an Italian Hospital, illustrating that firstly, hidden variable discovery within probabilistic graphical models can act as an ideal framework to improve prediction of comorbidities by modelling complex disease progression; secondly, that inference methods can aid the understanding of the influences of these hidden variables; finally, that the obtained significant subgroups of patients can be explained and characterised using a combination of latent variable analysis and temporal association rules so that clinicians can be empowered to focus on early diagnosis and treatment in a personalised way.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, my deep gratitude goes to my primary supervisor, Dr Allen Tucker, my commander in chief, who with patience cleverly guided me through one of the toughest but still most exciting learning experiences in my life. Although nobody is perfect, you made me chase perfection. Thank you for trusting in me and leading me as a smart spirit and a bright mind hero through the dark moments of soaring to my dreams. Secondly, but still a winner of mine, a massive recognition goes out to Dr Stephen Swift, for everything you have done for me, all the help you have given me – in both professional and private matters, as a supervisor, and as a friend. Special thanks also to my hero Dr Mahir Arzoky who have made my life happy in so many ways and been an important part of my life. Thanks for your compassionate heart. Another special honorary mention goes to Prof Martin Shepperd, Prof Steve Counsell, and David Jones, who with great kindness generously helped and supported me far more than I could wish for. Goodhearted people like you continuously make the world a better place.

My sincere thanks to the Intelligent Data Analysis group at Brunel, for its inspiring and welcoming atmosphere and amazing people. My appreciation also extends to my colleagues and friends at Brunel for collaboration, professional sparring, and an amazing social sphere over the last four years.

On a personal level, a special thanks to my highly valued far distanced friends. Elmira, you honestly nurtured my soul and became one of my main motivations in life, you are a true blessing. Meysam, for being my friend, thanks for always being there for me. Amin, for the encouragement which was crucial for me in taking on this challenge, and hence getting to this point. And not least, Tobias, who has given the greatest gift which moulded me in a better person. Your kind words warmed my heart. Thanks for your compassionate heart. Lastly, but by far the most, I want to thank my parents for their priceless support. The love and appreciation I have for you are beyond words. I could not have done this without you.

This thesis is dedicated with endless love and gratitude to my parents.

Chapter 1

Introduction

This thesis explores Artificial Intelligence (AI) techniques for modelling the progression of disease whilst simultaneously stratifying patients and doing so in a transparent manner as possible. It uses diabetes as a case study. Diabetes is a chronic disease with an onset that is commonly associated with multiple life-threatening comorbidities (complications). Early prediction of diabetic complications and the behaviour of associated risk factors can reduce patients' suffering time. Therefore, models of time-series diabetic data (which can often be imbalanced, incomplete and involve many complex interactions) are needed to better manage the disease. Unlike earlier work in modelling diabetes, here the focus is upon a combination of both descriptive and predictive data mining methodologies to accurately predict complications through explainable patient models. Firstly, the thesis describes how best to enhance the prediction performance and reduce bias in the models inferred from complex time-series data; secondly, it deals with imbalanced clinical data by using various re-balancing approaches, whilst determining the precise influences of unknown risk factors (hidden/latent variables) within a probabilistic network framework; finally, it considers how to group the patients into meaningful subgroups by means of latent variables in order to discover how complications can interact differently on some patients. This introductory chapter describes the motivation behind the proposed methodologies and contributions of the research, sets out some initial background and outlines a roadmap of the thesis.

1.1 Motivation

Clinicians attempt to help patients by using quality care for a range of life-threatening diseases, while they monitor the associated comorbidities. Despite the recent improvement in general practice, it has been reported that nearly half of patients still do not receive this expected care [57]. Clinicians also predict disease and related complications based on their prior knowledge and an individual patient's clinical history. Many studies have attempted to find automated ways of helping clinicians to predict disease progression. However, data that is required to learn predictive models are often biased or limited, though better predictions could save the National Health System (NHS) billions of pounds [92]. This thesis looks at Type 2 Diabetes as a case study, which is often known as a "silent killer". It is increasingly seen as a serious, worldwide public health concern. The World Health Organisation (WHO) claims that diabetes is a major cause of blindness, heart attacks, kidney failure, stroke, and foot damage. It also reported that Type 2 Diabetes Mellitus (T2DM) accounts for at least 90% of all types of diabetes. Based on another investigation carried out by the WHO, it revealed that in the next ten years, there would be about 550 million people suffering from this disease, and it would be the 7^{th} leading cause of death. The first step in the development of T2DM appears to be a condition where cells of the body develop insulin resistance. This is because the patient's body does not respond well to insulin—insulin should be informing cells to request blood sugar, but in insulin resistance, cells tend to ignore this signal. Eventually, after a few years of insulin resistance, the outcome is T2DM with high blood sugar levels (which can negatively affect the nerves, blood vessels, and cause more complications to be developed) [31]. It has been observed that patients with T2DM are also at increased risk of micro-vascular comorbidities, including kidney damage (nephropathy), nerve damage (neuropathy) and eye disease (retinopathy) [102].

Patient mortality often occurs due to complications caused by the disease and not the disease itself. Nevertheless, for a long time, these life-threatening complications have remained un-diagnosed because of the hidden patterns of their associated risk factors [138].

Lack of prediction of the onset of associated diseases/complications can negatively affect a patient's health in many ways. They can be numerous and interact in complex non-linear

ways throughout the disease process. Patients must switch to different medications as more complications develop. T2DM is potentially reversible, treatable, and manageable if caught early enough. Early diagnosis and management of the disease can reduce the risk of complication development [9]. As a result, clinical data needs to be considered as a time-series so that the progression of the disease can be captured as early as possible. However, dealing with time-series patient records is known to be a significant issue in the prognosis of complications [9]. This is because predicting a target complication can be challenging without the consideration of associated historical complications. Many of the state-of-the-art AI techniques used in modelling disease are often in the form of a "black box", such as deep learning approaches where the internal workings and complexities are extremely difficult to understand, both from practitioners' and patients' perspectives. Unlike simpler models such as logistic regression, these approaches are complex and not easy to explain. For example, the complexity of countless hidden layers in a deep neural network and their interconnections makes it challenging to determine precisely how predictions are being made.

1.2 Contributions

This thesis contributes in several ways to our understanding of AI models can be used to generate accurate predictions, whilst remaining explainable. It demonstrates how a graphical model approach with latent variables can provide a basis for better prediction of disease complications while assessing whether the controlled explicit modelling of unmeasured effects is an appropriate way for "Opening the black box" in disease prediction. These contributions can be summarised as follows:

1. **Utilising appropriate data mining (supervised and unsupervised learning) approaches to modelling disease:**

    - **Modelling disease by using probabilistic AI models:** Probabilistic graphical models such as Dynamic Bayesian Networks (DBNs) are chosen because both explanation and prediction are key. These models are more informative from the

qualitative point of view, have demonstrated much promise in the modelling of disease progression and can naturally incorporate hidden variables. In this research, a time-series predictive model was explored for the early prediction of the comorbidities from the diabetic patients' follow-ups at the IRCCS Istituto clinic scientific (ICS) Maugeri of Pavia, Italy.

2. **Dealing with highly unbalanced clinical data:** In traditional disease prognosis, there are too many false positives / false negatives due to clinical data often being highly imbalanced. Using a class balancing method along with the DBNs, would clearly be beneficial. This thesis demonstrates how to extend DBN models to handle highly unbalanced time-series clinical data:

   - Time-Series Bootstrapping: A bootstrap technique was used that has been specifically designed for longitudinal data where the occurrence of the positive class occurs far less than the negative (typical in complications for patients diagnosed with diabetes). The results of this study have illustrated that re-balancing data demonstrated an improvement in prediction performance.
   - Pair-sampling Strategy: Pair-sampling was exploited to effectively address unbalanced time-series medical data. This method divided the dataset into positive and negative patient instances, from which the train and test data sets were generated.
   - Classifying Disease Complications: This work investigated the problem of discovering the relationships and interactions between binary T2DM complications whilst addressing the unbalanced nature of the data when stratifying patients.

3. **Modelling complex interactions among both observed disease risk factors/-complication and unmeasured effects using a targeted hidden variable approach:** This thesis explored the explicit modelling of relationships between latent variables and clinical features within a DBN framework. The discovered latent variables help to reduce the uncertainty in the prediction process by identifying the relationship between T2DM complications and risk factors.

- Discovering a hidden variable and finding its precise location within the DBN structure.
- Obtaining an optimal number of hidden variables in a stepwise approach to avoid creating overly complex models that risk overfitting and becoming "black box" in nature: This study proposed a novel methodology for using multiple hidden variables in a DBN structure so that unmeasured effects could be captured. A stepwise hidden variable approach was developed. The extensive set of experiments showed that the proposed method improves prediction accuracy, whilst identifying the correct number of hidden variables, and targeting their precise location within the network structure (therefore aiding explanation).
- Incorporating a Combination of the IC* algorithm and a Mutual Information based scoring metric to identify the strength of relationships between the latent variables and clinical risk factors: The hidden variable structure was learned by a constraint-based method which calculated several conditional independence tests. A measure of link strength was exploited to calculate the overall strength of the dependent links. These combined methods helped to focus on the most powerful dependencies between T2DM risk factors, enabled us to observe the specific impact of each discovered edge in a DBN, and obtained a reliable structure.
- Exploiting appropriate data mining approaches to model complex interactions among the complications: The temporal association rule mining was proposed to extract the meaningful temporal patterns and sequence of time-series complications.

4. **Personalising and handling the variability of progression in patients by identifying subgroups of patients that share similar behaviours (via a latent temporal phenotype):** Unsupervised learning was used to identify cohorts of patients with similar sub-classes of disease trajectory and complications co-occurrence pattern. Subgroups of patients based upon a time-series clustering and hidden variable discovery approach were found and aligned with the discovered clusters from the association rules into knowledge. This research was a first attempt to combine hidden variable discov-

ery with temporal association rules for investigating the relationships among the T2DM complications. The following methodologies were utilised:

- The characterisation of temporal phenotypes from discovered hidden variables.
- Using a combination of time-series clustering with dynamic time warping and the Jaccard index to group patients.
- The discovery of temporal phenotypes was combined with Temporal Association Rule Mining to find similar subgroups of patients that aids explanation.

5. **Focusing on explainable AI approaches to help "Open the Black Box":** Throughout the thesis, a focus was kept on transparent models in the form of a probabilistic graphical framework, a controlled approach to discovering hidden variables and a method for the interpretation of their influences and semantics within a clinical perspective.

   - This thesis attempted to interpret hidden variables. By exploiting methods to explicitly model unmeasured risk factors within a graphical structure, not only could disease progression modelling accuracy be improved, but it also allowed clinicians to better understand their meaning.
   - Discovering temporal phenotypes by identifying underlying sequences of temporally associated complications for specific patient subgroups.

   The promising experimental results demonstrate that patient personalisation using the proposed methodology could provide better prediction accuracy and interpretability.

1.3 Thesis Outline

In Figure 1.1, the thesis methodology is outlined and organised as follows: Chapter 2 provides a comprehensive literature review of the various machine learning approaches in the disease prediction process. It then describes the T2DM dataset and some preliminaries (with notation) for the thesis content. An explanation of the time-series methodologies on diseases is presented, followed by an introduction of the patient modelling approaches. Chapter 3 shows

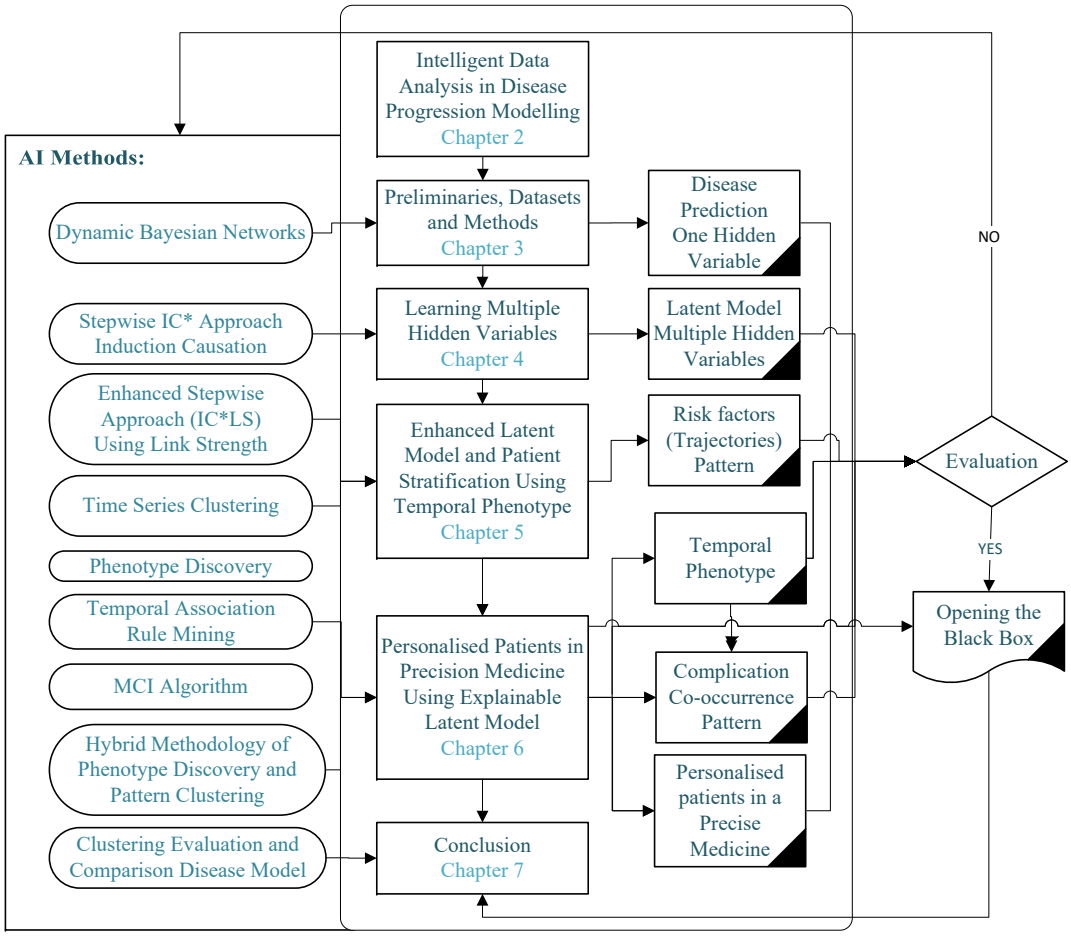

Figure 1.1: Methodology Process Diagram.

the key contribution of this work to model the time-series of clinical data with a graphical probabilistic model. It mainly focuses on the state-of-the-art in the use of the DBNs and how they can model Diabetic patients. It also includes the evaluation of the most common temporal pattern discovery methods with respect to understanding and uncovering the hidden variables. Part of the main findings in this chapter has been published in [140]. Chapter 4 presents a key contribution of the thesis based on the identification of hidden variables in a DBN framework to both improve the predictive accuracy and understand the relationship between hidden variables and the observed risk factors. It discusses how to obtain multiple hidden variables in a stepwise approach while monitoring improvement in the prediction performance. It contributes to how the discovered hidden variables explain the DBNs-based patient model. Most part of this chapter has been published in [141, 144]. Chapter 5 presents another key contribution based on patient stratification and modelling. It extends the work in the previous chapter by utilising a time-series bootstrapping approach to balance data. It learned an optimal number of the hidden variables in an enhanced version of the stepwise approach proposed in the previous chapter and contributed a journal paper currently under the peer review [139]. Chapter 6 employs both predictive and descriptive model and further characterisation of different patterns in diabetes progression. It discusses the novelty of mining Temporal Associated Rules for stratifying patients into meaningful sub-groups. The patients' clusters found from the temporal phenotype (obtained in chapter 5) are validated by using the clusters obtained based on temporal association rules. Part of this work has been published in [141, 142, 143]. The validation also aims to ensure a more meaningful characterisation of the subgroups of patients can be identified (explaining the behaviour of the latent phenotype). Part of this work has been published in Computational Intelligence journal [142]. Finally, chapter 7 concludes and provides directions for future research.

1.4 Publications

- Predicting comorbidities using resampling and dynamic bayesian networks with latent variables. In 2017 IEEE 30th International Symposium on Computer-Based Medical

Systems (CBMS) (pp. 205-206). IEEE.

- Predicting complications in Type 2 Diabetes Mellitus Using Dynamic Bayesian Networks. IDA Springer, The 17 International Symposium on Intelligent Data Analysis.

- Predicting disease complications using a stepwise hidden variable approach for learning dynamic bayesian networks. In 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS) (pp. 106-111). IEEE. “The best paper and presentation award.”

- Predicting complications in Type 2 Diabetes Mellitus Using Dynamic Bayesian Networks, 2018 IEEE on Computer-Based Medical Systems (CBMS) Sweden PhD Consortium.

- Opening the Black Box: Discovering and Explaining Hidden Variables in Type 2 Diabetic Patient Modelling. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1040-1044).

- Opening the Black Box: Exploring Temporal Pattern of Type 2 Diabetes Complications in Patient Clustering Using Association Rules and Hidden Variable Discovery. In 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 198-203), won "Vice-Chancellor’s Travel Prize”.

- Computer Science Brunel PhD Symposium (CSBPS 2019) and Poster Presentations, "Opening the Black Box in Disease Prediction".

- Opening the Black Box: Patient Personalisation Using Hidden Variables in Type 2 Diabetes Prediction. Journal of Computational Intelligence (Special Issue of Biomedical (Big) Data Science).

- Predicting Type 2 Diabetes Complications and Personalising Patient Using Artificial Intelligence Methodology. A chapter of Type 2 Diabetes book (2020), publisher IntechOpen.

- Identifying Latent Variables in Dynamic Bayesian Networks with Bootstrapping Applied to Type 2 Diabetes Complication Prediction. Journal of Intelligent Data Analysis (IDA). Unpublished journal paper (2022).

Chapter 2

Intelligent Data Analysis in Disease Progression Modelling

2.1 Introduction

This chapter reviews the current literature on some of the most common AI methodologies, including probabilistic modelling, association rule mining, and latent variable discovery. Intelligent Data Analysis (IDA) is a subcategory of AI that is focused on data analysis and modelling. These methods are known to be highly successful in combining advantages of modern data analytics, classical statistics and the expertise of scientists and experts [14, 16, 56]. IDA techniques have already proved successful in clinical modelling [10]. A large and growing body of literature has investigated IDA approaches that have shown excellent results modelling cross-sectional clinical data for classification. There has also been substantial modelling on longitudinal data using IDA techniques. However, there is still an urgent need to improve these models to take account of the variability of disease progression from person to person, and explicitly model the time-varying nature of the disease. Many studies have attempted to find automated ways of helping clinicians predict disease progression [108].

For many clinical problems, the underlying structure of unmeasured variables may play an essential role in the progress of the disease. However, it is still a relatively unexplored

area. Identifying these unmeasured variables as hidden or latent variables is key. What is more, understanding the semantics behind these unmeasured risk factors can improve the understanding of the disease mechanisms and thus better improve clinical decision making. Interpreting these latent variables is complicated; however, as they may represent different many types of unmeasured information such as social deprivation, missing clinical data, environmental factors, time-based information or some combination of these. To gain trust in any AI model, it is mandatory to understand/explain influencing factors of disease that guide predictions or decisions. This is because clinicians expect to understand AI diagnoses to be able to make decisions. There is a great deal of debate over the importance of explanation in AI models inferred from health data. In particular, there is a balance that needs to be made between the accuracy of complex "deep" models such as convolutional neural networks (in predictive strategy) and the transparency of models (in descriptive strategy) that aims to model data in a more "human" way such as expert systems.

A combination of explainable and "deep" strategies rather than either one of them alone would have a better prognostic value. Furthermore, in order to obtain a more accurate and explainable prediction of progression, the predictive models need to be personalised based on how an individual patient matches historical data by identifying patient subgroups.

2.2 Probabilistic Model for Time-Series Analysis

Understanding the pattern of complications associated with the disease has been used significantly in the clinical domain [132]. It provides an insight into the prediction and relative prevention of the associated complications which are expected to occur in a patient follow-up [48]. It generally can lead to less suffering time for patients while saving time and cost to healthcare. However, that is highly dependent on the stage of disease along with the prior occurring complications, which is associated with time-series analysis. In time-series analysis, every disease risk factor and complication is determined by various features in previous patient visits (time interval). At every medical visit, all diabetic patients have a unique profile of symptoms and complications that change over time, regardless of the phase of the disease.

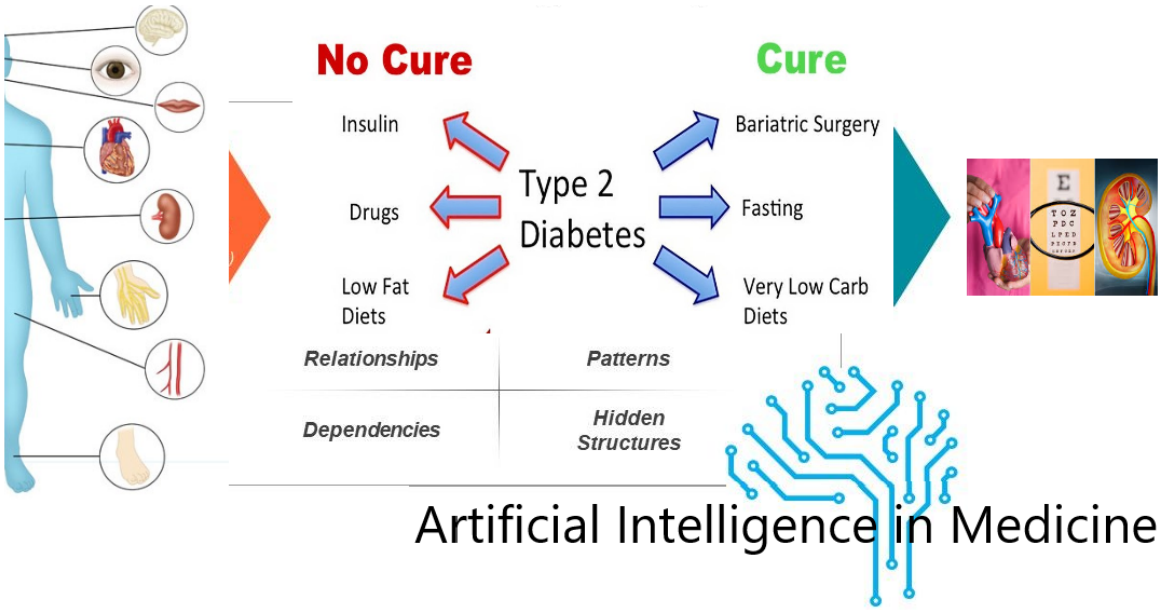

Figure 2.1: The organs/muscles affected by the common complications associated with Type 2 Diabetes

This non-stationary characteristic of clinical data collected as part of the monitoring of T2DM creates a difficult context for effective forecasting [131]. Clinical data needs to be considered as time-series data in order to provide a description of the progression of a disease over time. Nevertheless, dealing with time-series patient records is known to be a major issue in the prognosis of comorbidities [130], particularly when time-series data is imbalanced and contains few examples of patients without comorbidities that are common to all patients. In Type 2 Diabetes, for example, once patients are diagnosed with T2DM, half of them show signs of complications [128]. Unfortunately, these life-threatening complications remain undiagnosed for a long time because of the hidden patterns of their associated risk factors [138]. If T2DM is not appropriately managed, the development of serious complications, such as neuropathy, retinopathy, and hypertension lead to disability, premature mortality and financial cost [41]. The prediction process is complex due to the interactions between these complications and other features, as well as between complications themselves. More importantly, each patient has a unique profile of complication occurrence and the status of T2DM risk factors during a patient's time-series is subject to change, as their levels may rise and fall over time. Early diagnosis and prevention techniques are needed to reduce the associated mortality and morbid-

ity caused by T2DM complications [79]. Although there are various methodologies for T2DM prediction, for example, risk-prediction equation and Markov models [87], studies that enable early predictions of diabetes using predictive models are limited [61]. The risk-prediction equations suffer from uncertainty as well as performing only one-step-ahead predictions, while Markov models are limited to a small number of discrete risk factors.

There are various methodologies for T2DM prediction, e.g, risk-prediction equations and Markov models [87]. In addition, various studies on longitudinal data mining literature suggest an association between T2DM comorbidities and risk factors, e.g., [1]. Research on T2DM prediction has often been restricted to modelling a limited number of visits. The most existing literature on investigating the prognosis of T2DM complications, e.g., [36, 37] focuses particularly on logistic regression and Naïve Bayes. For example, for investigating the prognosis of T2DM complications, [37] focuses on logistic regression and Naïve Bayes methods. In Naïve Bayes, there is an assumption of independence among the risk factors whereas all features are independent of one another. Dagliati et al. [36] presented a Hierarchical Bayesian Logistic Regression model to anticipate patients changes when the individual model parameters are estimated. The major limitation of the previous work in T2DM literature derives from time discretisation in temporal time slices per year. For example, they were limited to the external and internal heterogeneity to model T2DM patients for predicting comorbidities in cross-sectional data with just three horizons of time. They presented a parameter estimation Markov Chain Monte Carlo (MCMC) approach, which might not be suitable for large datasets and time-series modelling was not employed in the individual measurements. Moreover, logistic regression does not perform well when there are multiple or non-linear decision limitations. Therefore, this study considers all T2DM patient's follow-up visits regardless of year basis while precisely monitoring the location of change within the unequal number of visits.

The mentioned research differs from the work presented in this thesis in terms of modelling strategies, handling of unbalanced data, and the combinations of predictors. Nevertheless, extensive research has been carried out on the prediction of diabetic progression, no single study exists which has attempted to interpret the impact of hidden variables on the predictive model of diabetic disorders. Overall, such studies seemed to be unsatisfactory for modelling the

complex T2DM complications/risk factors. Therefore, this thesis suggests that AI in Medicine can provide useful techniques to analyse patient data to be able to find cure for the disease or reduce patient's suffering time (see Figure 2.1).

2.2.1 Dynamic Bayesian Networks

In the field of medical informatics, probabilistic IDA techniques are exploited to obtain different clinical solutions. To improve patient quality of life, there is an urgent need to extend and explore probabilistic IDA methods to answer to investigate the disease complications from a clinical point of view. Bayesian Network models appear to be well suited T2DM progression modelling, because of their flexibility in modelling spatial and temporal relationships as well as their ease of interpretation [97]. Thus, a Bayesian Network (BN) decision model was exploited in [107] for supporting the diagnosis of dementia and Alzheimer disease and mild cognitive impairment. It has been reported that Dynamic Bayesian Networks (DBNs) are simple BNs for modelling time-series data and popular for modelling uncertain noisy time-series clinical data [89]. More importantly, DBNs are probabilistic graphical models that can handle missing data and hidden variables.

Previous work on learning DBNs have inferred both network structures and parameters from (sometimes incomplete) clinical datasets [89]. For example, a recent study presented a DBN method but to analyse fisheries data [125]. Authors in [54] proposed a Bayes Network to predict diabetes on the Pima Indian Diabetes dataset. However, the study failed to consider the time-series analysis. Similarly, authors in another study [81] simulated the health state and complications of type 1 diabetes patients by using partially and entirely learned Bayesian models. Apart from using a different type of Diabetes, this thesis is utilising a different approach from the above studies for the representation of the relationship between T2DM risk factors. Many diseases involved structural changes based upon key stages in the progression, but many models did not appear to take this into account. There has been some work in extending DBNs to model underlying processes that are non-stationary [126]. In [126], clinical features were modelled using a second-order time-series model while time-invariant temporal dependencies were assumed. Among this, some studies, for example, Marini and co-authors

conducted research [81] that variables were connected within two-time-series and within the same time slice assumed that the temporal dependencies were time-invariant. In addition, in Marini's paper for learning the network structures, a Tabu search was used based on the Hill climbing algorithm for Bayesian Networks but with no use of latent variables. However, the approach was useful for stratifying patients according to the probability of developing complications, the major limitation of the Marini's work derived from time discretisation in time slices of one year.

Another work in [53] retained the stationary nature of the structure in favour of parameter flexibility, arguing that structure changes lead almost certainly to over-flexibility of the model in short time-series. Alternatively, a paper [104] formalised non-stationary DBN models and suggested MCMC sampling algorithm for learning the structure of the model from time-series biological data. Similarly, authors in [118] estimated the variance in the data structure parameter with an MCMC approach, but the search space was limited to a fixed number of segments and indirect edges only, which is not suitable for T2DM data. Such studies remained narrow and limited by constraints on one or more degrees of freedom: the segmentation points of the time-series, the parameters of the variables, the dependencies between the variables and the number of segments and the ignorance of the incomplete data and latent variable.

2.2.2 Dealing with Time-Series Imbalanced Data

Another common problem with classifying complications in longitudinal data is that there may be many more cases where the complication does not manifest compared to those where it does. Early prediction of T2DM complications while discovering the behaviour of associated aggressive risk factors can help to improve a patients quality of life [77]. This study suggests that while there is an association between the latent variable and joint complications in the prognosis of T2DM patients, this relationship is complex. In T2DM data analysis, another challenge can be to classify/group patients in imbalanced clinical data with several binary complications. Models of the time-series data are needed to manage diabetic complications and deal with their imbalanced and complex interactions. In particular, mining time-series is one of the challenging problems in the prognosis of disease. In addition, it has received

considerable critical attention in data mining especially when there are rare positive results [137]. It has been reported that a class imbalance in the training data caused by one class (here positive cases) massively outnumbers the examples in another class (negative class) [72]. This situation may occur where the number of positive clinical test results for a complication is not equal or even close to the number of negatives. That can be solved by applying an appropriate balancing strategy in a multi-class classification problem. Different learning techniques deal with imbalanced data, such as oversampling, undersampling, boosting, bagging, bootstrapping, and repeated random sub-sampling [62]. This thesis in order to prepare T2DM data for the prediction has utilised these strategies and customised them based on dataset nature (time-series patients records with the unequal number of visits). As a result, various balancing strategies such as pair-sampling, bootstrapping undersampling and over-sampling have been proposed in [140, 144]

The bootstrap approach can be used to identify the significant statistics from classifiers learnt from such data. For example, in a study [71], Li and co-authors provide an extension to the temporal bootstrap approach while applied on cross-sectional data. Similarly, a study conducted in [125], the bootstrap strategy is extended to longitudinal data by sampling pairs of time points, thus enabling the (first-order) temporal nature of the data to be inferred. However, these solutions only can be applicable when the imbalance ratio for all binary complications is similar. Otherwise, it can be more difficult if we need to over-sample one class value and under-sample others in order to reduce bias from data.

Overall, the observed balancing strategies from the prior studies have not been sufficient for analysing more than one complication at a time, whereas it was almost impossible to obtain a satisfactory prediction performance enhancement for all complications. As well as modelling unmeasured factors, hidden variables can also be used to model non-stationary processes. This thesis attempts to address this issue by using hidden variables discovery approaches based upon T2DM risk factors/complications dependencies. Before explaining these strategies, it is necessary to understand unmeasured variables and analyse their dependencies that are generated by causal structures.

2.2.3 Causal Structure Learning and Latent Variable Discovery

Moreover, various studies on longitudinal data sets have suggested an association between complications and risk factors of the disease. To discover probabilistic dependencies given clinical data, it is necessary to search the space of belief networks or casual models, which is called casual discovery of BNs [146]. These patterns of dependency with no model based solely upon the observed variables can be explained by using a latent variable. The casual discovery indicates dependencies that are generated by casual structures with unmeasured factors, i.e., hidden variables. Hidden variable modelling, introduced in [113], has a long tradition in casual discovery. One of the research gap in the previous literature of disease prediction is the existence of the unmeasured or latent variables. This is because clinicians cannot measure all risk factors and carry out all kinds of tests, so there are some unmeasured factors that clinicians fail to measure, which need to be discovered at the early stage of diabetes.

Furthermore, Factor Learning (FL) was introduced in [82],which has been known as one method for learning a probabilistic model from data. It can also be helpful to understand latent variables and measure their hypothetical impacts. FL contrasts with most other BN learning methods in that it learns a factor structure. As Martin and co-authors in [82] stated that FL for hidden variables could identify the most probable structures of factors have given the data and suitable priors. However, with a large number of variables, FL methods might be prohibitively expensive. Again in the same research these authors provided a factor structure for learning methods that efficiently utilised hidden variables. Factor structure indicates the joint probability distribution among discrete observed variables. It also contributes an explanation across a small number of variables. Although factor structures are suitable for polynomial time inference, they can cause a reduction in the prediction accuracy and precision; they contribute an explanation across a small number of variables. Nevertheless, these techniques failed to consider prior belief in the factor structure, and therefore, it could be hard to rely on the final structure.

Factor structure indicates the joint probability distribution among discrete observed variables. Interestingly, each factor in a factor structure corresponds to a completely connected dependency graph. Although they are suitable for polynomial time inference, caused reducing

accuracy and precision. By contrast, they are not able to decide precisely whether or not latent variables are present, and in consequence there has been some controversy about that status of exploratory versus confirmatory factor analysis. In this regard, casual discovery methods in AI have the advantages as they can discover the actual dependencies and independencies in the data.

The causal discovery of BNs is a critical research territory, which depends on looking through the space of causal models for those which can best clarify a pattern of probabilistic conditions appeared in the data [146]. As a result, [19] showed the integration of structure-search algorithm with a latent variable in a DBNs model. However, the method did not consider the discovery of the long-range dependencies with an equal number of time slices. Similarly, in [34], Bayesian belief networks was used to find the most probable structure, using the K2 algorithm, while adding a hidden variable. Nevertheless, Cooper in [34] applied the K2 method that needs an ordering on the nodes. Witting focused on using hidden variables in a known structure [136]. Cooper in [34] used Bayesian techniques to find the most probable structure and can use this technique to add hidden variables. In principle, exact Bayesian methods for hidden variables could identify the most probable structures of factors given the data and suitable priors. However, with a large number of variables, exact methods are prohibitively expensive. Furthermore, in [110] Silva highlighted the weakness of DAG (Directed Acyclic Graph) models in the marginalisation of Hidden factors and representing the independencies over a subset of features in a DAG with more links. They suggested that Directed mixed graphs (DMGs) are a solution to this drawback. Therefore, they represented how to perform Bayesian inference on two DMGs, such as Gaussian and Probit, which is not the focus of this thesis.

Nevertheless, such studies remained narrow and limited by constraints on one or more degrees of freedom: the segmentation points of the time-series, the parameters of the variables, the dependencies between the variables and the number of hidden factors. As a result, Chicharro in [26] analysed causal influences to find the relationship among different brain regions in several disorders. Similar to this thesis, Chicharro's research made use of Inductive Causation (IC*) algorithm in the latent process to analyse Granger causality and Dynamic

Causal Modelling. However, Chicharro's study did not consider DBNs to understand causal influences.

Difficulties arise, however, when an attempt is made to implement a Bayesian Network structure as authors in [8] have argued that the number of potential DAGs over the disease risk factors is super-exponential. Additionally, the real cause-effect relationship DAG is not distinguishable while from equivalent structures when learning only using from observational data. This issue will be worse, especially when each expert has a unique probability of correctly labelling the inclusion or exclusion of edges in the disease structure. As noted by Amirkhani [8], some scoring functions are provided with that score each suitable graph based on the data and experts' knowledge. Another research in [105] shows that networks with the fixed structure containing hidden variables can be learned automatically from data using a gradient-descent mechanism similar to that used in neural networks.

A few algorithms have been created to understand the structure for Bayesian Networks from both fully observed models and those with hidden variables. Structure Expectation-Maximization (SEM) has been produced for learning Probabilistic system structure from information with latent factors and missing data. A structure learning algorithm has been created for non-stationary dynamic probabilistic models. For example, REVEAL (REVerse Engineering ALgorithm) has been utilised as a structure learning algorithm, that learns the optimal set of parents for each node of a network independently, based on the information-theoretic concepts of mutual information analysis. However, the two-stage temporal Bayes network (2TBN) cannot be well recovered by the application of REVEAL. A normally utilised structure learning algorithm depends on REVEAL which takes in the ideal arrangement of guardians for every hub of a system autonomously, in light of the theoretical data ideas of common data examination. Be that as it may, the two-arrange fleeting Bayes organise as the 2TBN which cannot be all around recuperated by use of REVEAL. Rijmen in [103] exploited an HMM to study the temporal pattern of symptoms burden in brain tumour patients. He showed that the discovery of symptom experience over time is necessary for treatment and follow-up of patients with symptom-specific intervention.

In general, Bayesian learning methods could determine network structure and how the net-

work's variables should be represented along with the causal links among them. Moreover, it addressed the difficulty of qualifying causal relationships in terms of Conditional Probability Tables (CPTs). Witting focused on using hidden variables in a known structure [136] as the knowledge of the latent variable in predictive modelling is important for an understanding of the complex AI models. Discovering latent variables can potentially capture unmeasured effects from clinical data, simplifying complex networks of interactions and giving us a better understanding of disease processes. In addition, it can improve classification accuracy and boost user confidence in the classification models [49]. Elidan and co-authors in [46] emphasised the importance of the presence of hidden variables. In addition, they determined a hidden variable that interacted with observed variables and located them within the Bayesian Network structure. They also showed that networks without hidden variables are clearly less useful because of the increased number of edges needed to model all interactions, which caused overfitting. Despite the productivity of exploring trees of hidden variables to render all observable variables independently [95], these hidden variables were non-optimal with independencies among observable variables.

Overall, previous works on learning DBNs have presented both network structures and parameters from clinical data sets and learning parameters for a fixed network of incomplete data, in the presence of missing data and latent variables [89]. Much of the current literature on disease prediction have argued that a complex AI model, with many unexplainable hidden variables, also has several serious drawbacks. Therefore, this thesis has chosen AI DBNs model to learn parameters and latent variables to predict complications. The next section intends to emphasise the explainability of the proposed methodology in order to uncover the meaning behind the latent AI model.

2.3 Black Box Models and AI in Medicine

Investigating unmeasured risk factors can improve the modelling of disease progression and thus enable clinicians to focus on early diagnosis and treatment of unexpected conditions. However, the overuse of hidden variables and lack of explainability can lead to complex models, which are

not well understood (being black box in nature). Models need to be understood by clinicians to facilitate transparency and trust.

Neural Networks (NNs) are a robust methodology to approximate complex functions in IDA literature and are known as the black box models. Black box AI models in decision making are mostly based on deep learning techniques with many latent variables. Black box AI models in decision making are mostly based on deep learning techniques with many latent variables. Deep learning approaches attempt to model complex interactions in data by using a considerable number of hidden variables. For example, NNs is a representation of function with some parameters and latent variables which are weights of Neuron.

In the 1980s there was a huge wave of excitement in NNs. Then Boltzmann machines were published in 1985, the back-propagation paper that appeared in nature came out in the 1986 and Parallel Distributed Processing (PDP) appeared in 1987. Neal's article in [91] explicitly linked feed forward NNs, which were called Connectionist NNs and graphical models in belief networks. In particular, in sigmoid belief networks feed forward NNs was used with sigmoid activation functions instead of considering those units as deterministic and considering them as binary random variables.

Despite their ability to provide an accurate predictive model, NNs often fail to provide any insights on the network structure in regard to explain the approximation function and the latent layers. In 1987 Danker, in a NN was conducted to choose probability distribution over the weight and to map the weight space on two function spaces represented by a NN. This was a more intuitive factor to consider than the distribution of weights, which was an arbitrary nuisance parameter. Danker contribution attempted to algorithmically figure out how to compute the posterior over the parameters. In addition, the parameters were optimised and computed the second derivative of the likelihood with respect to the parameters which was Hessian to estimate the posterior with a diagonal version of Laplace approximation.

In 1989, Tishby [123] applied Bayes rule on NNs as an interesting application to demonstrate the utility of the average prediction error for determining a sufficient size of the training set as well as selecting the optimal architecture of the network. Additionally, Tishby's study was used Bayesian inference to figure out the optimal architecture of NNs. Further, in 1991, Buntime

[21] clearly expressed Bayesian inference on NNs.

In 2006 Hinton [59] showed the importance of using "complementary priors" to reduce inference complexity in belief networks with many hidden layers. He derived a greedy algorithm from learning deep and directed belief networks one layer at time. Neal in [90] attempted to find a single "optimal" weight vector in conventional network training that led to over-fitting and weak generalisation.

2.3.1 Explainability in Deep Learning

This stage outlines and discusses the limitations of Deep Learning approaches that have been proposed, so far, to gain deeper insights into the understanding of black box AI models. AI medical machine such as Deep Learning has become ubiquitous to provide a high-performance prediction.

Nevertheless, understanding their mechanisms has become a significant concern worldwide whereby the goal is to gain clinicians and patients trust. The reason behind this is due to several obstacles that arise to interpret the findings, such as the scale of big data, complex interactions, and high-dimensional internal state.

Google's Novel Approach Most medical algorithms proposed by [27], such as "AI Doctor" designed to reproduce current problem-solving methods (e.g., the detection of cancers). In addition, the concept assignment can help people to strengthen their skills and talents for a computer system that showcased superhuman effectiveness and efficiency.

Google's AI Doctor can be demonstrated how they could be used to provide an explanation further into predictions generated by local classifiers, first from conventional image classification networks to a focused clinical application.

The concept attribution approach in AI Doctor offers several promising avenues for future work. In addition to this, the concept assignment can help people to strengthen their skills and talents for a computer system that showcases superhuman effectiveness and efficiency.

The concepts of explanatory power are outlined by Google under three principle assumptions/limitations: firstly, comprehension for whatever hidden layer and artificial neurons would

offer. This is based on most of the information in a deep neural network consists of hidden layers. Secondly, it recommends that acknowledging the numerous hidden layers and understanding their design on a meta-level would lead to more in-depth modelling. Finally, to comprise how nodes become active, it considers groups of interconnected neurons that trigger at the same time and space. These principles are defined instead of explaining the structural nature of each neuron in each network. This is because the stratification of a network for the categories of interconnected neurons would enable its configurations even more abstractable. This is the main weakness of the black box models.

One of the most highlighted ones is Google's approach to resolve the explainability issues while enabling human-like description of the internal state of a deep network by employing Concept Activation Vectors (CAVs). While medical systems are mostly designed to reproduce current decision-making methods such as the classifier used in the detection of cancers, Google has claimed that its novel strategy can interpret existing clinical data.

Although Google has made a claim that the CAVs can directly relate to one's anticipated theories, to draw conclusions about the decision-making process, it needs to consider the human needs of a higher level of understandability. Evidently, Google suggests that these aspects may not require to be understood at the early project stages, and can, therefore, be swiftly verified through a set of instances throughout the longitudinal study. To achieve this, Google proposed the CAVs Testing (TCAV) approach that demonstrates whether CAVs are used to measure the extent whereby the given definition needs to be applied. This technique is employed in TCAV can be a move towards establishing a humanistic understanding of the internal state of a Deep Learning model.

In the detection of various functionalities for the deep neural networks, TCAV may only have a few implications of causal inference. As a result, Google AI model revealed that by monitoring individuals' eyes, it might be possible to estimate the risk of developing the cardiovascular disease for individuals. Google AI algorithm classified these patients 30 percent of the time compared to a traditional predictor (SCORE) that had less classification error by 28 percent. For Google, the research is much more than just a modern cardiovascular risk assessment tool. This is because the methodology of Google seems reliable, as the retina has

a long history of cardiovascular risk prediction.

Nevertheless, cardiac specialists have been critical of the conclusions derived by Google in the clinical domain. With the proper information, AI is optimistic that innovative, unique healthcare insights might be created without human intervention. Unfortunately, this new approach is only established based on extensive and adequate datasets. This is presumably part of the explanation of why Google has established projects as its benchmark research proposal is capturing detailed patients' history of 100000 population across four years. However, the investigation conducted out by Google did not necessarily indicate that the suggestion was entirely distant. Such as image classifiers that could be applied to low-level structures.

The central concept and assumption are to consider a neural network as additional assistance that can cause issues related to the internal representation. As a result, the clinicians commented on the deep explanatory networks. They questioned the hypotheses, by stating that although the AI algorithms and Deep Learning could improve current prediction methods of clinical domain, the research would not be trustworthy unless it had been assessed with caution while a broader range of disease had been explored. Difficulties arose, when an attempt was made in order to implement the principles and these assumptions. It seemed to be evident that their approach was overconfident and yet to be trusted.

Prototyping Examples In Artificial Neural Networks In order to introduce a different perspective on Deep Learning models' interpretability, Zintgra and co-authors [66] conducted a study to simplify the black box structure of Artificial Neural Networks (ANNs). They made use of "prototypic examples" method that indicate tools in order to diagnose trained ANNs. In general, ANNs analyse discrete decision-making processes and obtain high-performance prediction results.

The prototype examples may be computationally intractable, including a pre-determined normal distribution to prevent the proliferation of unreasonable prototype cases. They provided an explanation of tools to train ANNs based on two datasets. Moreover, it can often be like such a losing battle to describe precisely how ANNs operate mathematically. Therefore, a much more comprehensive pre-processing methodology could also be used in a related development

(e.g., generative adversarial network proposed by Goodfellow et al. in [27]). Furthermore, experimental results and hypotheses in ANNs were portrayed and tested only on two datasets. Alternatively, a more detailed analysis is required to rely on the empirical results, which might be achieved by including rich data containing imbalance issue, different types of features. Selection bias was another potential concern because it could involve possible measurement errors. It could be extended through more set of data with various features.

Finally, conclusions and interpretations of data were drawn from an inevitably subjective mechanism on the investigator's basis. This was because to examine whether the produced case studies should satisfy the investigator's standards about the phenomena of been modelled (e.g., decisions could be only made by the time it came). This was established based on approaches or standards for collecting and analysing concepts that might be more unbiased. As a result, this could also enable investigators/analysers to understand the implications and weaknesses of the use of ANNs for the discrete decision-making process, which might enhance the strictness of the approach. However, many healthcare methods are required to reconstruct conventional prediction methods (e.g., the identification of cancers), but so far, different ideas to interpret previous clinical records have been discovered.

Visualisation in Deep Learning: For the time being, the possibility of an AI physician planning to roll new prognosis without direct human intervention is a significant distance in which the more presumably in decades rather than a few years later. Recent developments in several technologies in the Deep Learning area have been powered by the steadily declining expense of computing and storage. That being said, realistic apps, including certain integrated smartphone and electronic devices, have intensified explainability issues for Deep Learning in the black box resource-limited environments.

Liu et al. in [75] introduced the leading solution to address these issues where a deteriorated image of Binary Convolutionary Networks caused by binarising Filtres. They offered a range of Circulant Filtres (CiFs) and a Circulant Binary Convolution (CBConv) to strengthen efficiency and to tackle those limitations for Binary Convolutionary functionalities through their proposed Circulant Backpropagation (CBP). Then, CiFs effortlessly was integrated into the current deep

neural networks (DCNNs). Enormous research has indicated that perhaps the output difference amongst one-bit and total-precision DCNNs could be reduced by extending the variety and distributing the filtres. Zintgraf et al. in [148] identified numerous tools to test the model and understand how DCNNs could provide a reliable outcome by using the visualisation method.

Overall, the existing explanatory Deep Learning approaches would need to be adapted for further sophisticated longitudinal modelling strategy (rather than with a multivariate distribution). This would result in better outcomes, for example, in pixel values which could be estimated reliably by everyone's environment while it skewed down much more. By providing the black box models with sufficient data, machine learning seemed to be overconfident that completely different health knowledge could then be generated without user intervention. The black box models of Deep Learning can be simplified in several aspects. For example, if an object is detected, an image detection machine can breakdown back and towards specific attributes including shape, colour and texture of the image, and then reduce the predictions to a mathematical method by checking the classification error and then background diffusion to improve the practises.

In particular, in the world that it is possible to fully allocate decision making to computer systems, confidence in AI systems will be hard to achieve. In the future work, one approach that can be applied to the small-sized T2DM dataset can be the use of Bayesian Neural Networks, which will deal with uncertainties in data and model structure by exploiting the advantages of both Neural Networks and Bayesian modelling. To conclude, AI can improve current methods of medical diagnosis in terms of interpretability but cautioned that the technology would need to be more evaluated to be trusted by both patients and practitioners.

In black box models, it can be challenging to determine what is coordinating the visible patterns. Such models are problematic not only for lack of transparency but also for possible biases inherited by the algorithms from clinician's mistakes [98]. This issue is caused based on the human errors and biased sampling of training data as well as the underestimation of the impact of the risk factors underlying behaviour/pattern.

In general, as observed from prior studies, it is difficult to obtain performance enhancement while simultaneously trying to explain hidden factors. Lakkaraju in [69] suggested that there

is a trade-off between patient personalisation (in a descriptive analysis) and prediction performance (in predictive analysis). Generally speaking, an improvement in explainability is often possible through a less accurate model or at a higher cost of the predictive accuracy (in a Black box model) [132]. There are quite few research studies on predicting T2DM complications and T2DM black box models. However, studies on explaining an unknown risk factor/latent phenotype by using a hybrid data mining methodology (including descriptive and predictive) are rare to find in literature. Therefore, this study attempts to open the AI, black box model by using both predictive and descriptive strategies.

2.4 Patient Personalisation and Explanation

Most of the previously published studies in diabetes prediction have tended to focus on all patients as one integrated database rather than separating patients [36]. It can be challenging to stratify patients based on their longitudinal data in order to determine what is triggering the visible patterns that may be specific to one cohort of patients. There is some research, such as [24] that assesses the disease prediction performance based upon different IDA techniques. For example, the onset of the disease is modelled in [78] while other studies focus on patient modelling [99]. The approach described in this thesis aims to personalise patients by using unsupervised methodologies to group time-series patient data. The proposed descriptive strategy in this thesis has been regarded as a useful tool known as association rules to detect interesting relationships among T2DM complications.

2.4.1 Time-series Clustering

Time-series clustering is often problematic [2], especially when we need to analyse risk factors from matching patterns across time. The literature on time-series clustering and pattern discovery has highlighted several studies [7]. There have been some qualitative measures for clustering time-series data, which captured similar risk factor patterns in dynamic temporal data, regardless of whether the correlation between them was linear or not [29]. However, they did not seem to be very suitable for a long and an unequal number of time-series data (e.g.,

T2DM data). For instance, authors in [7] proposed an algorithm to cluster patients based on clinical data whilst utilising the clustering information for identifying distinct patterns.

Altiparmak in [7] provided a slope-wise comparison method (SWC) to find the correlation between local distance vectors of patient's visits, and group clinical test results into different sub-groups, based upon the related risk factors, by using feature selection. In their method each cluster of patients was considered as a transaction data that included a pattern indicating which cluster belonged to each patient. Authors in [40] used a similar method [7] in clustering, but they clustered fixed length time-series. Ceccon and coauthors [25] exploited a variation of the naive Bayes classifier with a hidden variable for segmenting patients into disease sub-types. Ceccon's study intended to enhance the classification performance of Glaucoma patients based upon visual field data. Nevertheless, they only focused on standard/static BNs (instead of DBNs) to infer the parameter in a cross-sectional dataset. Moreover, they failed to analyse the influences of multiple hidden variables on the prediction results.

2.4.2 Pattern Discovery and Association Rules Mining

It has previously been observed that patients with T2DM are also at an increased risk of microvascular comorbidities, including nephropathy, neuropathy, and retinopathy [102]. The underlying pattern of T2DM complications and how their co-occurrence is followed/caused/related by other complications associated with the disease, known as the major source of mortality and morbidity in T2DM [88]. That is because predicting a target complication can be challenging without the consideration of the effects of its associated complications. Similar to Diabetic type 1 patients, although genetic factors impact on developing T2DM, it is believed ignorance of developing complications harms patients' life. What is more, T2DM patients develop a different profile of complications and features, which changes over time per follow-up visit. One of the most important factors in the high number of dependencies among T2DM features and complications is the appearance of unmeasured risk factors. Surprisingly, the effect of understanding unmeasured variables, which play an important role in disease prediction, does not seems that closely examined.

Understanding these associated patterns has a remarkable actual value and can significantly

being used in the clinical domain [132]. It provides an insight into the prediction and relative prevention of the associated complications which are expected to occur in patient followups [48]. It also leads to less suffering time for patients while saves time and cost to healthcare. However, that is highly dependent on the stage of disease along with the prior occurring complications, which is associated with time-series analysis. In time-series analysis, every disease risk factor and complication is determined by various features in previous patient visits (time interval). To better understand the complications of the disease and their effects, this thesis clusters patient the associated rules among the complications. It attempts to address this issue and present an informative rules/ordering pattern of patient behaviour, with an aim to capture the complexities of the associated complications' over time. The proposed descriptive strategy has been regarded as a useful tool known as association rules (ARs) to detect interesting relationships among T2DM complications.

Association Rules (ARs) originated from learning patterns from supermarket transaction data, and they were introduced by Agrawal in [3]. Temporal Abstraction (TA) was employed in [86] for the segmentation and aggregation of a time-series into a symbolic representation. TA has appeared to be a suitable solution for decision making and data mining. With a slightly similar objective to this thesis, Moskovitch and co-authors [86] conducted a study in which time-interval mining method obtained informative temporal patterns for finding relationships in the transitivity inherent of time-series diabetic patients. They also exploited TA to mine and aggregate time-series into a symbolic representation. Although Moskovitch's paper coincides our study by using supervised learning in time-series Diabetes data, it differs from this work in finding meaningful time-series patterns only based on gender not complex temporal patterns from a longitudinal clinical dataset with the appearance of latent risk factors.

Temporal Association Rules (TARs) [133] is an extension to association rules [3] to analyse basket data that includes a temporal dimension to order related items. Many algorithms with temporal rules work by dividing the temporal transitions database into different partitions based on the time granularity obliged. For example, different mining algorithms were reformulated and presented to reflect the new general temporal association rules. These include Progressive Partition Minder (PPM), Segmented Progressive Filter (SPF), and TAR algorithm

[3, 70, 133].

Various algorithms have been proposed for the incremental mining of temporal association rules, especially for numerical attributes [50]. Allen's rules [6] generalised abstracted time-series data into a relation (PRECEDES) to find TARs in [106]. Various ways were proposed to explore the problem of temporal association rules discovery [4]. Nevertheless, previous studies performed discovering association rules on a given subset specified by the time [63], whilst not considering the specific exhibition period of the elements.

Association Rule Mining (ARM) finds frequent patterns by mining ARs with the use of two basic parameters of support and confidence [147]. The majority of the previous ARM algorithms worked by dividing the temporal transitions database into different partitions based on the time granularity obliged. Then mining temporal association rules are employed by locating frequent temporal item subsets within these partitions. Whereas the incremental mining of temporal association rules for numerical attributes cannot be easily adapted to a transaction database. Despite all efforts, no method exists today that can find meaningful subgroups of patients based on the underlying pattern of complications in the existence of the latent risk factors.

Difficulties arise with TARs when there are some rare rules of particular interest [83]. Many studies have employed the most common filtering metrics rather than support and confidence in order to detect interesting rules [119]. There is a controversy to this, as a study in the literature argued that a conservative ARM methodology only based on a fixed and rigid threshold for the filtering metrics could be problematic. A few studies attempted to mine frequent underlying patterns of diabetic complications [44]. The frequent pattern mining research significantly affects data mining techniques in longitudinal data. A post-processing approach in [33] attempted to extract interesting subsets of temporal rules within T2DM data. However, it only considered characteristic patterns of administrative data without the appearance of latent variables. Other researchers have undertaken association rule mining of clinical data [43, 93]. Lee et al. attempted to address the issue in [70] and have led to the proposal of the concept of general TARs, where the items were allowed to have varying exhibition periods, and their support was made based on that accordingly.

Another research conducted by Plasse et al. in [100] looked at finding homogeneous groups of variables. They suggested that a variable clustering method could be applied to the data in order to achieve a better result in pattern discovering methodology. However, their strategy to mine ARs differed from this thesis in which the number of rules was reduced only based on hierarchical clustering applied to items, not to multiple identical binary attributes. Among these, some methods uncovered temporal patterns and relationships among clinical variables, including causal information [80], numeric time-series analysis [112]. Nevertheless, considering all of this evidence, none of the above studies has clustered uneven time-series clinical data based on a hidden variable for extracting temporal phenotype and behaviours of patients.

2.5 Summary

This chapter has described the research gap in the modelling and explaining of complex disease processes and thus given the motivation behind this work. The previously discussed methods suffer from some limitations in addressing imbalance issues, complex and temporal relationships between (sometimes unmeasured) factors, and the identification of different underlying characteristics of disease for different subgroups of the population. There is considerable research on predicting T2DM complications. Among these, studies on explaining unknown risk factors and identifying temporal phenotypes by using hybrid methods (including descriptive and predictive) are rare to find in literature. This thesis attempts to address these issues. In the next chapter, after describing the case study data as a starting point, DBNs are explored as a framework for modelling real time-series clinical data. In the following chapters, the identification of informative hidden factors is investigated followed by methods to cluster patients into meaningful subgroups along with the identification of a latent temporal phenotype and the characterisation of these groups using temporal association rules.

Chapter 3

Preliminaries, Datasets and Methods

3.1 Introduction

This chapter describes some existing key methods that can be updated or combined to model multiple diabetes complications in the presence of unmeasured factors. It focuses on rule-based methods for an explanation of patient subgroups and a probabilistic framework for modelling data explicitly. Intelligent systems, whether biological or artificial, require the ability to make decisions under uncertainty using the available evidence. Several computational models exhibit some of the required functionality to handle uncertainty. These computational models in AI and Machine Learning are judged by two main criteria: ease of creation and effectiveness in decision making. For example, NNs which represent complex input/output relations using combinations of simple nonlinear processing elements, are a familiar tool in AI and computational neuroscience. Alternatively, probabilistic networks (also called Bayesian Networks) are a more explicit representation of a domain through modelling the joint probability distribution (the probability of all possible outcomes in a domain). This is achieved by providing a topological description of the conditional independence relationships among variables. This chapter firstly describes the dataset that will be the main focus of the thesis. It then goes on to explain the definitions and algorithms that are key to the work. Temporal association rules are described in detail followed by an introduction to Bayesian Networks and a full formal definition of the Dynamic Bayesian Network. Re-balancing techniques are also explained in

the context of the T2DM data. Finally, some initial results are demonstrated on the diabetes dataset where class imbalance poses a problem due to the rare occurrence of different individual complications. This is dealt with using a DBN with fixed single hidden node combined with a bootstrap technique that has been specifically designed for the longitudinal data to identify targeted complications.

3.2 Type 2 Diabetes as a Case Study (Data Selection)

The World Health Organisation (WHO) reported that Type 2 Diabetes Mellitus (T2DM) accounts for at least 90% of all diabetes types. Another study in WHO revealed that T2DM patients are at increased risk of long-term vascular comorbidities, which is known as "underlying cause of death" and severe phenotype of the disease [52]. It has previously been observed that patients with T2DM are also at an increased risk of microvascular comorbidities, including nephropathy, neuropathy, and retinopathy [52]. As T2DM is a rising public health concern worldwide. It is a chronic disease with an onset that is commonly associated with multiple complications, such as retinopathy and nephropathy. Models of the time series data (which are often imbalanced and involve complex interactions) are needed to better manage diabetic comorbidities. Although extensive research has been carried out on the prediction of diabetic progression [12, 13, 36, 37, 81], no single study exists which has attempted to interpret the impact of latent variables in the presence of diabetic disorders.

T2DM data was chosen as a case study for this work as it suits the characteristic of complex and small-sized dataset with uneven number of visits per patient, after performing the centre profiling and a detailed analysis of the literature reported in [11]. In T2DM dataset with each visit, a patient has a unique profile of symptoms and complications, regardless of the phase of the disease.

3.2.1 Data Collection

This work presents an evaluation of the analysis on T2DM datasets collected in the Pavia data warehouses funded by the 7th Framework Programme (FP7-ICT 600914) and information

regarding the data is retrieved from MOdels and Simulation techniques for discovering diabetes Influence faCtors (MOSAIC) project website [22]. The Data is belonged to the MOSAIC European Union project reported most of the information presented in this section retrieved from MOSAIC website [22]. The data collection accumulated in the Italian Pavia region in the MOSAIC project contains data including 1000 patient populations and extracted from external ways such as: the Hospital Fondazione Salvatore Maugeri (FSM), that mostly captures epidemiological records relevant to normal healthcare sector, and by the Local Healthcare Agency (Agenzia Sanitaria Locale, ASL), that accumulates measurements for institutional and technical transparency.

The combination of different sets of data gives a full summary of the medical knowledge of people with diabetes participated in the research. Different factors included within the study involve demographic information (age, sex, period of diabetes), medical information from both the comprehensive FSM, and the ASL data (haemoglobin glycated, systolic blood pressure, lipid profile, smoking habit, and body mass index) and administrative data gathered via ASL (drug consumption). The periods of participant follow-ups used throughout the data can be seen in Figure 3.1 and scale between 1-year and 18-year duration, by a mode of 5-year period.

Throughout the MOSAIC project, designers originally investigated the prediction issue of choosing appropriate probability calculators for complications to be included in the patient population. A cardiovascular risk score, the validated "ProgettoCuore score16", was accessible [22]. A such scoring was tailored to the Italian community, based from the "Framingham study17". FSM, though, identifies individuals with dysfunction in metabolic processes and heart failure cases. For Indications, which is why, in the FSM Clinical data, microvascular complications are more likely to occur after the first visit than macrovascular complications. In addition, patients typically tend to be treated with FSM (and their data collected) only after onset of the disease. Because of the existence of the evidence gathered in Pavia, no microvascular risk model might be found which could have been implemented without a robust scaling drift [12, 13].

Mining clinical data is a challenging task given the mixture of clinical test results (such as blood pressure or cholesterol level), complication types (categorical, numerical, ordinal data

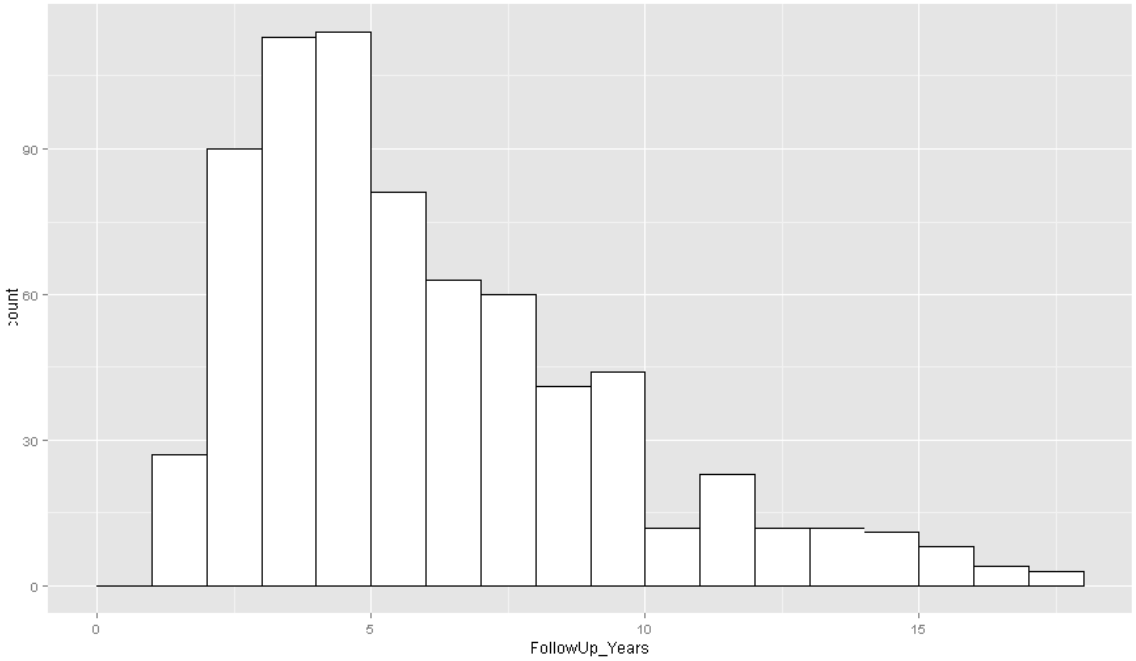

Figure 3.1: Follow Up Duration [36]

types), unequal length of patients visits, highly correlated risk factors, unmeasured factors, heterogeneity, biased data, and more. It aims to build upon this thesis by conducting dynamical analyses on time-series clinical data to improve classification accuracy. With a probabilistic graphical model, it is easy to interpret and provide information regarding the qualitative structure of the clinical domain. For instance, it helps to find out whether risk factors are direct causes or influencing factors of a disease or complication. Moreover, it seems extremely important to learn a model from the clinical data with a small amount of training data with many parameters and few patients (samples). More importantly, in clinical domain each parameter is supported by little evidence and estimation of the parameters is not robust.

3.2.2 Data Description

The data for this study consists of pre-diagnosed T2DM patients aged 25 to 65 years (inclusive) that were recruited from clinical followups at the "IRCCS Instituti Clinic Scientifici" (ICS) Maugeri of Pavia, Italy. The MOSAIC project funds the data under the 7^{th} Framework

Program of the European Commission, Theme ICT–2011.5.2 Virtual Physiological Human (600914) from 2009 to 2013. The dataset consists of physical examinations such as cholesterol and blood pressure and laboratory data, including HbA1c measurements and lipid profile. For this study, certain complications and risk factors (predictors) were selected based on existing literature on diabetes [5, 30, 44, 124, 129] and using recommendations from the clinicians at ICS. The selected T2DM complications are Retinopathy (RET), Hypertension (HYP), Nephropathy (NEP), Neuropathy (NEU) and Liver Disease (LIV). Here, the predictors are identified and selected from the dataset: Body Mass Index (BMI), Systolic Blood Pressure (SBP), High-Density Lipoprotein (HDL), Glycated haemoglobin -HbA1c- (HBA), Diastolic Blood Pressure (DBP), Cholesterol (COL), Smoking habit (SMK) and Creatinine (CRT). It is necessary here to clarify exactly what is meant by Control Value and Discretised Value.

In T2DM data, the worsening level of the micro-vascular diseases and HYP is known as a significant cause of death [35]. Even though micro-vascular complications such as RET, NEP, NEU are less frequent comparing to HYP, an inadequate estimation of them causes long-term suffering and life-threatening comorbidities [88]. Fowler and co-authors in [48] researched type 2 Diabetic American patients. This research utilised T2DM key risk factors such as HbA1c, SBP, and DBP to investigate relationships among complications such as HYP, NEP, RET, and NEU. In addition, LIV is a severe phenotype of diabetes and associated with T2DM complications, especially NEU [122]. Litwak analysed Russian diabetic patients in [74] which referred to the influence of macro-vascular and micro-vascular disease on one anther. For example, important features in T2DM dataset such as blood pressure, HDL, lipid, BMI, and HbA1c influence diabetic patient's complications. They also revealed that HDL has a negative effect on HYP, NEP, NEU, and RET, whereas HbA1c negatively associated with HYP. Again, a study conducted by Ramachandran [101] referred to the high prevalence of NEU and RET in Type 2 diabetes in India. Similarly, research in [44] suggested that most of the diabetic patients have objective evidence for some variety of NEU, but only a few of them have identified by symptoms. This research also showed that there is a strong association among NEP, NEU, and RET.

Tables 3.1-3.2 illustrate Control Values for T2DM complications and risk factors, respec-

tively. In Table 3.1 shows the binariased complications with two clinical level of High and Low. This study only concentrates on five binary complications as the predictive target classes in a binary classification problem (with two categories of classes: "high" or "low" risk). Furthermore, a complication class value of low risk (zero) represents a patient visit in which the complication is not present; otherwise, it is at high risk (one). For instance, a complication class value of zero represents a patient visit in which the complication is not present; otherwise, it is one.

Alternatively, in Table 3.2, T2DM risk factors associated with a patient (symptoms/clinical tests) are abstracted in the multi-class classification problems with more than two targets risk patient, according to a diabetes expert's definitions [12, 13]. Clinical risk factors are consists of three clinical level of risk, namely low (0), medium (1) and high (2). Node ID column is used as the risk factor identifier. For instance, in order to help distinguishing the clinical features of smoking habit where Node ID equals to 13, discretised into three categories (0,1,2), namely non-smoker, ex-smoker and smoker. The term 'HDL' is used here to refer to High-Density Lipoprotein as well as Lipid Mechanism (which was analysed in this Chapter's experimental analysis).

Table 3.1: The description of T2DM Target Complication, Clinical Node Control Values, and Discretised States.

Node ID	Target Complication	Diagnosis Outcome	Clinical Risk Class
2	Retinopathy (RET)	{Negative,Positive}	{low,high}
3	Neuropathy (NEU)	{Negative,Positive}	{low,high}
4	Nephropathy (NEP)	{Negative,Positive}	{low,high}
5	Liver Disease (LIV)	{Negative,Positive}	{low,high}
6	Hypertension (HYP)	{Negative,Positive}	{low,high}

3.2.3 Value of the observed and Unmeasured Data

In this case study, the association of non-binary risk factors have not been considered directly in order to extract rules among T2DM complications. Alternatively, non-binary features are involved indirectly. The reason behind this is that the overall behaviour of these T2DM risk factors is captured by utilising a latent variable discovered by using the constraint based

Table 3.2: The description of the T2DM Clinical Features, Risk Factors, Control Values, and the Discretised States.

Node ID	T2DM Risk Factors	Control Value (Mean±SD)	Discretised Value
1	HbA1c (HBA)	6.6 ± 1.2 (%)	{low,medium,high}
7	Body Mass Index (BMI)	26.4 ± 2.4 ($kg/m2$)	{low,medium,high}
8	Creatinine (CRT)	0.9 ± 0.2 (mg/dL)	{low,medium,high}
9	Cholesterol (COL)	0.9 ± 0.2 (mg/dL)	{low,medium,high}
10	High-Density Lipoprotein (HDL)	1.1 ± 0.3 ($mmol/l$)	{low,medium,high}
11	Diastolic Blood Pressure (DBP)	91 ± 12 ($mmHg$)	{low,medium,high}
12	Systolic Blood Pressure (SBP)	148 ± 19($mmHg$)	{low,medium,high}
13	Smoking Habit (SMK)	{0,1,2}	{low,medium,high}

algorithm in a DBN framework (which is called "temporal phenotype" and will be explain in Chapter 5).

These non-stationary device dynamics is modelled with hidden variables. They reflect a transition in the relationships between the environmental factors experienced across time. In this case, the significance for the latent factor is set to refine the model fit to the data while the model is parameterised by data (such as the log likelihood). If in the time series, e.g., the slope of an association between two components increases, the value of the hidden variables correlated with these components will differ as the trends for the observed ecosystem components change. One, some or all of the observed environment components within the model may be connected to a latent variable. The discovered latent variables may represent a different type of predictions, such as life expectancy, quality of life, or the spread of specific disease or comorbidities. The latent variable value then depends on all the process elements of which it has been related, and a shift in trends means that process relationships have shifted. This is incredibly beneficial in T2DM dataset in which non-stationary risk factor dynamics are widespread and complicated.

3.2.4 Temporal Pattern of Complications and Data Notations

It has previously been observed that patients with type 2 diabetes mellitus are at increased risk of microvascular comorbidities including retinopathy, neuropathy, and nephropathy. Predicting the comorbidities has long been a question of great interest in a wide range of medical fields.

This study have tried to predict the future state of a patient per visit by utilising a set of observed test based on the complication patterns.

Tables 3.1-3.2 represent the selected T2DM complications (comorbidities), risk factors and their clinical control values. T2DM dataset is discretised into qualitative states (binary complications and non-binary features) of ordinal clinical risk by using statistical parameters such as mean, median, and Standard Deviation (SD).

As mentioned earlier, the main objective of this work is to predict future T2DM complications model architecture. Having provided the health state of the patient on the first visit, there is a need to foresee whether nephropathy, neuropathy or retinopathy will continue to progress in the long term. For each patient, the posterior probability in predicting a target complication is predicted at time t with the observed evidence (prior knowledge) from $t-1$ to estimate the risk of developing complication for the corresponding patient patient. Therefore, considering how the state of patient during each visit changes can be an important challenge for physicians preparing for future visits. The main goal of this thesis is to understand the underlying patterns of associated binary complications.

To study diabetes, a random sample of patients with different comorbidities was recruited from Pavia clinical data in Italy. However, due to visit constraints, this research cannot provide a comprehensive review of these comorbidities which were diagnosed before diabetes onset. This study is unable to encompass the entire comorbidities dataset. Therefore, this method is particularly stated on studying the microvascular complications (retinopathy, neuropathy, and nephropathy). The purpose of this study is to show how to improve the quality of life by anticipating the future stages of the diabetes complications for different patients at their various visits.

From diabetes health status records, the T2DM dataset is accumulated (which is denoted here as DS) from pre-diagnosed diabetic patients. For each patient in T2DM dataset defined the following notations:

Let π demonstrates a distinct patient where i identifies the patient in which $i \leq p$, and $p = 356$ denotes the maximum number of patients in DS. In addition, time-series between the first visit and the j^{th} visit of i^{th} patient (π_i) is represented by V_{ij}. For each of the patients

in DS, over which a linear order of in $[V_{iv}\ V_{iz}]$ is defined to represent all visits between V_{iv} and V_{iz} for π_i where $v \leq z$ and V_{iv} occurs before or is earlier than V_{iz} $(V_{iz}) = \sum V_{iv}, ..., V_{iz})$. The number of visits is not necessarily equivalent to each patient π_i and varies between two and 300 ($2 \leq T_i \leq 300$). Hence, there is a total of $T = 3959$ visits/instances/time-series in DS, which contains the temporal observations of the occurring complications for all T2DM patients. Let $\pi_i = \sum_{j=1}^{T_i} V_{ij}$ and $V_{ij} = (V_{i1}, V_{i1}, V_{i2}, ..., V_{iT_i})$ be a set of visits for i^{th} patient with T_i time-series. In order to clarify the dataset, Equations 3.1-3.2 are defined to illustrate dataset based upon individual patient or patient's time-series (visits) and the corresponding complications pattern, respectively.

$$DS{=}\sum_{i=1}^{p} \pi_i{=}(\pi_1, \pi_2, ..., \pi_i, ..., \pi_p), 1 \leq p \leq 356 \tag{3.1}$$

$$DS{=}\sum_{i=1}^{p} \sum_{j=1}^{Max(T_i)} V_{ij}{=}([V_{1T_1}], [V_{2T_2}], ..., [V_{iT_i}], ..., [V_{pT_p}]), 1 \leq T \leq 3959 \tag{3.2}$$

$$
\pi = \left[
\begin{array}{ll}
\pi_1 = & \left[
\begin{array}{ll}
\textbf{Visits} & \textbf{Complications Pattern} \\
V_{11} & \{\} \\
V_{12} & \{HYP\} \\
V_{13} & \{HYP\} \\
V_{14} & \{HYP, LIV\} \\
V_{15} & \{HYP, LIV, NEU\} \\
V_{16} & \{HYP, LIV, NEU, NEP\} \\
V_{17} & \{HYP, LIV, NEU, NEP\} \\
V_{18} & \{HYP, LIV, NEU, NEP\}
\end{array}
\right], \\
\pi_2 = & \left[
\begin{array}{ll}
V_{21} & \{\} \\
V_{22} & \{RET\} \\
V_{23} & \{RET, HYP\} \\
V_{24} & \{RET, HYP, NEU\} \\
V_{25} & \{RET, HYP, NEU, LIV\} \\
V_{26} & \{RET, HYP, NEU, LIV\} \\
 & . \\
 & . \\
 & .
\end{array}
\right], \\
\pi_i = & \left[
\begin{array}{l}
V_{i1} \\
V_{i2} \\
. \\
. \\
. \\
V_{iT_i}
\end{array}
\right], \\
\pi_p = & \left[
\begin{array}{l}
V_{p1} \\
V_{p2} \\
V_{pT_p}
\end{array}
\right]
\end{array}
\right] \tag{3.3}
$$

3.3 Descriptive Data Analysis

This section, firstly describes the clinical data and descriptive analyses used throughout this thesis. It then explains the solutions, which are explored in this study to deal with missing data and class unbalance problems, as well as our model design options.

One amongst MOSAIC's major objectives is to use knowledge discovery methods to help explain the processes driving the complications of diabetes via the study of particular testimonials, time events as well as behavioural influences.

Due to the initiatives carried out by its experts in past initiatives and research, several medical resources on an Eu commission have been provided to the MOSAIC committee. MOSAIC data have been used to build predictive model that combine knowledge associated with environmental and biomedical influences, including physiological, biological, epigenetic modification and behavioural inputs. The aims of such modelling techniques are to identify specific therapeutic processes in medical history and to stratify the patient at risk of suffering T2DM as well as its associated complications. The goal is to incorporate these into existing predictive analytics frameworks to improve decision-making in patient practice.

The Knowledge Discovery in Databases (KDD) process is exploited, in this study, to identify co-occurrence of the complications at different levels of abstraction from the T2DM dataset. It started the first stage of KDD with T2DM raw data and ended with extracted knowledge captured as a result of the following stages as seen in [47]:

1. Data selection and determining subset of patients with unequal number of visits.

2. Pre-processing for cleansing and removing all uninteresting and uninformative information about patients such as dates of their visits.

3. A transformation phase for transforming common complications of T2DM patients to the associated rules with respect to the temporal patterns and sequence of occurrences.

4. Data mining in extracting the meaningful relationship among the associated complication rules.

5. Interpretation and evaluation for understanding the temporal phenotype, the discovered H, and co-occurrence of the complications pattern into knowledge.

Having considered this thesis objectives, this part mainly focuses on description of four goals listed below:

1. Knowledge Discovery and the design of computational efficient algorithms capable of strengthening existing diagnostic methods and guidelines for T2DM, IGT and IFG.
2. strengthening the diagnosis of patient populations with these metabolic diseases.
3. helping to determine the likelihood of having T2DM as well as its serious comorbidities.
4. the final objective of the study is to deploy resources that will lead to the advancement of disease management, along with the reduction of complications, by identifying efficient medical and behavioural strategies.

3.3.1 Pre-processing and Relational Models

In order to represent data, before modelling the clinical data, different pre-processing techniques are used. First, Relational Models in DBMS were used to to design a database and ensure that the data is understandable. DBMS was defined to create and maintain a database by using Data Definition Language (DDL). Then, Relational Algebra in the DBMS was employed to build one single table from integration of primary tables (five table were intuitively collected based on each complications individually) in the database. Furthermore, Relational Calculus (Structured Query Language (SQL) query) was used to formulate the definition of the joint table in terms of relationship among the primary tables. For example, it also employed Microsoft Access and SQL Server to store the collected data at Pavia. This was utilised to interact with the database, retrieve, manipulate and extract the useful information gathered from all preliminary tables by employing the Data Manipulation Language (DML). The uninformative and bias records (e.g., a patient with only one visit) were truncated to filter out unnecessary data.

Table 3.3: Percentage of Missing Values for T2DM variables at the first visit compared to all visits.[22]

Missing Values		
Variable	missing values in visit 1	missing values in all visits
Time to diagnosis (t)	2.4%	0.5%
Body Mass Index (BMI)	0.1%	1.1%
Glycated Hemoglobin (HbA1c)	16.9%	7.4%
Total cholesterol (Chol)	34.1%	44.9%
HDL cholesterol (HDL)	40.3%	48.1%
LDL cholesterol (LDL)	74.5%	77.9%
Triglycerides (TRG)	36.2%	12%
Smoking Habit (SMK)	0%	0%

3.3.2 Missing Values and Data Imputation

As explained previously, the missing data is a serious concern. This section clarify the strategies was employed to cope with missing details, issues with class imbalances and our model development decisions. In choosing the predictor variables to be used in these models, this study mainly focuses on the analysis of the literature mentioned earlier and found the variables which were usually related and accessible in the data with a significant risk of microvascular complications. Having obtained this, as can be seen in Table 3.3, there were several details missing from most of the measurements. For example, lipid-related data are indeed very likely to involves missing values (which is shown in Table 3.3). There are two options to deal with this issue: first, whether data are imputed or not and then, whether lipid based data is included or not.

In data imputation, the MOSAIC study tested two direct analytical techniques (i.e. the mean and median of each attribute) and one Random Forest technique to the data imputation strategy. This last idea is built on Stekhoven and Buehlmann's Random Forest imputation algorithm [115] and established as missForest.

To assess the efficiency of the imputation technique, only cases lacking missing data were taken into account. The entire set of statistics was then changed by deleting value records randomly. The rate of missing values in the initial collected data was in particular determined with each variable, and the same percentage was omitted arbitrarily again from collected data,

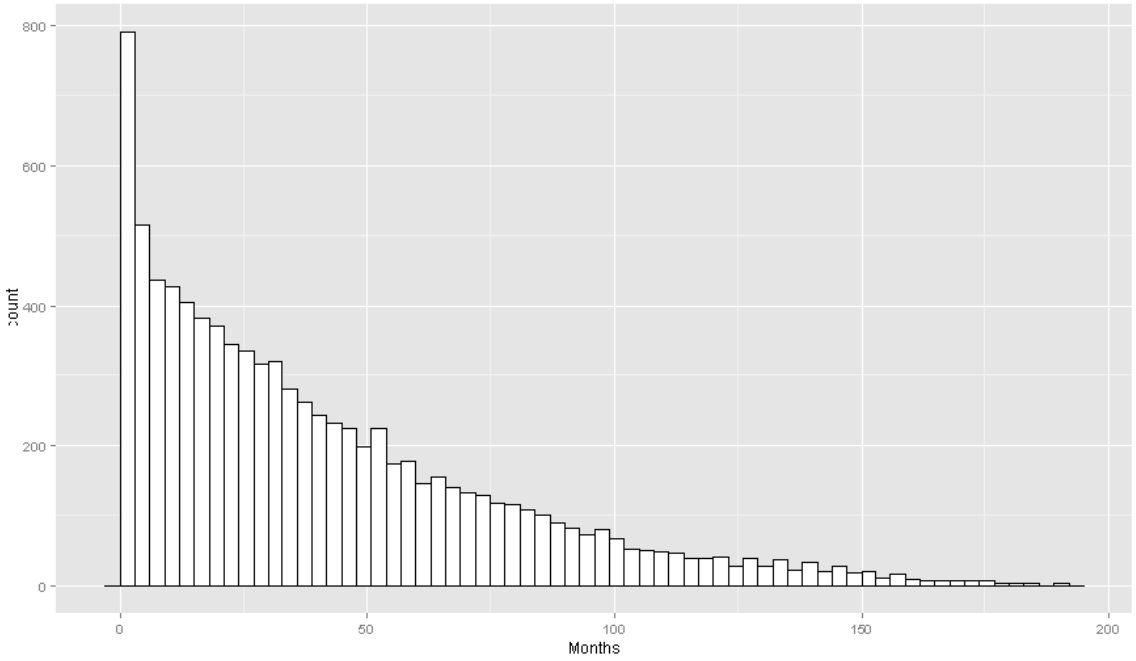

Figure 3.2: Time between follow-up, in months. [36]

thereby generating fictitious missing values to evaluate the ability of imputation

Median, mean and missForest imputed the deleted values. The criteria picked for missForest were 100 trees and a limit of 100 iterations. Then the efficiency of the imputation contrasted by calculating the average root squared error (RMSE) and the normalised average RMSEN error by the synthetic missing data. In order to measure RMSE, only numerical variables and risk factors with missing values was used. As seen in Table 3.4, MissForest outperformed the mean or medium imputation and was consequently preferred as the main tool in providing the imputation technique.

Missing values for smoking status were imputed assuming that patients do not change their smoking habit during their follow up period, otherwise by the most frequent value observed in the dataset (for patients for which the smoking status was missing at any visit). The initial formulation of HbA1c was presented in $mmol/mol$ and it was converted into percentage, while the subsequent formulation translated as HbA1c $\% = (0.0915 x HbA1c mmol/mol) + 2.15$.

Missing values for continuous variables (triglycerides, systolic blood pressure, body mass

Table 3.4: RMSE of mean, median and missForest on numerical features.[22]

RMSE						
	BMI	Hba1c	COL	HDL	LDL	Triglycerides
missForest	0.57	3.56	22.2	9.77	23.09	48.04
Mean	3.23	11.51	35.36	14.25	31.12	72.45
Median	3.23	11.81	35.36	14.37	31.14	74.47
RMSEN						
	BMI	Hba1c	COL	HDL	LDL	Triglycerides
missForest	0.01	0.03	0.07	0.04	0.08	0.05
Mean	0.09	0.11	0.12	0.06	0.11	0.08
Median	0.09	0.12	0.12	0.06	0.11	0.08

index, and total cholesterol) is imputed using the k-Nearest Neighbour, where the function was implemented in "$DMwR$" within the R package. Lack of smoking status values was apportioned considering during a follow-up visit in which the person may not switch their addiction, followed by its most frequently reported statistic (for those patients in T2DM population that the smoking status was missing at any of their visits). Continuous variables also were categorised in the discretisation algorithm into three stages as obtained in three percentiles of numerical series and considered as random effects. Similarly, smoking status was observed using representative variables with "never-smoker" becoming a low-risk group, whereas "ex-smoker" and "current-smoker" also were moderate and high-risk, respectively.

Another data imputation is when fewer than two visits appointments were reported, sufferers were omitted. Unless accompanied by at least 1 return during a 12-month period, single follow-up were exempt. Figure 3.2 illustrates the time period variability in the patient group among visits. In addition to this, Centre Profiling is aimed at assessing the hospital characteristics in terms of population (number of patients with complications, time to diagnosis of the complications) and of patterns of care (e.g. centres that are used to deal with more complex cases, centres that perform an initial intensive diagnostic program to discover complication early after the first visit).

Furthermore, in definitions of an increase risk over a 12-month period of each follow-up was used to conclude that HbA1c percentage was greater or equal to 0.6 as the dependent risk

factor. Whereas, it was classified and encoded to make of the existence of either two, one or zero. The behavioural characteristics such as time gap between follow-ups and Smoking habit, as well as clinical measures (e.g., BMI, Triglycerides and HbA1) were regarded as independent variables to be used in a T2DM evaluation and disease progression analyses. Equations 3.1-3.2 illustrate the time series representation of the analysis schema.

3.4 Classification and Imbalanced Data

To predict a target complication, patients are classified into two categories (cases): positive and negative cases. The outcome of the prediction or classification (Y) can be considered as a vector of disease risk factors represents by $Y = (X, C_i)$, where X is the vector of symptoms, and C_i shows a target complication class selected from $C = \{HYP, NEU, NEP, LIV, RET\}$. In this study, C_i only takes on binary values ($C_i = \{0 \mid 1\}$) as the main focus is to predict only one complication at time. For example, if a patient is diagnosed negatively (not having the complication), the class value becomes zero ($C_i = 0$) otherwise it sets to one ($C_i = 1$) in which it shows that a patient is diagnosed positively (having a target complication).

Considering a specific (target) complication at each time point, by detecting any "one" in the class over all patient visits is directed to join to the positive case otherwise the patient becomes a member of the negative case. This explains the "Patient-based" analysis. Once a patient has been identified as a positive case ($C_i = 1$), the patient stays in the corresponding case throughout their time-series. As a result, those patients who are already at a high risk of developing complications, it is assumed that they do not switch from positive case to the negative case. Alternatively, in "Visit-based" each time point (as a single visit for a patient) is scored individually as "zero" or "one" for each patient.

3.4.1 Re-balancing Strategy for the Time Series Complex Data

T2DM dataset is highly imbalanced based on common complications. Different learning techniques deal with imbalanced data, such as oversampling, undersampling, boosting, bagging, bootstrapping, and repeated random sub-sampling [62]. To be able appropriately re-balance

longitudinal design, this study suggest time series bootstrapping and pair-sampling methods considering cell arrays of patient follow-ups.

There are several aspects that might influence the performance achieved by Bayesian network learning. It has been reported that one of these aspects is related to a class imbalance in which examples in the training data belonging to one class (here negative cases) heavily outnumber the examples in the other class. In this situation, which is found in clinical data describing an infrequent but important event, the learning system may have difficulties in learning the concept related to the minority class (number of positive cases). In fact, the problem seems to be related to learning with too few minority class examples in the presence of other complicating factors, such as class overlapping.

In practice, T2DM data includes repeated measurements (in follow-up visits) multiple times for single patient. The biggest focus is to reflect the way in which the behaviour evolves over time as well as the risk factors. The key structures for longitudinal data processing are considered as marginal, combined and transition [32], which are challenging to be re-balanced. Imbalance class variable distribution is defined as the associated binary classification issues. One choice to deal with this issue is to disregard class imbalances and actually go through stages of studying and evaluation.

Another solution is to re-sample the instances or repeated measurements. To deal with unbalanced data, bootstrapping approaches are exploited to regenerate our observed time series visits per each patient. This method is utilised for inference by re-sampling and concatenating different pairs of visits. To provide this, the algorithm was conditioned on a newly structured dataset which is re-balanced by oversampling the minority class.

Therefore, this thesis employed "cell arrays" structure in MATLAB (as was seen in equation 3.1) which will help to address the above research gaps. In the proposed structure in pre-processing approach, each patient represented by a cell array of the visits. This was used to avoid any changes in the ordering of visits in time series analysis. In addition to this and to be able to keep the ordering of uneven number of visits for each patient, an appropriate structure of data was created. Thus, it constructs T2DM data as a cell arrays of patients, where a cell array represent all patients data. A patient cell array also includes a cell array of

Table 3.5: List of all complications of T2DM considered in the dataset.[22]

Complication	Type	Number of cases
Nephropathy (NEP)	Microvascular	121
Neuropathy (NEU)	Microvascular	126
Retinopathy (RET)	Microvascular	119
Fatty Liver Disease (LIV)	Not Vascular	227

the corresponding visits (a set of relevant patient visit across a time series). Thus, the final cell array of cell arrays (which was called T2DM dataset (DS)) is re-balanced by utilising two re-balancing techniques: the time series bootstrapping approach and pair sampling.

Table 3.5 represents the unequal range of complex patient populations while considering a complication was present. This table also illustrates the complications type and number of cases in the original Diabetes dataset. The vector of the binary class in the models refers to if the complexities are equivalent to the predefined threshold in a few years' time following the first visit. In particular, the previous probabilities of the class for BN are calculated on the initial imbalance collected data, whereas the marginal probabilities are determined in the training samples balanced with oversampling. In this approach, the latter probabilities of the experiment are modified to the current class distribution and can be used to readjust the model for the new patients.

The unbalancing ratio was defined here based on the raw natural unbalance rate of T2DM dataset. An unbalanced ratio was calculated as the ratio of negative to positive cases for a specific complication to ensure a balance. The unbalanced ratio of 3.2, 3.2 and 2.2 are used, in this study, for oversampling the positive cases that developed retinopathy, liver disease and hypertension, respectively.

3.4.2 Time Series Bootstrapping Approach

In this study, bootstrap approach is adapted to identify the significant statistics from classifiers learnt from such data where the occurrence of the positive class is far less than the negative. This is because Bootstrap re-balancing methods generally have been found to produce more accurate and reliable statistics [111].

Bootstrapping facilitates the acquisition by drawing the subsets from the measured data and calculating the statistics for each component of the subsets, of an unspecified feature of an uncertain distribution. Bootstrap also helps to achieve an estimated representation of the beliefs.

Having considered the temporal and complex nature of T2DM data, the bootstrap approach in the longitudinal dataset is extended by re-sampling consecutive time points, thus enabling the (first-order) to be inferred. The proposed oversampling Time Series Bootstrapping methodology is called "TS Bootstrapping" which employs a variant on the re-sampling approaches introduced in [85, 125, 130].

It re-samples the rare complication class with a replacement with respect to the dynamics of progression. The bootstrap pairs of time points are selected with replacement to ensure more states where the complication is present ($C_i = 1$) than in the original data. Thus, the re-sampling approach of the data involves a bootstrap process to re-sample observed time-series/visits of a patient with the replacement whereby the original training data is sampled in pairs of consecutive time points, t-1 and t. It also assumes that patient status at time t depends on the corresponding hidden variable at a previous time $t-1$ (Markov properties). As a result, the bootstrapped data contains an equal number of positive and negative cases for the target complication at time t. TS Bootstrapping approach was chosen as it seemed appropriate for T2DM dataset in the prediction of disease non-stationary models of data was difficult. Moreover, predicting rare cases in clinical data with an unbalanced distribution of a target complication was challenging, where common statistical methods such as standard regression is not appropriate. This is because it only models average score over the different structures throughout the time series. Another method is re-sampling, which can be applied on the learning data and trigger its distribution based on the bias in the data [85]. In the next section, the time series re-balanced data is analysed by DBNs learning models.

The final model is evaluated on a test set which preserves the original class quantity. The key drawbacks of this methodology are focused on the development of the final model for therapeutic use: a basic oversampling phase of the initial data should be preceded by the training of this model.

3.4.3 Evaluation Strategies for Re-balanced Data

This study is designed to provide a model of the prognosis for the major comorbidities of patients diagnosed with diabetes. Here the aim is to analyse the care received by patients with T2DM and specific comorbidities in the Pavia hospital, Italy. In addition to this, another aim is to build upon this thesis is conducting dynamical analyses on time series clinical data to improve classification accuracy. To obtain valuable results considering sensitive clinical data, this study suggests appropriate evaluation techniques to define various analysis strategies that rely on an initial assessment of the centre. Such assessment is aimed at understanding the hospital characteristics in terms of population (number of patients with complications, time to diagnosis of the complications) and of patterns of care (e.g. centres that are used to deal with more complex cases, centres that perform an initial intensive diagnostic program to discover complication early after the first visit). The purposed profiling centre is to estimate hospital features concerning patient behaviours (e.g. centres that interact with more complicated cases and carry out an initial intensive diagnostic process to find a complication in the event of the first visit. The proportion of patients with the complication, diagnostic periods of complications are used to analyse the overall data.

The data is separated to include the equal number of train cases and separate pairs of measures. To provide the train set indices, re-sampling approach with replacement was obtained and to obtain test set indices, the acquisition of those values which have not been re-sampled. The bootstrapping technique, first, splits the data into training data (to be fed into model of learning Bayesian network by using the model parameterization EM algorithm) and testing set (to be considered in the model validation). This procedure was performed 250 times, so that statistical validation (prediction performance calculation) was found in the model predictions. For each model the model output was evaluated as regards the amount of the squared error (SSE).

Non-parametric bootstrap (re-sampling by training set replacement, [49]) was then performed 250 times per modelling method to achieve significant statistical test results in the forecasts for each complication (number of iterations was found to be optimum through experimentation).

In order to assess the predictive model, appropriate validation analyses are conducted to predict the onset of T2DM complications (e.g., accuracy, sensitivity, specificity and precision by using 95 percent confidence interval). "Visit-based validation strategy" assesses the proposed predictive models based on the re-balanced time series train and test data. In Chapter 3 and Chapter 5, these models are created considering an unequal number of time series visits based on their regular follow-up. In the visit-based, an appropriate re-balancing strategy is adapted whereby the original training data is extended by adding bootstrapped pairs of time points. As a result, the bootstrapped data contains an equal number of positive and negative cases for the target complication at time t. Alternatively, "patient-based validation strategy" tests the model based on the re-balanced train data on different patient data as test set, which is randomly retrieved from the original T2DM dataset (mainly in Chapter 4 and Chapter 6).

3.5 Bayesian Networks

Various studies on longitudinal clinical data suggest an association between complications and risk factors of disease. In addition, there are still uncertainties around understanding the relationship between observed clinical data, complications, and unmeasured variables. Uncertainty is inherent in modelling data: there is uncertainty in the sampling of the data, in the parameters and structure of the models, and in the number of hidden variables. There is also uncertainty in the labels in supervised learning. It is possible for some physicians to believe that a certain diagnosis is accurate with a 100 percent certainty while another physician suggests those beliefs are valid with a zero percent certainty based upon the individual prior. A reasonable approach to tackle all of these issues is to take a Bayesian approach where a prior belief in a model is updated with data. Bayes theorem was invented by Thomas Bayes in 1763 and stated in Equation 3.4.

$$P(\theta|X) = P(\theta)\frac{P(X|\theta)}{P(X)} \tag{3.4}$$

where the Probability of success, network nodes (the observed data and prior) and parameters (the updated prior) are shown by X and θ, respectively. Knowledge or the state of

knowledge about hypotheses needs to be expressed through a probability distribution before data is observed and analysed. The hypotheses are evaluated by scoring the probability given data under the Bayes rule based likelihood. This rule comes from the basic rules of probability theory, which shows the way to combine the prior and the likelihood with multiplying them and re-normalising over the space of plausible hypotheses. For learning from the data, Bayes theorem tells us how to update the beliefs about the certain disease in the arrival of new evidence, but it cannot tell us how to set the prior beliefs (as shown in Equation 3.4). If one assumes that θ as the prior success, in the case study a posterior distribution represents the probability of a patient is diagnosed correctly by a disease/complication, which is shown in Equation 3.5.

$$\text{Posterior distribution: } P(\theta|X) = \frac{\theta L(\theta|X)}{\int P(\theta)L(\theta|X)d\theta} \quad , \tag{3.5}$$

Where $P(\theta)$ is known as the prior distribution in the prior model of $P(X|\theta)$ with a likelihood function of $L(\theta|X) \sim P(X|\theta)$.

In this study, uniform distribution estimates an uninformative prior. In the data cleaning process, the clinical test results are normalised and then discretised to prepare T2DM data for the Bayesian analysis.

A Bayesian Network consists of two components. The first is a Directed Acyclic Graph (DAG) with arcs (also referred to as links or edges) between nodes representing random variables in the domain. In particular, if there is an arc from node X to node Y in the BN, X becomes a parent of Y where Y is a child or descendant of X. Informally, a directed link between nodes $X \rightarrow Y$ indicates the existence of a direct influence from X on Y. The strength of this influence indicates whether the conditional probabilities quantify the directed link. The second component is a set of Conditional Probability Distributions (CPDs) associated with each node. The CPDs can be modelled by either a continuous distribution or with Conditional Probability Tables (CPTs) of discrete-valued variables. Bayesian approaches suggest a promising model to help clinicians to predict disease while there are still uncertainties around understanding the relationship between observed clinical data, complications, and unmeasured variables. In

biomedical science and clinical decision support, BNs have become a popular representation for dealing with uncertainty domain knowledge [39]. In particular, BNs can represent probabilistic relationships between complications and symptoms using Bayes theorem. They can be used to combine existing knowledge with data and interpreted for non-statisticians. They are probabilistic graphical networks that model longitudinal data considering noise, missing data and uncertainty in the data collection process.

3.6 Time-Series Probabilistic Models in Clinical Domain

Mining clinical data is a challenging task given the mixture of clinical test results (such as blood pressure, cholesterol level, etc), complication types (categorical, numerical, ordinal data types), unequal length of patients visits, highly correlated risk factors, unmeasured factors, heterogeneity, biased data, and more. With a probabilistic graphical model, it is easy to interpret and provide information regarding the qualitative structure of the clinical domain. For instance, it helps to find out whether risk factors are influencing factors of a disease or complication or if they are independent. Probabilistic models can be invaluable where we do not want too much reliance on the training data which risks overfitting with poor generalisation capabilities. Therefore, this thesis now describes some probabilistic graphical methods including the Hidden Markov Model and the Dynamic Bayesian Network (DBN), which ideally suits the problem of modelling complex clinical data from both qualitative and quantitative clinicians' point of view while handling uncertainty.

3.6.1 Hidden Markov and State Space Models

Hidden Markov Model (HMM) is a ubiquitous tool for modelling time-series data. The HMM obtains its name from several assumptions. Firstly, it assumes that the observation at time t is generated by several processes where a state (S_t) is hidden from the observer. Secondly, it assumes that the state of the hidden process satisfies the Markov property that given the previous state (value of S_{t-1}). It also declares that the current state of S_t is independent of all other states prior to $t-1$. In other words, the state at a certain time encapsulates

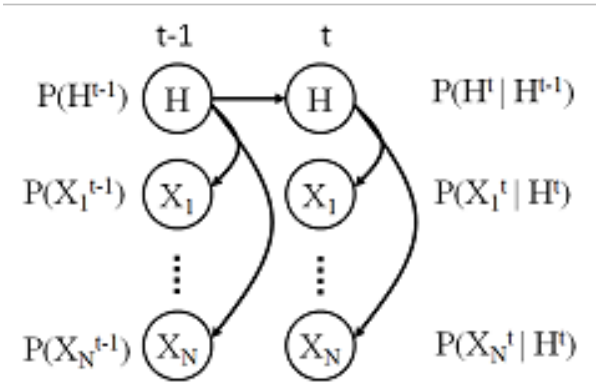

Figure 3.3: Space State Model illustrates the interactions among a hidden factor as a H and the observed nodes (X_i) in two time-series. The max number of patients is shown by N ($p = N$)

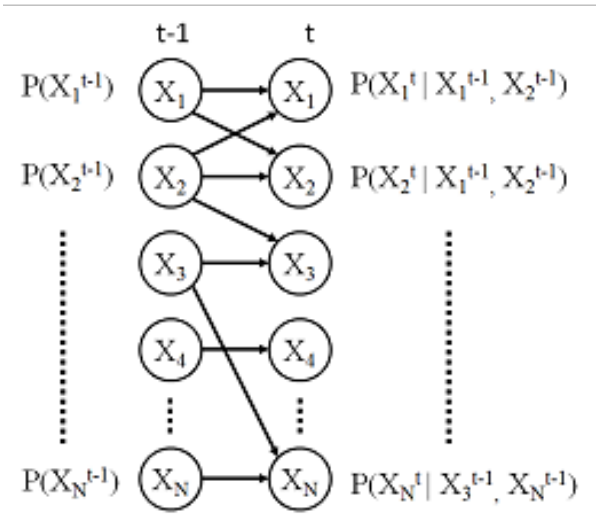

Figure 3.4: The Space State Model with dynamic interactions among time-series nodes.

all we need to know about the history of the process in order to predict the future of the process. Thirdly, HMM is generalised by representing the state using a collection of discrete state variables, illustrated in Figures 3.3,3.4 and stated in Equation 3.6, each of which S_t can take on $K^{(m)} = M$ values donated by the integers $1, ..., k$ [127].

$$S_t = S_t^{(1)}, ..., S_t^{(m)}, ..., S_t^{(M)} \tag{3.6}$$

Finally, to define a probability distribution over sequences of observations all that is left to specify is a probability distribution over the initial state of $P(S_1)$, and $K \times K$ state transition matrices and output models which are not considered to be dependent on t as the model is time-invariant (except for the initial state). If the observable variables are discrete symbols taking on one of H values, the output model can be fully specified by a $K \times H$ observation (or emission) matrix.

Nevertheless, several inference problems are associated with hidden Markov models. For example, the probability of observing a sequence of unequal patients' visits is an inference problem that is associated with HMMs as each patient has an unequal number of visits. To calculate this probability, all possible state sequences are aggregated given the parameters of the model. Assigning a time index t to each variable, one of the simplest causal models for a sequence of the patient visit $Y_1, ..., Y_T$ represents the first-order Markov model, in which each variable is directly influenced only by the previous variables as represented in Equation 3.7.

$$P(Y_{1:T}) = P(Y_1), P(Y_2 \mid Y_1), ..., P(Y_T \mid Y_{T-1}). \tag{3.7}$$

Where it satisfies a Markov property and S_t, Y_t are independent of the states and observation at all other time indexes. Taken together, these Markov properties mean that the joint distribution of a sequence of states and observations can be factored in the following way (see Equation 3.8).

$$(S_{1:T}, Y_{1:T}) = \frac{P(S_1)P(Y_1 \mid S_1)}{\prod_{t=2}^{T} P(S_t \mid S_{t-1})P(Y_t \mid S_t)} \tag{3.8}$$

Having observed $Y_1, ..., Y_T$, the first order Markov model will only make use of Y_t to predict

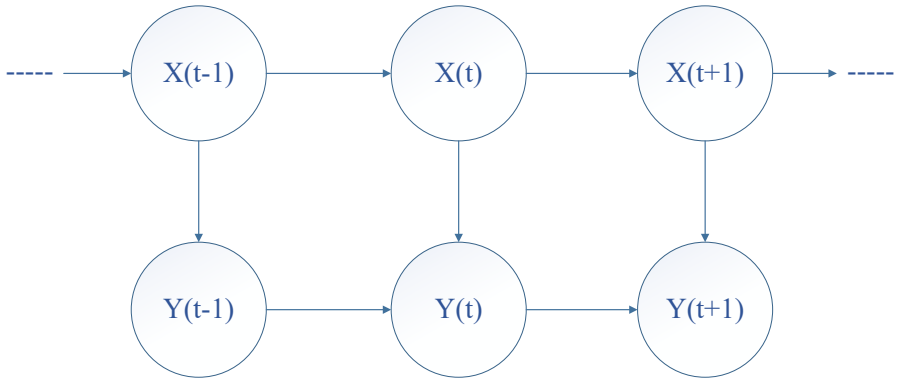

Figure 3.5: The dependency graph of HMM.

Y_{t+1}. One simple way of extending Markov models is to allow higher order interactions between variables, for example, n$^{\text{th}}$ order Markov model allows arcs from $Y_{t-n}, ..., Y_{t-1}$ to Y_t. Another way to extend Markov models is to posit that the observations are dependent on a hidden variable which is known as a state among the sequence of the states in a Markov process. This state of a Markov process is not directly observable, and it is often hidden, or partially observable. In state-space models shown in Figures 3.3-3.4, a sequence of K-dimensional real-valued observation vectors $X_1, ..., X_T$ is modelled by assuming that each time step (Y_t) is generated from a K-dimensional real-valued hidden state variable H_t. Additionally, a sequence of $X(H_{1:T} = H_1, ..., H_T)$ defines the first-order Markov process calculating the likelihood of H in Equation 3.9:

$$P(H_{1:T}, Y_{1:T}) = P(H_1)P(Y_1 \mid H_1) \prod_{t=2}^{T} P(H_t \mid H_{t-1})P(Y_t \mid H_t). \tag{3.9}$$

For real-valued observation vectors, $P(Y_t \mid S_t)$ can be modelled in many different forms, such as a Gaussian, mixture of Gaussian, or a neural network [89]. For high-dimensional real-valued observations, a beneficial output model is obtained by replacing the Gaussian by a factor analyser. Factor Analysis (FA) is a method for modelling correlations in high-dimensional data and is closely related to principal components analysis (PCA). The factorisation of the joint probability means that BN for a State-Space Model is identical to the HMMs except that

the hidden variable S is replaced by X. This factorisation of the joint probability can be drawn graphically in the form shown in Figure 3.5. This graph is also known as a BN, belief network, probabilistic graphical model, or probabilistic independence network. It represents the dependencies among variables in the model. A node in the graph represents each variable, whereas each node receives directed arcs from another node. Therefore, these nodes within the network are conditionally dependent on the factorisation of the joint distribution of the related nodes.

Boyan in [20] stated, HMMs suffer from important limitations when it comes to modelling real-world time-series data. If the state transition matrix is unconstrained, any arbitrary nonlinear dynamics can also be modelled. It is assumed that an HMM captures the underlying state space by using M different K-dimensional variables. Thus, the HMM requires $K^{(M)}$ distinct states to model the system. This representation is not only inefficient but difficult to interpret. More seriously, an unconstrained HMM with $K^{(M)}$ states has of order $K^{(2M)}$ parameters in the transition matrix. Unless the dataset captures all these possible transitions or a prior knowledge is used to constrain the parameters, severe overfitting may result. Taking into account Y_{t-1} while predicting Y_t as well as H_t aims to relax HMM assumptions and results in an auto-regressive HMM (AR-HMM) with a higher likelihood (see Figure 3.5).

3.6.2 Dynamic Bayesian Networks

To consider different aspects of medical domain knowledge, which are of a causal, complex, incomplete and temporal nature, an extension of the standard BN, known as the Dynamic Bayesian Network (DBN), is required to model temporal processes [89]. DBNs are belief networks that represent the stochastic process of a set of random variables. Instead of answering the question of importance by using either qualitative or quantitative methodologies, DBNs use a mixture of both. As a result, building a DBN requires two different parts: a quantitative specification for learning local and cross-correlations (in terms of conditional probability distributions) and a qualitative specification for the structure definition. The DBN extends BNs to model a set of temporal random variables by using probability and joint distributions with explicit time-stamps. See Figure 3.7 for an example of a DBN structure for diabetes

data. Notice that links can exist from any of time points to the next time points (where the conditional distribution is defined by $P(Z_t|Z_{t-1})$) in the Equation 3.10 or within the same time point defined by $P(Z_t|Z_t^i)$.

$$P(Z_t|Z_{t-1}) = \prod_{i=1}^{N} P(Z_t^i|Pa(Z_t)^i) \text{ for all } t > 1, \tag{3.10}$$

where Z_t^i could be i$^{\text{th}}$ observed, hidden or outcome node with a parent of $Pa(Z_t)^i$. In a DAG of a first order Markov model (the case study), a parent can either be in the same time slice or in the previous time slice.

$$P(Z_{1:T}) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(Z_t^i|Pa(Z_t^i)) \tag{3.11}$$

In Figure 3.5, the variables within the DBN has been classified into different categories: Hidden variables (H), Complication variables (C) and observed variables (O), which are known also as disease risk factors.

Unlike HMMs, DBNs provide a more powerful approximation of inference as well as a reduction in the number of parameters [20]. Another very well-known model in this class is the linear-Gaussian state state-space model, also known as the Kalman Filter Model (KFM), which can be thought of as the continuous-state version of HMMs. Murphy in [89] claimed that HMMs and KFMs can be considered as the least complex DBNs. Murphy also argued that, although HMMs and KFMs are simple and flexible, they are restricted in their "expressive power" [89]. Therefore, DBNs are introduced as an extended and generalised form of HMMs and KFMs that allow arbitrary probability distributions.

In time-series modelling, the assumption that an event can cause another event in the future, but not vice-versa, simplifies the design of the DBNs in which directed arcs should flow forward in time (as shown in Figure 3.5). Learning DBN structures provides a principled mechanism to detect conditional dependencies in time-series data. However, to analyse the likelihood of one occurring complication, relationships between variables can change over time (in non-stationary processes). In order to deal with these, a new class of graphical model called a non-stationary DBN is used [23] in which the conditional dependence structure of the

underlying prediction process is permitted to change over time (see Figure 3.6).

3.7 Hidden Variables and Causal Structure Discovery of Bayesian Networks

Predicting a disease is non-trivial since clinical data is incomplete and often contains unmeasured factors. Clinicians cannot measure all risk factors and carry out all kinds of tests, so there are some unmeasured factors that clinicians fail to measure, which need to be discovered at the early stage of the disease. In machine learning literature, there has been attention in developing the network structure in the appearance of latent/hidden variables in probabilistic models such as DBNs. Elidan and co-authors in [46] noted that networks without considering the impact of hidden variables are clearly less useful because of the increased number of edges needed to model all interactions among risk factors.

More importantly, the marginalised network without a hidden variable needs more parameters that cause substantial data fragmentation as well as non-reliable parameter estimations. A Hidden variable can capture the variability of parameters by selecting a set of parameters and considering them as random variables in the state distribution. Alternatively, it treats the parameter variability as a random variable in discrete state-space models. An H in a DBN is shown by the state distribution of a set of random variables, H_t^1, ..., $H_t^{N_h}$. By contrast, the H in an HMM is shown as a random variable in the state space model. The Hidden variable discovery method provides one solution for learning data in which it can be an exceptionally distinctive approach to scan for "structural signatures" of the hidden factor substructures. This enhances the comprehension of the disease progression and as a preliminary remark helps to learn an explainable AI model towards the promising needs.

To discover probabilistic dependencies, including latent variables, we need to search the space of belief networks and possible hidden variables. It may not be possible to decide precisely where and whether latent variables are present. For example, the FCI and IC* algorithm [96, 114] are used to identify where potential latent variables exist based on conditional independence tests. These are explained in more detail in Chapter 4.

The causal structure of a BN consists of two components. The first is a Directed Acyclic Graph (DAG) with arcs (also referred to as links or edges) between nodes representing random variables in the domain. In particular, if there is an arc from node X to node Y in the BN, X becomes a parent of Y where Y is a child or descendant of X. Informally, a directed link between nodes $X \rightarrow Y$ indicates the existence of a direct influence from X on Y. The strength of this influence indicates whether the directed link is quantified by the conditional probabilities. The second component is a set of Conditional Probability Distributions (CPDs) associated with each node. The CPDs can be modelled by either a continuous distribution or with Conditional Probability Tables (CPTs) of discrete-valued variables.

3.8 Standard DBNs Model in Predicting T2DM Complications: A Case Study

Here, to infer a DBN structure, a predefined latent variable is fixed, and other links are learned over the T2DM complications based upon the K2 and REVerse Engineering ALgorithm (REVEAL) [73]. The structure is illustrated in Figure 3.7. The DBN captures the disease process over time. A choice of two potential observations are evaluated in the model: either a patient having the disease/complication or not. We want to predict the P(Complication| Risk Factors). An application of Bayes theorem to predict T2DM complications is defined in Equation 3.12 which represents the likelihood of having a target complication given T2DM risk factors as prior.

$$P\ (\text{Complication} \mid \text{Risk Factors}) \approx \frac{P(\text{Risk Factors} \mid \ \text{Complication})\ P(\text{Complication})}{P(\text{Risk Factors})} \tag{3.12}$$

For these experiments this thesis wishes to capture the trajectories of the T2DM patients with two assumptions: firstly the Markov assumption that the future is independent of the past given the present (so there are only two time slices within the DBNs); secondly that the

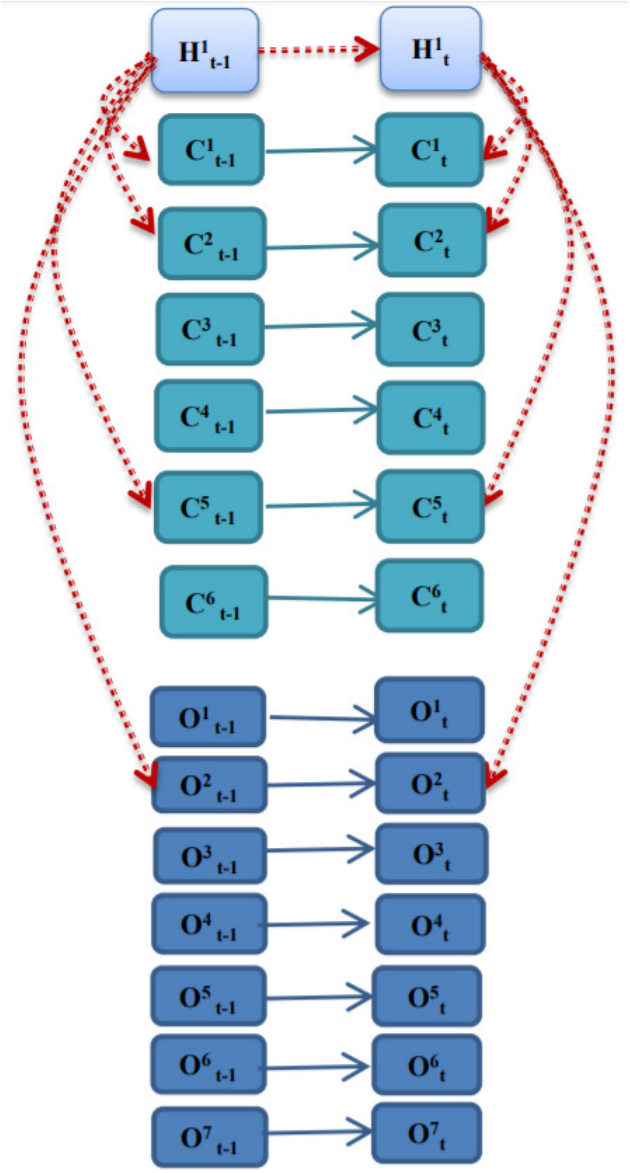

Figure 3.6: Two time-series structures using the K2 approach to identify the links from hidden nodes to other features and fully Auto-Regressive dynamic links. The H, C, and O illustrate Hidden, Complication, and Observed nodes, respectively.

processes involved are non-stationary (therefore $P(Y_t \mid X_{t-1})$ can possibly vary for some $t-1$. The learned DAG for T2DM data in Figure 3.5 demonstrates dynamic conditional independence as well as the joint distribution of the domain representing probabilistic relationships among the risk factors (shown by arrows in green) and complications (shown by arrows in red).

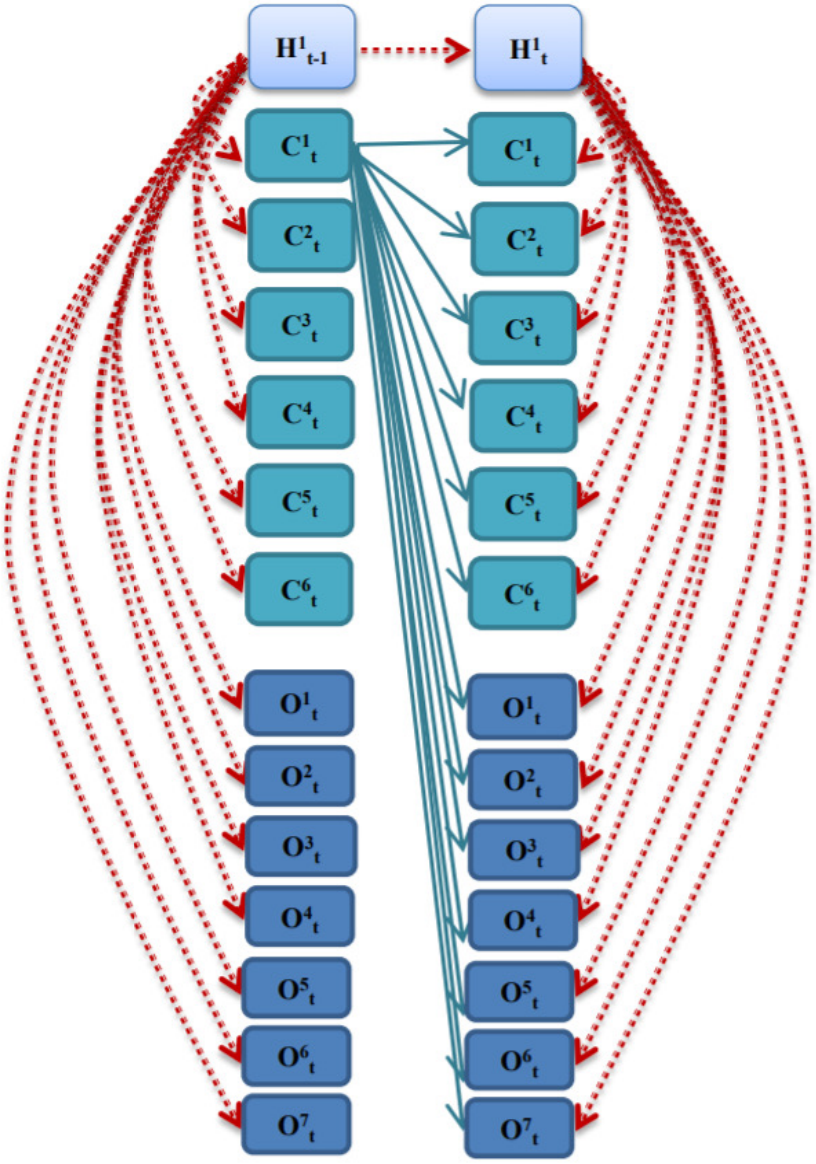

Figure 3.7: Two time-series structures using the K2 approach and dynamic links, which are learned from the REVEAL algorithm. The H, C, and O illustrate Hidden, Complication, and Observed nodes, respectively. The hidden variable is pointing to all complications and observed nodes.

For learning the parameters, the gradient ascent is suitable in the situation in which network structure is known. Otherwise, EM is appropriate while some variables are not observed (is hidden) [17]. Here, EM is performed to learn the DBNs parameters, including the hidden variable that is fixed to be connected to all data points. It is envisaged that this latent process

Table 3.6: ROC Statistics for the sensitivity analysis carried out on DBNs inferred on the original imbalanced data (IB) and on the balanced bootstrapped data (BBS).

Result	Lipid Metabolism	Liver Disease
IB AUC	0.5227 +− 0.04	0.6518 +− 0.05
BBS AUC	0.5809 +− 0.04	0.7141 +− 0.04
IB Sensitivity	0.775	0.996
BBS Sensitivity	0.855	0.891
IB Specificity	0.394	0.0
BBS Specificity	0.178	0.373

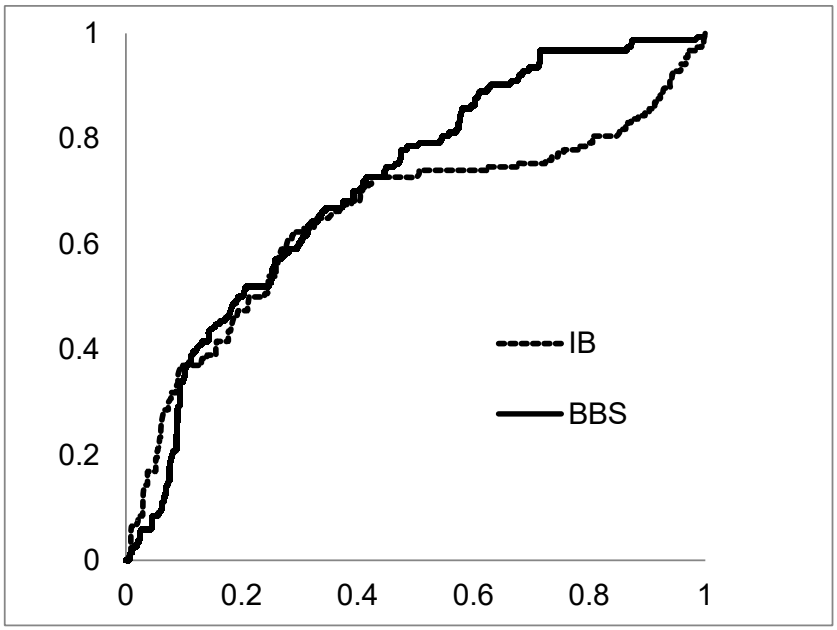

Figure 3.8: AUC Comparison of Liver Disease for the sensitivity analysis carried out on DBNs inferred on the original imbalanced data (IB) and on the balanced bootstrapped data (BBS).

will capture some of the complex dynamics of the comorbidities and how they interact with the clinical variables. The conditional distribution over the state in the model is shown at time t given the variables at time $t-1$ (as illustrated in Figure 3.5).

The DBNs were trained and tested using the TS Bootstrapping approach described earlier in this chapter. The test set was used to test the models' ability to predict each complication at the following time-point before the latent variables were explored for improving these statistics. The dataset for this experiment covers approximately 1000 patients and focuses on two comorbidities: disorders of lipid metabolism and non-alcoholic chronic liver disease. These

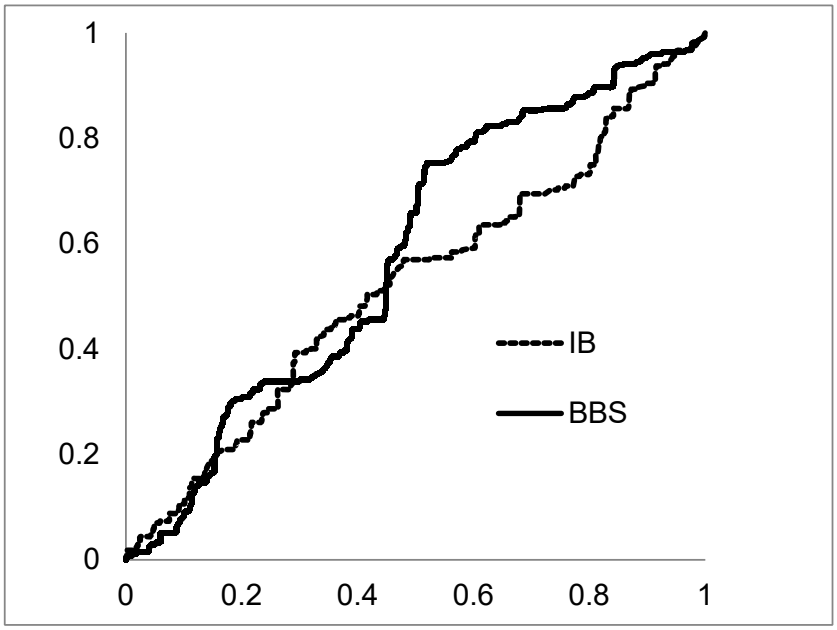

Figure 3.9: AUC Comparison of Lipid Metabolism for the sensitivity analysis carried out on DBNs inferred on the original imbalanced data (IB) and on the balanced bootstrapped data (BBS).

were selected as they seem to occur most commonly within the data.

As a results for assessing the prediction performance of (Liver disease and Lipid metabolism), the re-balanced data using Time series Bootstrapping (BBS) was compared to Imbalance data (IB) in Figures 3.8-3.9. In these figures, Receiver Operating Characteristics (ROCs) was used to measure the sensitivity and specificity of the prediction of these two complications and were illustrated at the subsequent visit. These were based on DBN models that applied on the original imbalanced data and DBNs that were trained on the bootstrapped time-series data. As can be seen in Figures 3.8-3.9, the resulting ROCs varied dramatically for detecting the false positive - whilst both methods are very similar for the smaller number of true and false positives (bottom left of the ROCs), the DBN results trained on the original data are much closer to random (on the diagonal) for more significant numbers of false positives, whereas this issue did not occur in the bootstrapped data. similarly, Table 3.6 compared the Area Under the ROC Curve (AUC) for the prediction performance of re-balanced data using Time series Bootstrapping (BBS) versus Imbalance data (IB), the sensitivity and specificity all reflected this issue. "IB" curve in both Figures 3.8-3.9.

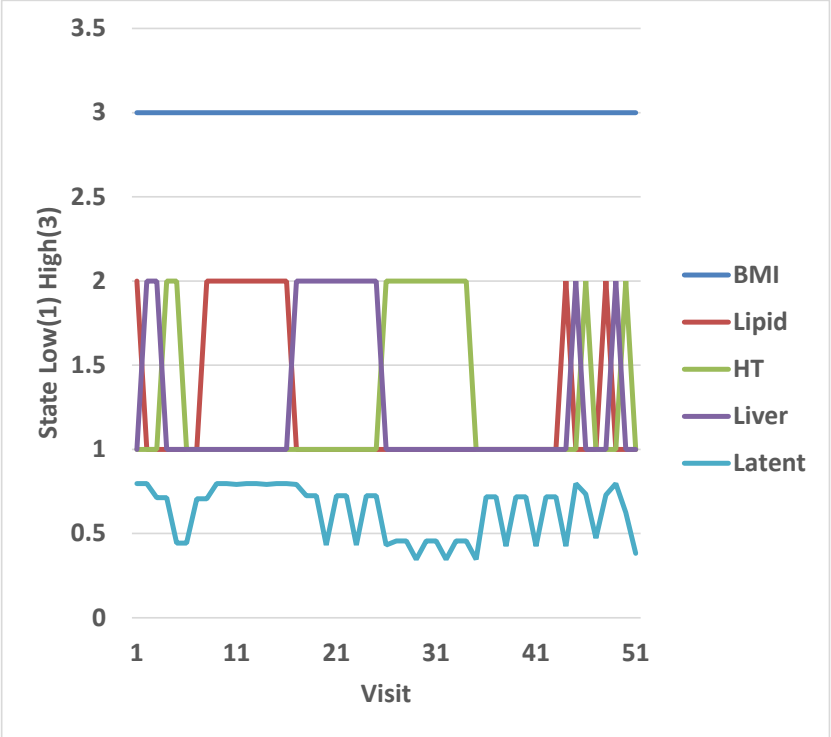

Figure 3.10: Latent variable examples for a randomly chosen patient based upon the DBN inferred from the bootstrapped data to overcome the class imbalance.

Looking at how the latent variable has behaved in the modelling illustrates a form of refactoring of the complication data has occurred. The latent variable appears in many cases to have captured a combination of clinical factors and complications where an increase in the probability of latent state one coincides with the increased likelihood of complications (particularly Lipid Metabolism and Liver Disease complications). To a lesser degree, it is associated with a change in BMI. Figures 3.10-3.11 show in two sample patients how the probability of latent state one is correlated with these factors but with a time shift one visit earlier, highlighting how it can enhance prediction.

Overall, the approach of re-balancing patients' visit data allows us to explore the temporal nature of how various complications progress while the data is unbalanced. The proposed visit-based temporal bootstrap approach identifies intermediate stages in T2DM process, and associated complications are exhibiting a variety of risk factor profiles subtly.

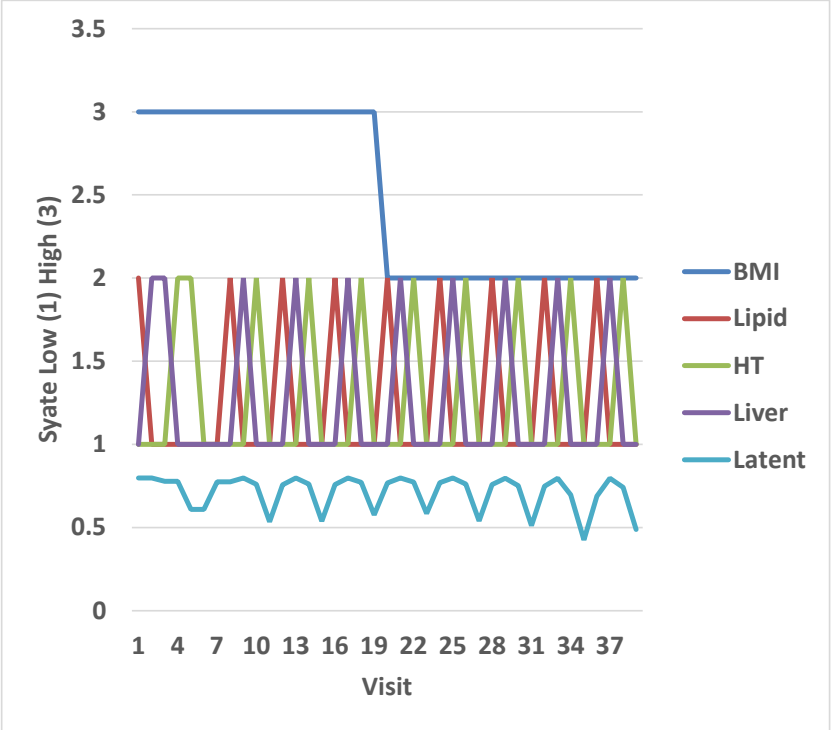

Figure 3.11: Latent variable examples for a randomly chosen patient based upon the DBN inferred from the bootstrapped data to overcome the class imbalance.

3.9 Summary

This chapter explored the key T2DM dataset and some key methods for modelling complex time-series. In particular, it has explored the use of DBNs with a single latent variable for prediction and the early detection of T2DM complications. The model was used to compute the probabilities of the presence of time-series comorbidities, given a set of risk factors. This chapter has also explored the combination of re-sampling to remove the bias of imbalanced data in time-series with a latent variable DBNs model and predict the onset of complications associated with a disease. The proposed Bootstrapping method was applied to the observed visits per patient in the pairs of consecutive time points to balance positive and negative patient cases. Additionally, it discussed how the explicit consideration of Bayesian models contributes to improved modelling of T2DM features. Results indicate that the re-sampling procedure could assist in the predictions, and the latent variables could also factorise the data into an

underlying hidden state that improves the prediction accuracy. The following chapters will involve exploring the extension of these models with more latent variables to capture a greater variety of factors that characterise key changes in the clinical and complication data. In the next chapter, multiple hidden variables are used to help identify different cohorts of patients who have different dynamics.

Chapter 4

Learning Multiple Hidden Variables

4.1 Introduction

The previous chapter attempted to understand the effect of one single unmeasured variable on the prediction of T2DM complications using the standard DBNs framework. As discussed in Chapters 2-3, it is becoming challenging to ignore the existence of numerous hidden variables. This chapter seeks to capture more hidden variables and classify patients by the complications, thus extending the framework. Discovering multiple hidden variables plays an important role in implementing better predictive models while differentiating patient subgroups. Unlike deep learning approaches that attempt to model complex interactions in data by using sometimes huge numbers of hidden variables, this chapter develops a novel algorithm that iteratively adds hidden variables to a DBN structure. In doing so it identifies the correct number of hidden variables, and targets their precise location within the network structure. This chapter is organised as follows: Section 2 provides an algorithmic framework for dependency discovery with latent variables in a predictive model of T2DM comorbidities. Section 3 introduces a method for addressing imbalanced binary complications. Section 4 introduces the stepwise methodology in which hidden variables are added iteratively. Section 5 evaluates and compares state-of-the-art methods for inducing meaningful structures in each iteration of the process. Section 5 discusses the experimental findings and their significance. Section 6 concludes and provides a direction to the next chapter where the discovered hidden variables and groups of

patients are evaluated.

4.2 Discovering Multiple Hidden Variables

There are likely to be many unmeasured effects that impact the disease progression of different patients. Unlike deep learning approaches that attempt to model complex interactions in data by using very large number of hidden variables in layers, this research adopts a different approach. As discussed in the previous chapter, probabilistic graphical models such as DBNs have demonstrated much promise in the transparent modelling of disease progression, and they can naturally incorporate hidden variables using the EM algorithm. In this section, the previous DBNs model is extended to incorporate a stepwise hidden variable structure learning process that incrementally adds hidden variables based on the IC* algorithm. This section discusses the pre-processing and re-balancing techniques and then looks at causality and dependency constraints among the disease risk factors.

4.3 Learning from Imbalanced Data using Pair-sampling

The prediction strategy is based on analysis of clinical test results to identify groups of patients who have been diagnosed with T2DM and are also at risk of developing complications associated with diabetes. As it was discussed in Chapter 3, since the clinical data is highly imbalanced, trust in the outcomes of the predictive model can not be maintained. Thus, to find hidden variables from the imbalanced clinical data, it is first necessary to re-balance the data. The previous chapter showed how it could carry out the experimental demonstration of the latent factors within time-series re-balanced clinical data. The re-balancing technique was based on the bootstrap approach. However, in this section a pair-sampling strategy is introduced to effectively address imbalanced data before learning hidden variables.

The pair-sampling method divides the dataset of T2DM patients can be divided into two types of cases, positive and negative. A positive case ($C_i = 1$) consists of the patients who have been reported positive in complication C_i at the subsequent visit. In contrast, a negative case ($C_i = 0$) is characterised by the binary diagnostic of the complication associated with

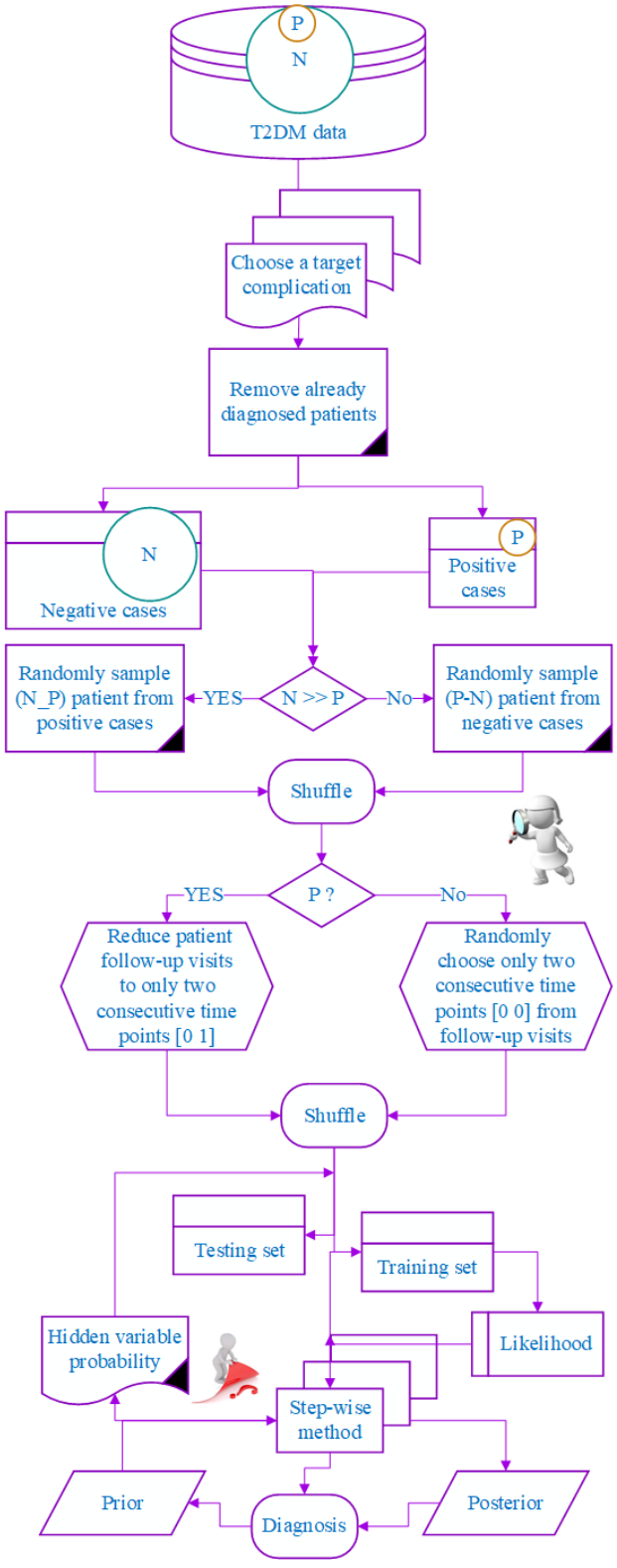

Figure 4.1: Diagram of Pair-sampling and the Stepwise IC* approach.

patients at lower risks in the second visit (V_{i2}). This is because there can not be a cure for ith patient in their first visit (V_{i1}) at high risk of developing complications as it has been reported that a patient with T2DM complications does not recover once diagnosed. Once a T2DM complication is diagnosed, it is recorded for the rest of the patients time-series. Therefore, the patient follow-ups are neutral, where the class value is consistent. This patient is known as a pre-diagnosed patient and needs to be removed before the prediction process, that has been already tested positive at the first visit. They are considered uninformative or biased samples.

An unbalanced ratio is calculated as the ratio of negative to positive cases for a specific complication to ensure a balance. Since the patient samples were highly imbalanced ($N \ll P$), the number of the positive cases (rare) was increased to the number of negative cases subtracted by the positive cases (N - P).

From the re-balanced dataset, train and test sets are generated. Half of the positive and negative cases are allocated to the test set while the remainder is assigned to a train set. Now data is partitioned into two halves of patient samples within each partition containing an equal number of samples for the target complication. Then, for each patient a pair-visit of [V_{i1} , V_{i2}] is defined, in which V_{ij} indicates that targeted complication which has not been diagnosed ($C_i = 0$). At the same time, V_{i2} represents the subsequent visit, which can represent either $C_i = 0$ or $C_i = 1$. For each of the un-diagnosed patients with a particular complication, there are two consecutive random visits which are illustrated by a pair-visit [V_{ij} , V_{ij+1}] to represent a switch from zero to one (pair-visit = $[0, 1]$). This method helps to be able to conduct early diagnosis, which can lead to appropriate management of these potentially serious complications.

4.4 Bayesian Networks and Latent Structures of Stepwise IC* Algorithm

Bayesian Networks represent the joint probability distribution over a set of variables (e.g., risk factors). These variables are represented as nodes within the Bayesian graphical structure and directed connections between these nodes capture independencies among them. This study assumes that positive diagnosis (the likelihood of developing a complication) must be greater

than a threshold of 0.1. The IC* algorithm provides a procedure to determine which causal connections among nodes in a network can be inferred from empirical observations, even in the presence of latent variables. IC* is a constraint-based method which applies conditional independence analyses to infer causal structures and learns a partially-oriented DAG with latent variables [114]. This algorithm can be used to analyse effective connectivity among T2DM risk factors. Here, a new methodology is proposed, which combines the basic principles of the IC* algorithm to obtain a DAG in addition to a dynamic process that is inferred using the REVEAL algorithm. A stepwise approach is used to incrementally add latent variables. Based on Pearl's Causality, a latent structure is a pair $\ell = < D, O >$ where D is a causal structure over V and where O is a set of observed variables. In general, the constraints that a latent structure imposes upon the distribution cannot be entirely characterised by any set of conditional independence statements. In the absence of hidden variables, tests for equivalence can be reduced to tests of induced conditional dependencies. More importantly, for every latent structure ℓ, there is often a dependency equivalent latent structure, the projection of ℓ on O, in which every unobserved node can be a root node with exactly two observed children. Each independence equivalence class is graphically represented by a pattern (PDAG), in which directed edges represent arrows that are common to every member in the equivalence class. The edges that are lacking direction represent ambiguous relationships: they are directed one way in some equivalent structures and another way in others. The IC* algorithm returns a marked pattern in a partial DAG in the form of a matrix that represents four types of edges over the variables:

1. a marked arrow $O_1 \overset{*}{\rightarrow} O_2$, signifying a directed path from two observed nodes (O_1 and O_2) in the underlying latent structure (and there is no latent common cause for these two nodes);

2. a bi-directed edge $O_1 \leftarrow \ell \rightarrow O_2$, signifying a latent common cause of two observed nodes (O_1 and O_2) in the underlying latent structure, or an inducing path between two variables; thus there is no directed path between them;

3. an unmarked arrow $O_1 \rightarrow O_2$, signifying either a directed path from O_1 to O_2 or a

bi-directed edge; and

4. an in-directed edge $O_1 -- O_2$, standing for either $O_1 \rightarrow O_2$ or $O_1 \leftarrow O_2$ or $O_1 \leftarrow\rightarrow O_2$.

For example, a "-2" in the IC* matrix represents a marked (*) arrow from node a to node b in the underlying latent structure and at the same time shows there is no latent common cause for those two nodes. There could be a marked link between liver and smoking habit, and there is no latent variable between these two nodes. Similarly, if it represents either of "-1" or "2", it can be inferred that there is a latent variable between these two nodes. A bidirectional edge is signifying a latent common cause in the underlying latent structure without a directed edge between two nodes.

4.4.1 Stepwise IC* Algorithm

An incremental strategy is employed in order to enhance the stepwise IC* algorithm, which is called stepwise IC*. A diagram showing the process of the pair-sampling and the stepwise procedure is presented in Figure 4.1. In this method, IC* is applied to a dataset. The probability of a high state of any learned latent variables at the current step is then inferred using the EM algorithm. The inferred probabilities of the hidden variable are treated as observations which means that it can be then treated as an observed variable in the subsequent step. In this next step, IC* is applied again to see if the new observed variable uncovers any new hidden variables. This is repeated until no other hidden variables are discovered. For example, a re-balanced dataset based on class values of retinopathy is provided using the Pair-sampling. The model is then trained using the structure obtained from the balanced dataset at the first step, to find a hidden variable. At the next step, there are 14 observed variables, including 13 different T2DM features plus one additional hidden variable probability that was inferred in the previous step. Then the structure is learned from this new 14 variable dataset, and a new hidden variable is discovered. Next, the prediction probability of the hidden variable is retrieved to generate another observed feature in the third version of the dataset. Later this obtained dataset is used to train and test the next step. Furthermore, the new hidden variable is pointing to neuropathy, HbA1c, liver disease, smoking, and BMI (see Figure 4.2). This

process is continued until the IC* algorithm is not able to find a new hidden variable, or there may not be any significant improvement in the prediction performance. In the next section, the use of the stepwise IC* approach to learning multiple latent variables was assessed concerning improvements in disease prediction.

4.5 Experimental Results

In order to evaluate the proposed approach, three T2DM complications with high prevalence in elderly patients are monitored. Then, a set of experiments was performed to compare the different stage of the stepwise IC* approach considering which step could predict complications with better prediction accuracy for the complications. It also was compared to the model with no latent variable. Furthermore, the experimental result indicates how by limiting the number of latent variables enabled a clearer understanding of their effects.

4.5.1 Understanding Hidden Variables

The causal discovery of BNs is a critical research area, which depends on looking through the space of models for those which can best clarify a pattern of probabilistic conditions in the data [146]. The causal discovery indicates dependencies that are generated by structures with unmeasured factors, i.e., latent variables. Hidden variable discovery in causal structures has been introduced in [113]. Latent variable models have a long tradition in causal discovery. Factor analysis and related methods can be used to position latent variables and measure their hypothetical effects. However, many do not provide clear means of deciding whether or not latent variables are present in the first place. As was discussed in Chapters 2-3, causal discovery methods in AI have many advantages [113]. One advantage of a latent variable is that they can better encode the actual dependencies and independencies in the data. For example, Figure 4.2 demonstrates a latent variable of 13 observed variables and one latent variable in T2DM data.

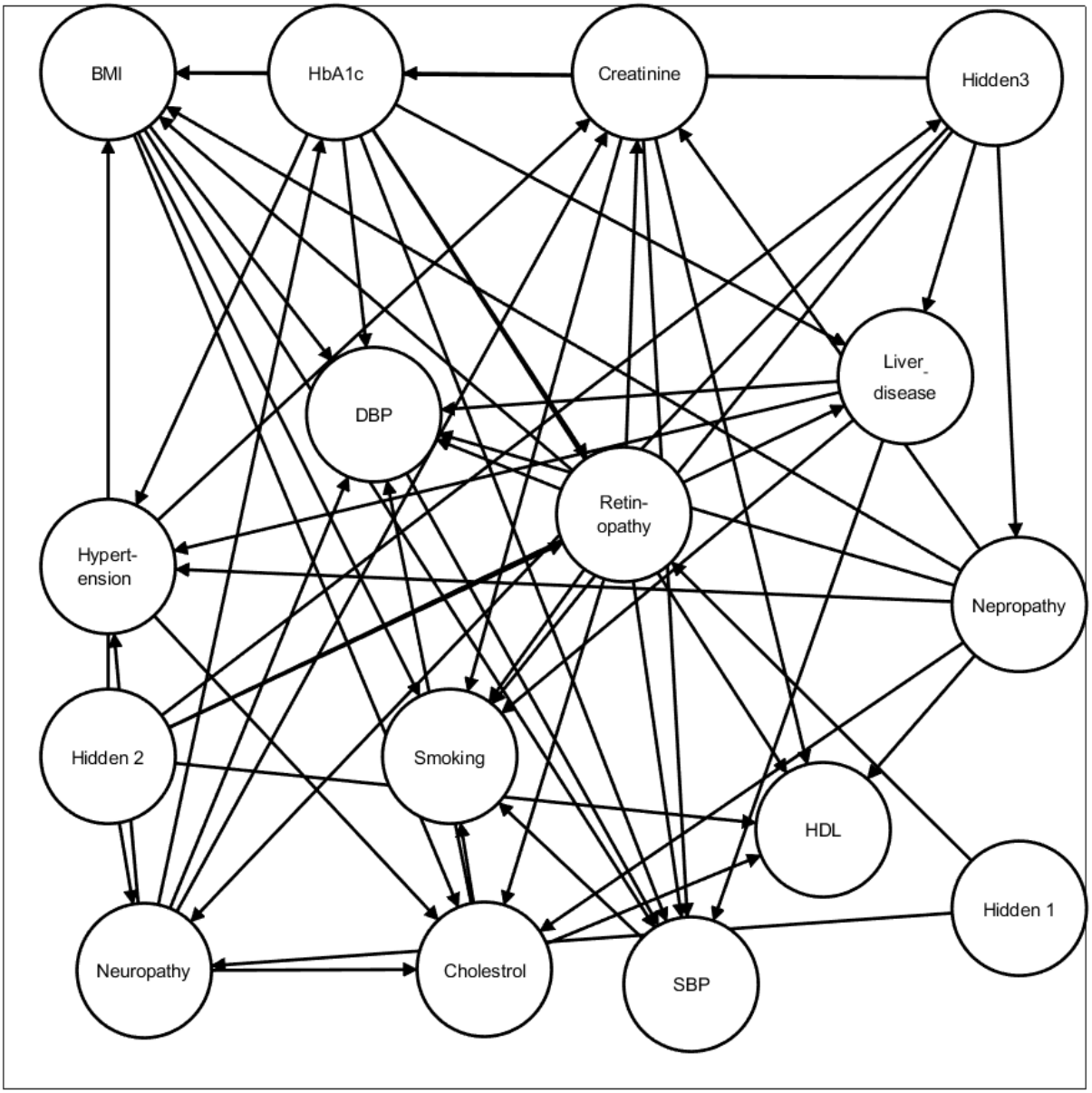

Figure 4.2: Graph of static relationships among T2DM risk factors by applying the third step of the stepwise IC* approach.

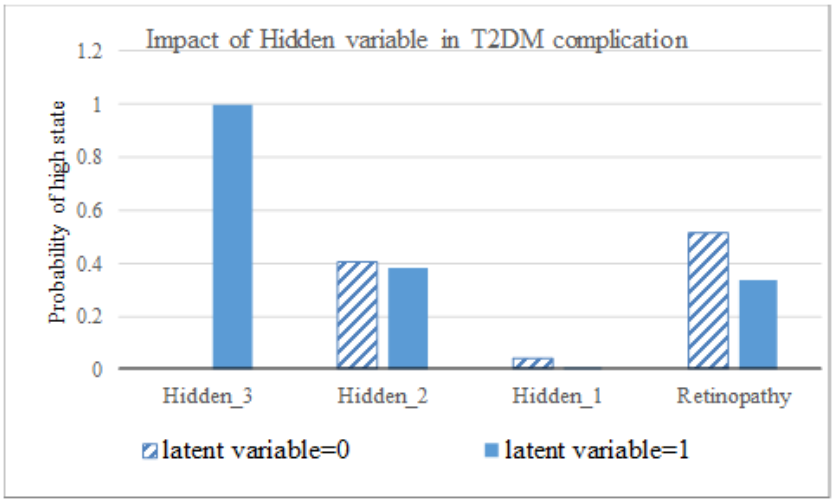

Figure 4.3: Changes in target complication (retinopathy) in response to different values of evidence (latent variable at the third step of the Stepwise approach.

4.5.2 Miss-classification Assessment

In assessing the miss-classification rate, the links between features in an adjacent matrix of IC* are described below:

- True Positive (TP), if there is a bi-directional link between two nodes.
- False Negative (FN), if the learned model lacks a bi-directional arc between two nodes.
- True Negative (TN), if the learned model has no bi-directional arcs.
- False Positive (FP), if there are one or more bi-directional links.

In order to assess the predictive model, appropriate validation analyses are conducted to predict the onset of T2DM complications (e.g., accuracy, sensitivity, specificity and precision by using 95 percent confidence interval). Thus, "patient-based validation strategy" tests the model based on the re-balanced train data on different patients data in test set, which is not consisting of the patients belonged to the training set, randomly retrieved from the original T2DM dataset (mainly in Chapter 4 and Chapter 6).

Table 4.1: Comparative performance analysis of the different steps of the Stepwise approach for three complications Comparative performance of retinopathy, liver disease and hypertension.

	Accuracy	Sensitivity	Specificity	Precision
Retinopathy/no Hidden	0.47	0.56	0.48	0.49
Retinopathy/step1	0.63	0.49	0.93	0.84
Retinopathy/step2	0.88	0.99	0.67	0.77
Retinopathy/step3	0.87	0.99	0.65	0.77
Liver disease/no Hidden	0.68	0.99	0.29	0.6
Liver disease/step1	0.83	0.99	0.79	0.83
Liver disease/step2	0.84	0.99	0.67	0.79
Liver disease/step3	0.88	0.99	0.77	0.87
Hypertension/no Hidden	0.52	0.95	0.18	0.58
Hypertension/step1	0.63	0.74	0.66	0.66
Hypertension/step2	0.58	0.75	0.39	0.54
Hypertension/step3	0.58	0.74	0.46	0.53

The FCI and PCA both rely on statistical significance tests to decide whether an arc exists between two variables and on its orientation. In addition, a default error rate ($\alpha = 0.05$) was used to find the correlation of T2DM risk factors using IC* algorithm.

As can be seen in Figure 4.2, the discovered relations among DBN nodes were shown to build the predictive retinopathy model at the three steps of the stepwise approach. As Friedman points out, a latent variable as a leaf/child or as root with only one child would be marginalised without affecting the distribution over the remaining variables. So there would be a latent variable that mediates only one parent and one child. It can be seen how the addition of the first hidden node influences the hidden node discovered at the second step (by the explaining away effect via NEU). The third hidden node was then added based upon being linked in part to hidden variable 2. Each of the components of T2DM exposure (HBA, H at the first step, and at the second step) were significantly associated with the risk of retinopathy progression. Similar results involving varying interacting hidden variables were observed when they were applied to other complications. The influence of a newly learned latent variable in each step of the enhanced stepwise algorithm was demonstrated by a bar chart in Figure 4.3.

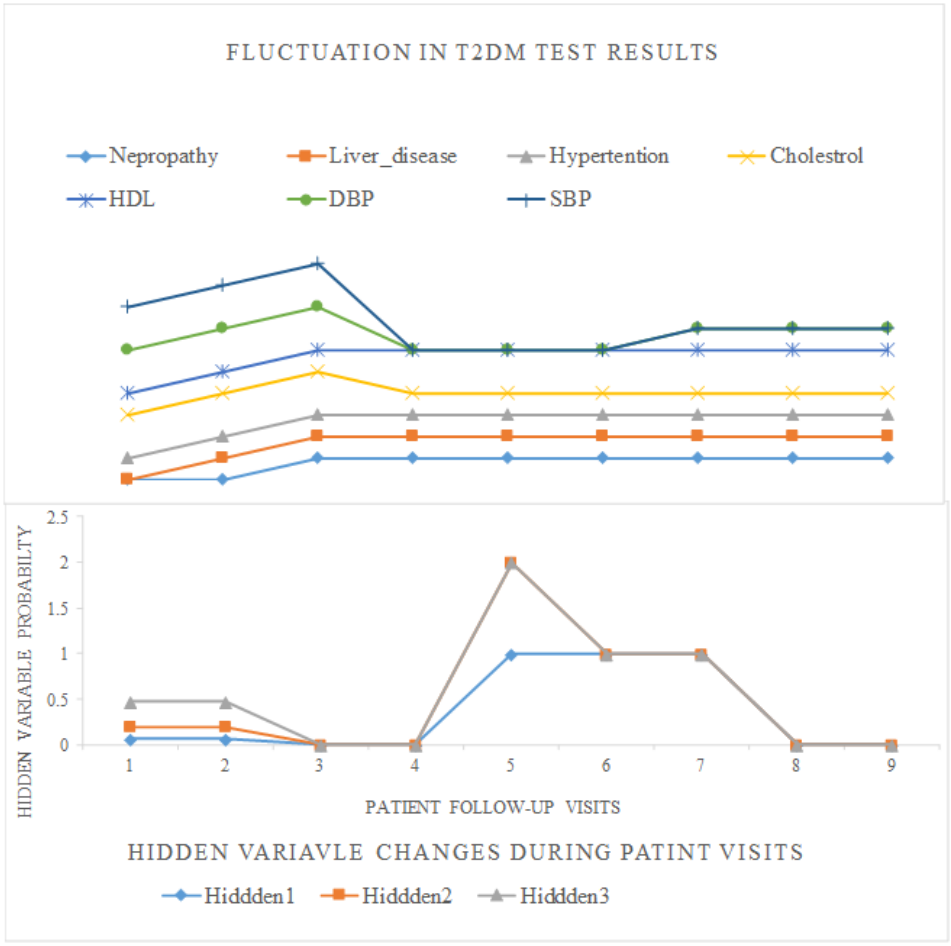

Figure 4.4: Impacts of understanding Hidden variable patterns at each step of Stepwise method on retinopathy prediction performance.

In Figure 4.3, the effects of the targeted latent variable were analysed by changing the prior as evidence of the observed latent variable in different states (0 and 1). Bayesian network inference was used to query target complication to capture the probability distribution. As can be seen in Figure 4.3, the prediction probability of retinopathy was dropped while evidence (latent variable in the first time slot) was switched to one. The hidden variable at the third step of the prediction (hidden 3) has been generally seen as a factor strongly related to retinopathy. In contrast, the probability of retinopathy being diagnosed was dropped from 0.5 to 0.3 by setting evidence from 0 to 1. Thus, it discussed whether discovering multiple hidden variables in the higher step of the enhanced stepwise approach could be a significant contributory factor to the development of retinopathy. Furthermore, it was also understood that the hidden variable at the second step of learning plays an important role in the diagnosis of retinopathy.

Figure 4.2 emphasised the power of the stepwise approach with a generally improving accuracy as several hidden variables were added. The accuracy plots indicated that the performance of the predictor for retinopathy from the no-hidden step to the first step of the approach was less significant than precision (compared to Figure 4.3-a-b). The sensitivity of retinopathy prediction was increased sharply from 0.47 to 0.99. However, sensitivity for prediction of liver disease and hypertension was remained constant after the first step of the approach. Despite the sharp rise in performance measures by adding a hidden variable at the first step of the procedure, for the rest of trend (learning more hidden variable), there was a slight increase from the first step (step1) to higher stages. Additionally, Figure 4.3-a-c illustrated that there was a sharp rise in prediction results (accuracy, precision and specificity) by exploiting a latent variable in the first step of the stepwise IC* method. Eventually, from Figure 4.3-a, it was evident that prediction accuracy has been improved from the first step (step1) to the third step (step3), especially for retinopathy and liver disease. The precision measure seemed easier to be interpreted, e.g, in Figure 4.3-b-a precision of 0.77 in the third step (step3) of learning hidden variable could immediately be understood as it was correctly diagnosed the positive occurrence of liver disease. Figure 4.4 explained how the changes in the hidden variable in three steps of the learning method reflect fluctuation in the observed variables at different points per patient visits. As can be seen in Figure 4.4, the latent variable which has been learned at the

third step of stepwise approach is on its peak and higher than the other steps and at the same time earlier than observed variables (SBP and DBP) rise points. Altogether, these outcomes emphasised the importance of hidden variable discovery in an early time disease prediction.

Having discussed the structure of the links considering the influence of the hidden variables could also be analysed by focusing on their Markov blanket. There seemed to be a strong relationship among T2DM key risk factors (BMI, liver disease, SBP, and DBP) in the market blanket, which was shown in Figure 4.2- Step4. The Links in Figure 4.2- Step4 showed that H2A1c was associated with an increased incidence of nephropathy, while H2A1c emerged as an independent risk factor for developing retinopathy, which could be validated by clinical evidence provided in [121]). Additionally, nephropathy and liver disease were independently associated with an increased incidence of hypertension in T2DM patients (clinical evidence was reported in [120]). These results suggested that the presence of macrovascular complications is positively correlated with the occurrence of microvascular complications such as neuropathy and nephropathy.

4.6 Summary

Predicting disease complications at the early stage of a longitudinal study has been known as a critical issue which has high practical benefits in clinical applications. For many clinical problems in patients, the underlying structure of risk factors (hidden factors) plays an important role in medical interventions. This chapter has made a start by developing an intuitive stepwise method to learn these latent effects based upon the IC* algorithm. More specifically, the proposed approach effectively integrated Bayesian methods with latent variables by adapting the prior probability of the event occurrence for future time points. To achieve this, in the data cleaning and pre-processing stage a new pair re-sampling strategy was employed, which helped to show how the hidden variables influence the T2DM risk factors. These results revealed that the proposed method is more accurate than using one of hidden variable step or no hidden variables at all. One limitation of the proposed approach could be the stopping rule to the stepwise approach, and in some cases, it seemed that accuracy starts to drop after the

final hidden variable is added. This might represent overfitting in the model. Classification accuracy could be monitored and used as a stopping condition (i.e., if it drops significantly). Although the IC* algorithm only learns static structure (another process is used to learn temporal links), there is potential to update the IC* algorithm to learn temporal associations. The Relationship of T2DM risk factors affects the risk of development and progression of complications in follow-up visits. A systematic understanding of how latent variables contribute to T2DM complications is still lacking. A new approach explained in the next chapter will use mutual information metrics to filter some of the hidden variable relationships where IC* results in uncertainty in choosing either a latent variable or a direct link between two nodes. In Chapter 5 the experimental findings and their significance will be tested statistically and by using the confidence interval. Eventually, it intends to explain the hidden variables which may enhance the stratification of patients and aids in understanding interactions between risk factors and unmeasured variables.

Chapter 5

Enhanced Latent Model and Patient Stratification Using Temporal Phenotype

5.1 Introduction

The primary goal of this chapter is to enhance the stepwise approach in previous chapter to fine tune the latent variables whilst determining the impact of latent variables within probabilistic networks generated from the observations. In the previous chapter, an intuitive stepwise method, based upon the IC* algorithm, was developed to learn the effects of multiple hidden variables on the prediction performance. However, the discovery of the optimum number of the hidden variables was not easy and sometime accuracy dropped as more were added due to overfitting. Thus, this chapter attempts to address these issues by proposing an enhanced variation on the stepwise IC* method (which is called IC*LS approach) for incrementally identifying hidden variables. It involves techniques for analysing the strength of relationships between clinical and hidden variables to better understand the meaning of the hidden variables within the complex disease model and explore their effect. Despite the importance of the latent variable discovery, there remains a paucity of evidence on understanding of how the discovered

latent variable contributes to explaining the complex patients model. Therefore, the main motivation behind this chapter is to stratify patient groups by means of latent variables to discover how complications in diabetes interact. This chapter explores how to cluster patients into different subgroups based on their latent variables (which is called "temporal phenotype"). Dynamic Time Warping Distance is used for time-series clustering to group patients based upon these hidden variables to uncover their effects on the complications. This distance metric is chosen because it can perform effectively to measure the patient dissimilarities based on their behaviour even with unequal lengths of time-series. This chapter is organised as follows: In Section 2, the process required to stratify patients is divided into two halves. The first half aims to increase the reliability of measures to produce the correct number of the latent variables by using the enhanced stepwise approach. It also contributes a re-balancing approach (which is called TS Bootstrapping). The second half explains how time-series clustering is utilised to group patients based on the temporal phenotype. Section 3 compares the enhanced methodology to the previous methods based on quantitative validation strategies, such as visit-based, patient-based, sensitive analysis, confidence interval. Section 4 interprets the quantitative results based on clinician point of view and medical articles before visualising the identified clusters in Section 5. Section 6 summarises and discusses how a greater focus on clustering patients in the next chapter will be maintained to produce interesting subgroups of patients that account more for understanding the latent variable.

5.2 The Enhanced Stepwise Approach

To rule out the possibility of whether a latent variable can be used to group patients, the most informative latent variables are discovered by adapting the stepwise approach. In Chapters 3-4 for learning the structure of the model, the K2 and stepwise IC* algorithms were used to create non-temporal (Intra) link, respectively. The main weakness of those algorithms was the failure to address how to learn a structure with the correct number of latent variables. Therefore, in this section, a combination of the IC* algorithm and the Link Strength methods are combined with learning the structure of DBNs. More explanation of the Link Strength measure, the

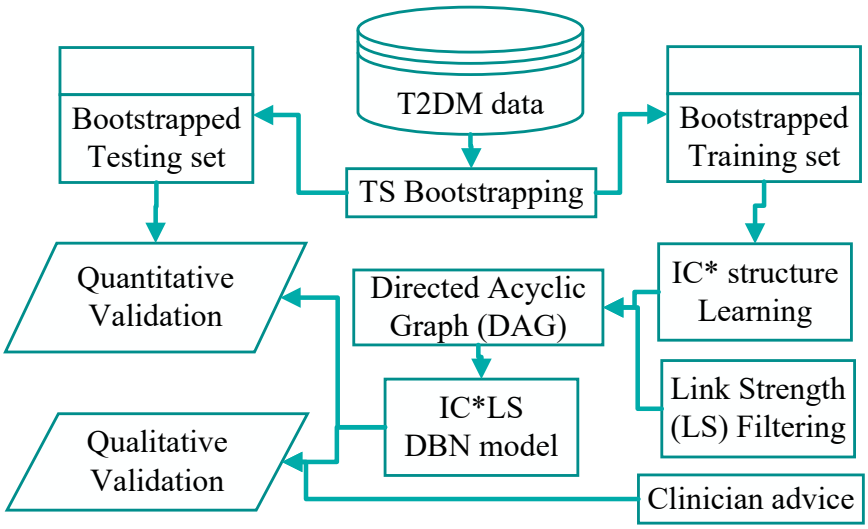

Figure 5.1: IC*LS Diagram: The overall strategy of the proposed predictive model.

Latent Structures and the key stages of implementing the enhance stepwise methodology are shown in Figure 5.1 and are explained as follows:

- **Pre-processing Stage: to Address the Imbalance Issue:**a time-series Bootstrapping approach is employed (which is introduced as TS Bootstrapping). This method is a variant on the re-sampling approach in Chapter 3 is utilised. It re-samples observed time-series visits per patient with the replacement, and the original training data is re-sampled in pairs of consecutive time points, $t-1$ and t.

- **Model Generation:** the discrete-time DBNs with two-time slots (t and t-1) are represented under the Markov properties assumption. These networks with temporal associations between the risk factors are inferred from the re-balanced T2DM historical patients. In the DBNs framework (as seen previously in Chapter 3, Figures 3.6-3.7), nodes represent variables at distinct time slots. A link represents the associations among nodes over time, so it can be used to forecast into the future. The bootstrapped data are trained and tested on their power to predict a complication at the next time point before the latent variables are explored. For instance, Figure 3.7 showed the first complication

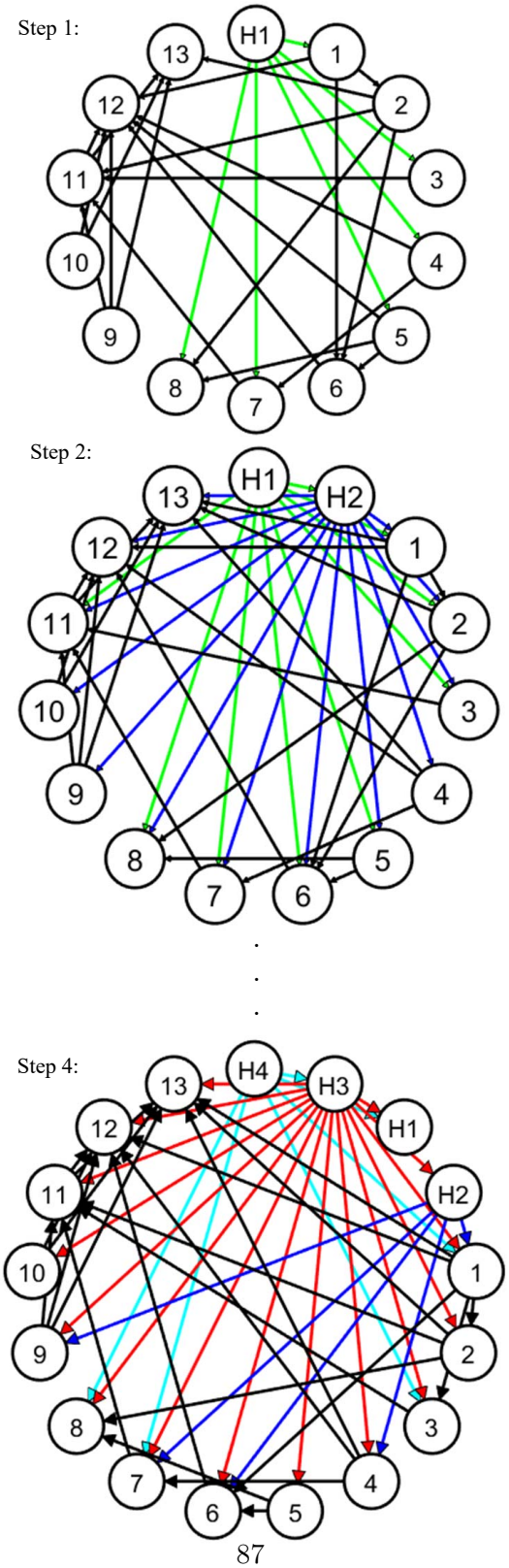

Figure 5.2: DAG of static relationships among T2DM risk factors by applying Step 1, 2 and 4 of the enhanced stepwise IC*LS approach.

at time $t-1$ affects the clinical states of other comorbidities and risk factors at t.

- **Link Strength Metric:** Link Strength (LS) [45] is a metric to calculate the overall strength of the dependent links. It focuses on the most powerful dependencies between T2DM risk factors and enables model to observe the specific impact of each discovered edge in a DBN. The percentage points of uncertainty reduction in a variable are utilised by knowing the state of another variable if the states of all other parent variables are known. True Average Link Strength (LSTA) calculates LS based on the average over the parent states using their actual joint probability. If there was a link in the IC* adjacent matrix with LSTA greater or equal to some threshold (here 20 percent), then a link in the final structure was retained; otherwise, it was deleted. This threshold was chosen to avoid providing overly connected networks and loops in the final DAG as well as to decrease the risk of edge overfitting. More explanations of LS and its measures are included in Appendix A.1.

- **Stepwise IC*LS Approach and Latent Structure:** an extended IC* stepwise approach attempts to identify the correlation among the latent variable and T2DM risk factors, which is called Induction Causation Link Strength (IC*LS) methodology (which also is introduced as "Enhanced latent model"). The probability of a high clinical level of the nodes and the learned hidden variables are then inferred using the BN inference. The CPTs and the EM algorithm are used to estimate the network parameters. The resultant CPTs indicate the probability of being in one state has given the states of all associated risk factors from the relationship graphs. In order to discover the correct number of hidden variables, extra checks are conducted on the learnt DAG on the stepwise IC* algorithm (which has been introduced in Chapter 4). As a result, the LS metric is applied to the stepwise IC* to provide a higher chance for DAG to learn optimal numbers of hidden variables; hence, a better stopping point can be obtained.

The next section aims to define similarity among patients by investigating the distance over either hidden variables in an unsupervised methodology.

5.3 Time Series Clustering

Having discovered the hidden variables and built a DBN predictive model, this section attempts to group patients to capture the status of the patient's risk factors during their time-series and investigate the relationship among them. For identifying patient groups (clusters) in the clinical time-series dataset, the latent variable probabilities for each patient are mapped to a vector of time-series. This vector should be considered as comparing pairs of patients. The concept of similarity in one cluster of patients is based on distances between two patients across their unequal follow-ups. However, calculating the right distance function to compare the pairs of patients would be a challenge. Capturing these local and dynamic correlations across a similar pattern among risk factors in the calculation of an average for each patient time-series would be another challenge. Nevertheless, the discovery of such clusters of patients is essential in revealing substantial correlations in T2DM risk factors in response to the disease over time. Thus, here an appropriate method is suggested. Dynamic Time Warping (DTW) [15] is used as a distance metric to find dissimilarities among patients. DTW distance is a suitable measure to evaluate the similarities and dissimilarities of time series concerning their shape. This metric can measure the discovered hidden variables probabilities to group patients into clusters. In this work, uni-variate DTW provides a warping function that compares a hidden variable vector of a patient time series to a hidden variable vector of another patient series, where these two vectors do not necessarily need to be equal. To achieve this, DTW keeps one patient hidden variable vector constant while stretching and shrinking the hidden variable vector to fit. This is then fed into hierarchical clustering (complete) to build sub-groups of patients based upon their hidden variables. This is also known as complete linkage cluster analysis since a cluster is formed when all the dissimilarities between pairs of patient visits in the cluster are less than a particular level. Thus, these sub-groups are distinguished by comparing the hidden variable patterns of patients. These pattern of patient behaviour can be thought of as a "temporal phenotype" for the cluster of patients. Based on the clustering model, two patients are similar if they exhibit similarity in the most common/frequent temporal phenotype. In order to characterise the profile of each discovered group, Medoid analysis [28] is applied

to the DTW distance matrices to extract a patient (pattern) with the smallest inter-patient distance among different sub-groups. This facilitates the overall risk factors behaviour across all patients in a subgroup. Mediod is calculated based on the Partitioning Around Medoids (PAM) algorithm to find a patient behaviour (as a temporal phenotype) in the centre of the cluster. The Mediod is chosen, in this study, because it can better handle noise and outliers. In addition to this, PAM estimates the most reasonable distance among items and eliminates a sum of pairwise dissimilarities among patients by using the Mediod instead of a sum of squared Euclidean distances which is used in the k-mean algorithm. This informative pattern determines a representative of a cluster, which is noted here as a "deep temporal phenotype". The next chapter, therefore, moves on to discuss the discovered clusters in more details while attempts to validate the clustering approach by comparing them to a different type of groups. Turning now to the experimental evidence on a set of models learnt from the data to evaluate the impact of adding latent variables and re-balancing the data via bootstrapping.

5.4 Experimental Results and Quantitative Validation Strategies

This section assessed the effectiveness of the bootstrap re-balancing method and the latent variable discovery approach in T2DM dataset. In Figure 5.2, a DAG for each step of the stepwise IC*LS approach is learned [1]. Conditional dependency for the hidden variable observed in the first, second and fourth steps are plotted in green, blue and cyan colours, respectively. The selected T2DM nodes (features and predictors) are labelled and ordered from 1 to 13 which are corresponded to complications and risk factors including: HBA, RET, NEU, NEP, LIV, HYP, BMI, CRT, COL, HDL, DBP, SBP and SMK, respectively (as shown previously in Tables 3.1-3.2). The initial hidden variable (H1) is closely linked to a small number of clinical factors, notably 1,3,4,5,7 and 8. However, as subsequent hidden variables are added, this structure changes. The second hidden variable (H2) is linked to more risk factors including H1 (see Figure 5.2).

[1] Another example for IC*LS DAGs is provided in Figure 5.9.

The proposed structure has been evaluated by performing the sensitivity analysis on the cohort based on two different perspectives: a "Visit-based" and a "Patient-based" validation test. The results were documented for the following comparative structures:

- "UNB-K2-REVEAL": the original data (unbalanced) was trained in the K2 algorithm for Intra links and the REVEAL algorithm for Inter links with the unbalanced data (which is not reliable due to the imbalance issue explained in Chapter 3).
- "B-K2-REVEAL": a latent variable and a fully learned structure from the K2 algorithm for Intra links and the REVEAL algorithm for Inter links with the balanced data using the TS Bootstrapping approach (shown in Figure 3.7).
- "NO-latent": the network is fully learned from the re-balanced data by using PC algorithm with no latent variable for Intra links shown in Figure 3.5. The dynamic structure for Inter links is Fully Auto-Regressive; each node is connected to the corresponding node in the next time slice shown in Figure 3.6.
- "IC*": the structure is obtained by using the IC* algorithm from the balanced data for Intra links seen in Figure 4.2 and Fully Auto-Regressive structure for the Inter links shown in Figure 3.6.
- "IC*LS": a combination of the IC* and LS filtering method is used to discover the structure for Intra links shown in Figure 5.2-Step4 and Fully Auto-Regressive shown in Figure 3.6.

In Table 5.2, the enhanced stepwise approach was compared to the previous stepwise approach in Chapter 4. It seemed evident that the enhanced stepwise method has achieved a better performance measurement in predicting retinopathy.

Figures 5.3-5.4-5.5 illustrated a case study to investigate how the latent variables have been interacted with other risk factors for predicting a complication in an individual patient. The early time prediction probabilities were represented in X-axis. In contrast, the targeted patient's visits were shown in the Y-axis. The predicted likelihood of liver disease was established in Figure 5.3-d seemed to be to very similar to its observed probability shown in

Table 5.1: Visit-based performance assessment percentages on the prediction results for three complications.

Performance Measure	UNB-K2-REVEAL	B-K2-REVEAL	NO-latent	IC*	IC*LS
AUC of Retinopathy	0.35	0.50	0.92	0.87	0.97
AUC of Liver Disease	0.38	0.51	0.68	0.90	0.97
AUC of Hypertension	0.60	0.51	0.63	0.81	0.97

Table 5.2: Comparison of enhanced stepwise IC* approach with its previous version in Chapter 4 and without latent variable in Chapter 3

Percentage:	Accuracy	Sensitivity	Specificity	Precision
No Hidden variable in Chapter 3	0.48	0.53	0.48	0.53
stepwise IC* (Step1) in Chapter 4	0.60	0.40	0.80	0.70
enhanced stepwise (Step1)	0.80	0.51	0.98	0.97
stepwise IC* (Step2) in Chapter 4	0.78	0.98	0.58	0.68
enhanced stepwise(Step2)	0.95	0.80	0.96	0.86
stepwise IC* (Step3) in Chapter 4	0.78	0.98	0.58	0.68
enhanced stepwise (Step3)	0.95	0.81	0.96	0.82
enhanced stepwise(Step4)	0.96	0.81	0.97	0.92
enhanced stepwise(Step5)	0.95	0.82	0.97	0.85

Figure 5.3-a, which indicated the complication occurrence slightly earlier than the prediction. The IC*LS latent approach, in Figure 5.3-c for liver disease, revealed a trigger around the clinician observation time, whereas the latent K2 process in Figure 5.3-b remained steady. A less significant predicted probability was also captured in Figure 5.4-d. This illustrated a fluctuation just before retinopathy has been monitored in Figure 5.4-a. Similarly, a trigger happened in two latent approaches in Figure 5.4-b-c. It revealed that the latent models had been appeared to be predicting the switches in most patient cases. However, with the small sample size, caution must be applied, as the findings might not be applicable and there have been a few cases where the model could not predict a complication earlier than the clinicians. As a result, the expected findings for predicting hypertension might differ from the conclusions presented here, as it was compared in Figures 5.5-d to Figures 5.5-a. It could be argued that the prediction results might be caused because of differences between complications. For example, hypertension has been reported as an easily detected macrovascular disease. In con-

trast, retinopathy as a chronic microvascular has been known very challenging to be caught at the earlier stage of the disease progression.

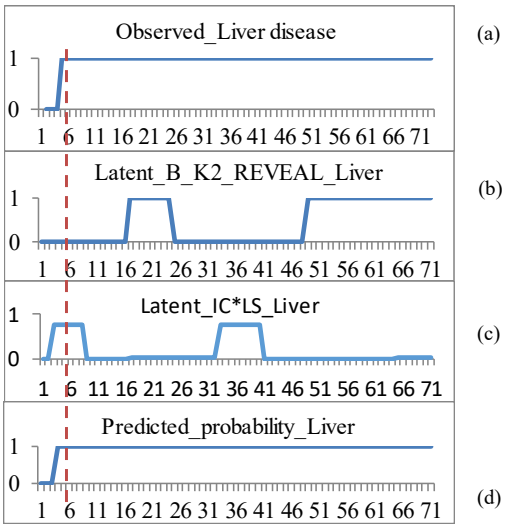

Figure 5.3: The latent Variable Behaviour for predicting the onset of Liver disease: A latent prediction pattern of liver disease over time (a patient follow-ups). The red dotted line represents marks the actual time of the disease occurrence.

5.4.1 Confidence Interval Results

To manage the uncertainty in the prediction, this study confined itself to Confidence Interval (CF) results derived from a randomly selected subset of T2DM patients. Here, the uncertainty in the structure and the predictive model was typically outlined by a confidence interval that has been declared to incorporate the true parameter value with a pre-defined likelihood. This was achieved by using the enhanced approach was compared to the previous results obtained in Chapter 3-4. In particular, T2DM patients data were randomly over-sampled for 250 times in predicting a target complication of T2DM (e.g., retinopathy). Clustered column charts in Figure 5.6 demonstrated the fluctuations of the average classification accuracy percentages of the randomly over-sampled cases, for five steps of the enhanced stepwise method. These results in Figure 5.6 revealed that the prediction accuracy of retinopathy in step one had been increased sharply by adding hidden variables at step two to four and then dropped slightly at

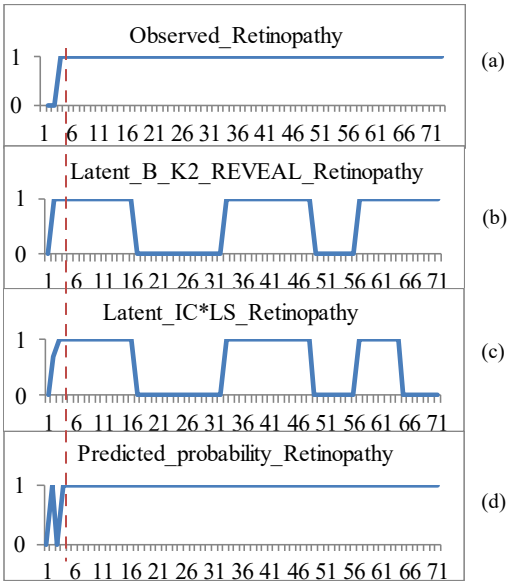

Figure 5.4: The latent Variable Behaviour for predicting the onset of retinopathy: Latent variable prediction pattern of retinopathy over time (a patient follow-ups).

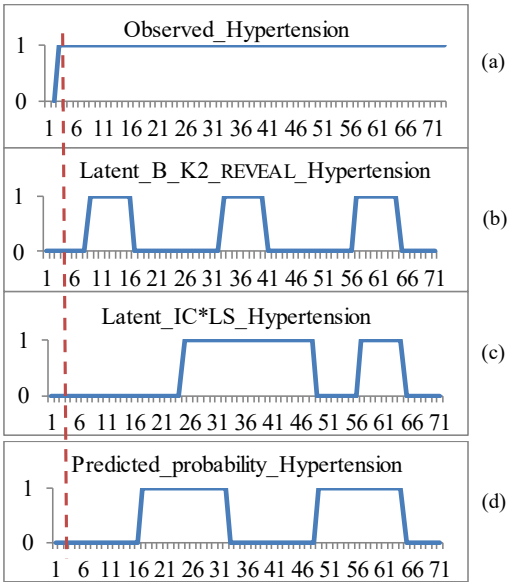

Figure 5.5: The latent Variable Behaviour for predicting the onset of hypertension: Latent variable prediction pattern of hypertension over time (a patient follow-ups).

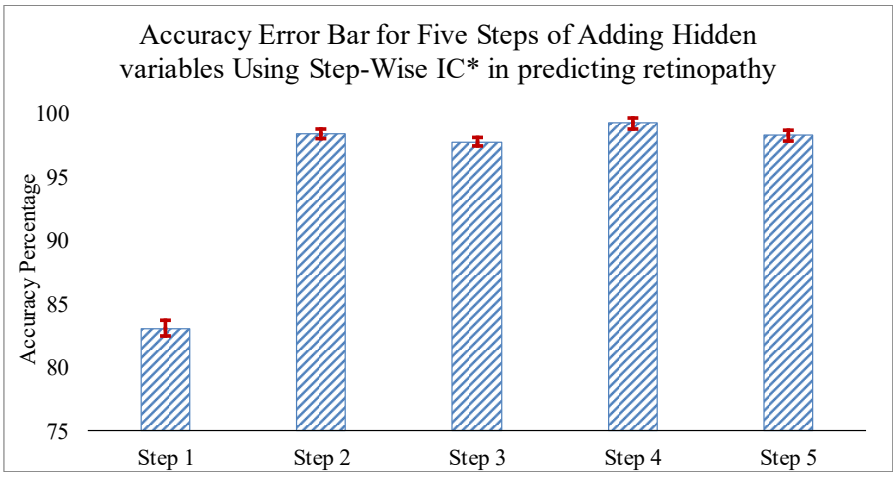

Figure 5.6: An error bar is obtained for calculating confidence interval for average classification accuracy (for 250 times) of predicting retinopathy at 5 steps of the enhanced stepwise IC* approach.

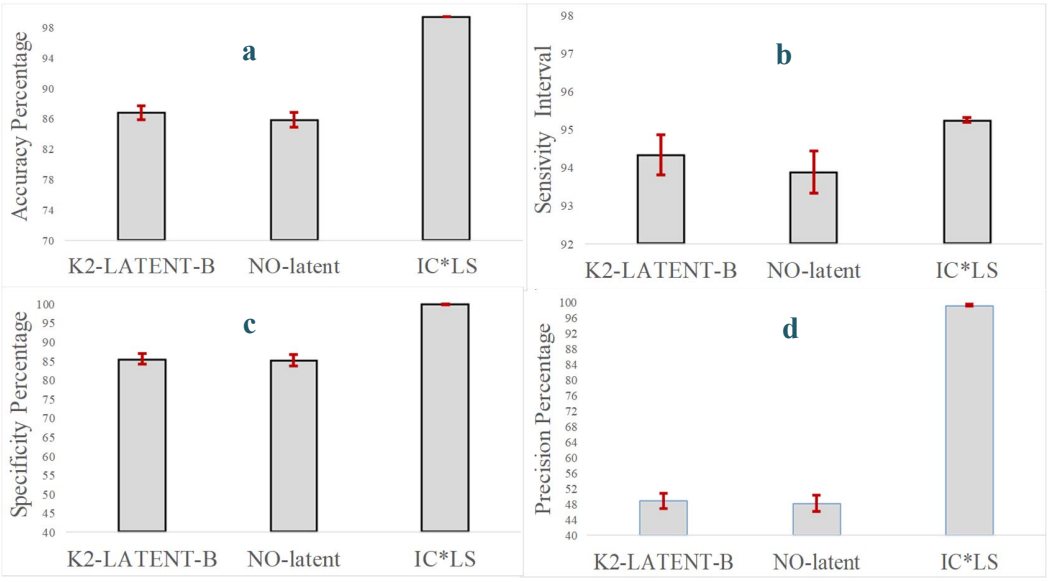

Figure 5.7: Bootstrap Confidence Interval: accuracy, sensitivity, specificity, and precision of liver disease prediction (Visit-based).

step five. Additionally, error bars on the top of the bar charts were illustrated. For example, the error bar in step one was more significant than the subsequent steps. The error bar in step two is quite large due to a more considerable confidence interval of the successive steps. Overall, it seemed that more significant than 95 percent that accuracy in retinopathy prediction using the stepwise IC*LS structure with more than one hidden variable appeared to be more accurate than a learning model with only one hidden variable. The influence of the latent variables on predicting liver disease was assessed in Figure 5.7. In this Figure, a very high percentage of 95% confidence interval was achieved by employing the IC*LS methodology, compared to the K2 and REVEAL algorithm and no latent variable approaches.

5.4.2 Qualitative Approach to Interpret the Predictive Model

The previous results showed how the targeted use of latent variables improves prediction accuracy, specificity, and sensitivity over standard approaches as well as aiding the understanding of relationships between these latent variables and disease complications/risk factors. Looking at how the different structures were performed within a DBN for predicting the appearance of complications, Figure 5.8 revealed that there could be a general trend to improvement in accuracy as more hidden variables have been added. Surprisingly, this improvement levelled out after adding the fifth hidden variable. To report more precise results, confidence intervals to manage the uncertainty in the prediction results derived from a randomly selected subset of T2DM patients. It was important to bear in mind the possible bias in the findings could not be extrapolated to all patients in the small-sized dataset. Although there was a direct link (correlation) between the latent variable and liver disease in Figure 4.2- Step4, these results should be interpreted with caution, as this did not necessarily mean that the latent variable caused liver disease. 5.8 showed how the probabilities of retinopathy, liver disease and hypertension, are influenced by the discovered latent variables. Surprisingly, in 5.8-b, a slight change was found in liver disease values, whilst the latent variable has the highest value. There was a significant negative correlation between the latent variable and hypertension, which was shown in 5.8-c. As a result of including this latent variable, there was a steep rise in the prediction accuracy of hypertension from 63% to 97% (in the IC*LS approach in

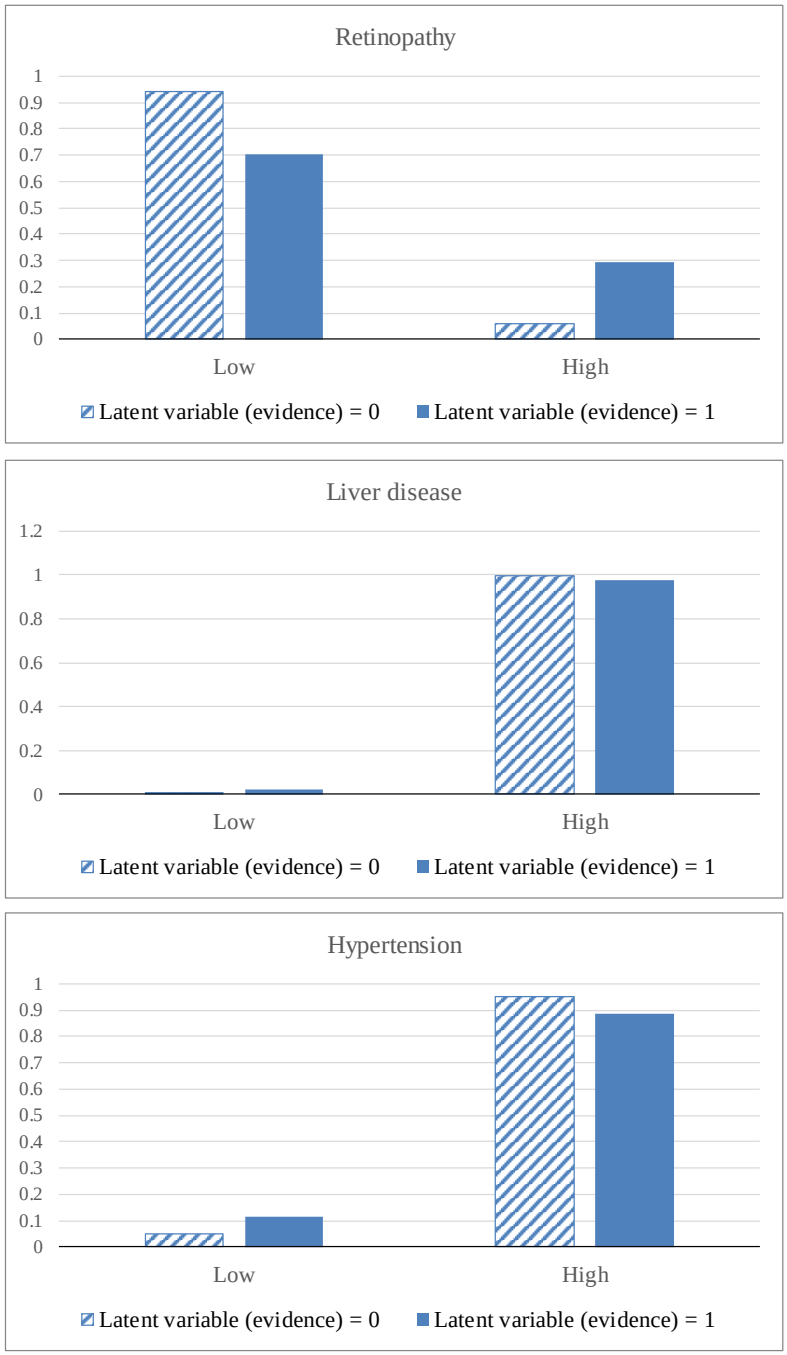

Figure 5.8: Prediction probabilities: The obtained posteriors for retinopathy, liver disease, and hypertension using a latent variable as the evidence.

Table 5.1 compared to NO-latent). Similarly, a positive correlation was found between the latent variable and retinopathy in Figure 5.8-a. It was apparent from Table 5.1 that retinopathy prediction was enhanced considerably from 92% (NO-latent) to 97% (IC*LS) by adding the latent variable. Together these findings have provided important insights into the latent variable effects, which helped to reduce the uncertainty in the prediction process by identifying the relationship between T2DM complications and risk factors. The AUC results obtained in Table 5.1-UNB-K2-REVEAL predicted hypertension accurately 60% of times comparing to 35% for retinopathy while data was imbalanced. This revealed the degree of improvement in the prediction performance from 35% to 51% for retinopathy and 38% to 51% for liver disease, whilst 60% to 51% for hypertension. The reason behind this could be argued that hypertension has been known as a macrovascular complication while retinopathy reported as a typical microvascular complication. Furthermore, hypertension appeared to be the easiest complication to be detected by clinicians due to the routine measurement of blood pressure. Alternatively, retinopathy and liver disease required either ophthalmology consultation or ultrasonography of liver.

The overall approach in this thesis is abstracted in Figure 5.10. In top of the Figure, first the patient's history (including the disease risk factors and complications) was learned and trained in a DBN model (in the middle). The obtained DAG was learned at each step of the stepwise IC*LS approach representing the links from a latent variable to other clinical risk factors. Then the inferred latent variable probabilities were employed to predict a target complication earlier than the actual occurrence time (bottom of the figure). This figure also revealed that the first latent variable (at visit $t-1$) was closely linked to a small number of clinical factors, while the second latent variable (at visit t) was connected to a larger number of risk factors.

5.4.3 Cluster Analysis

Here, the meaning of the hidden variables was explored beyond the DBNs structure within the DBNs using time series clustering and DTW distance. The discovered hidden variables were utilised to identify groups of patients based upon their temporal phenotype, which was

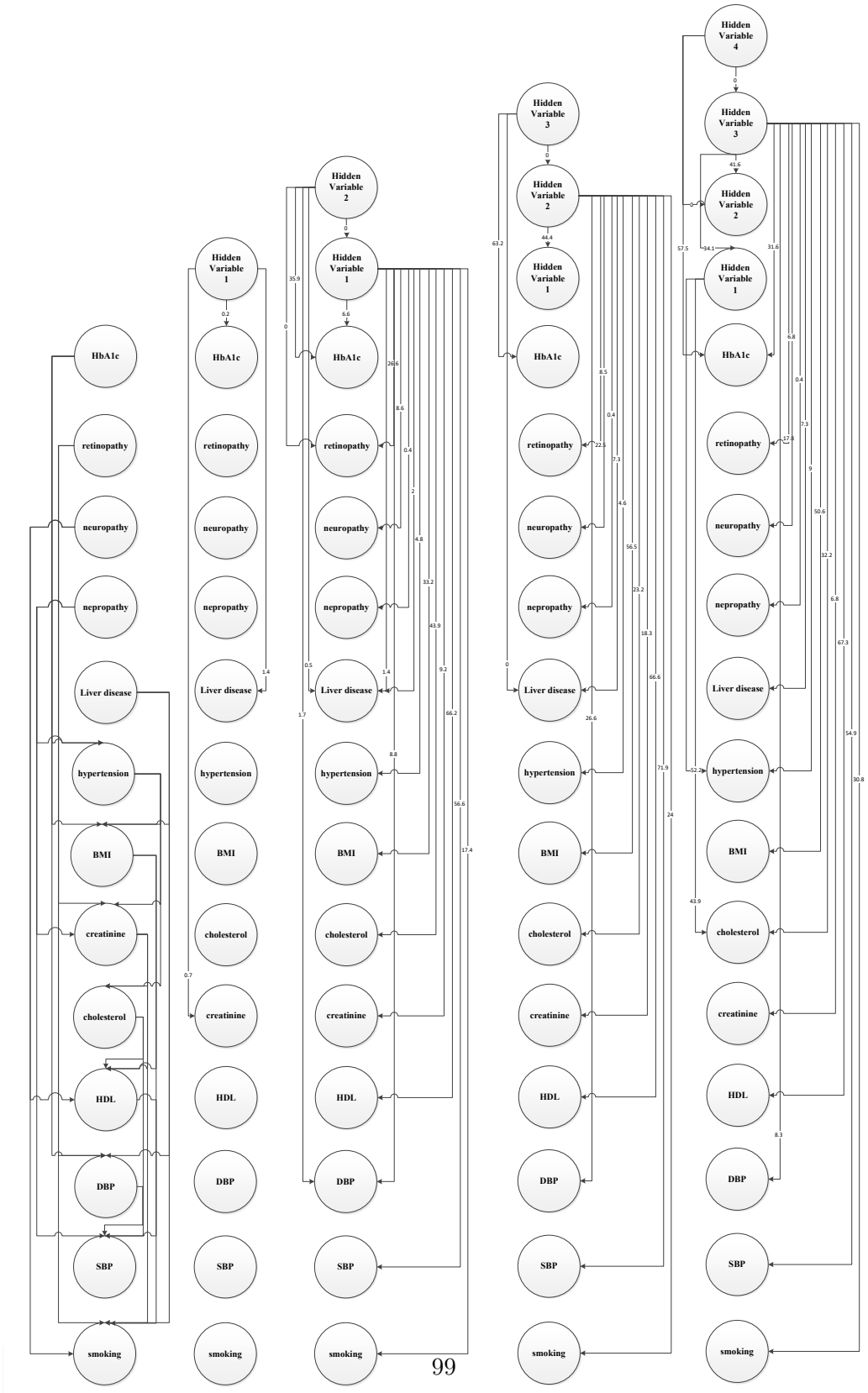

Figure 5.9: Hidden variables influence on clinical risk factors.

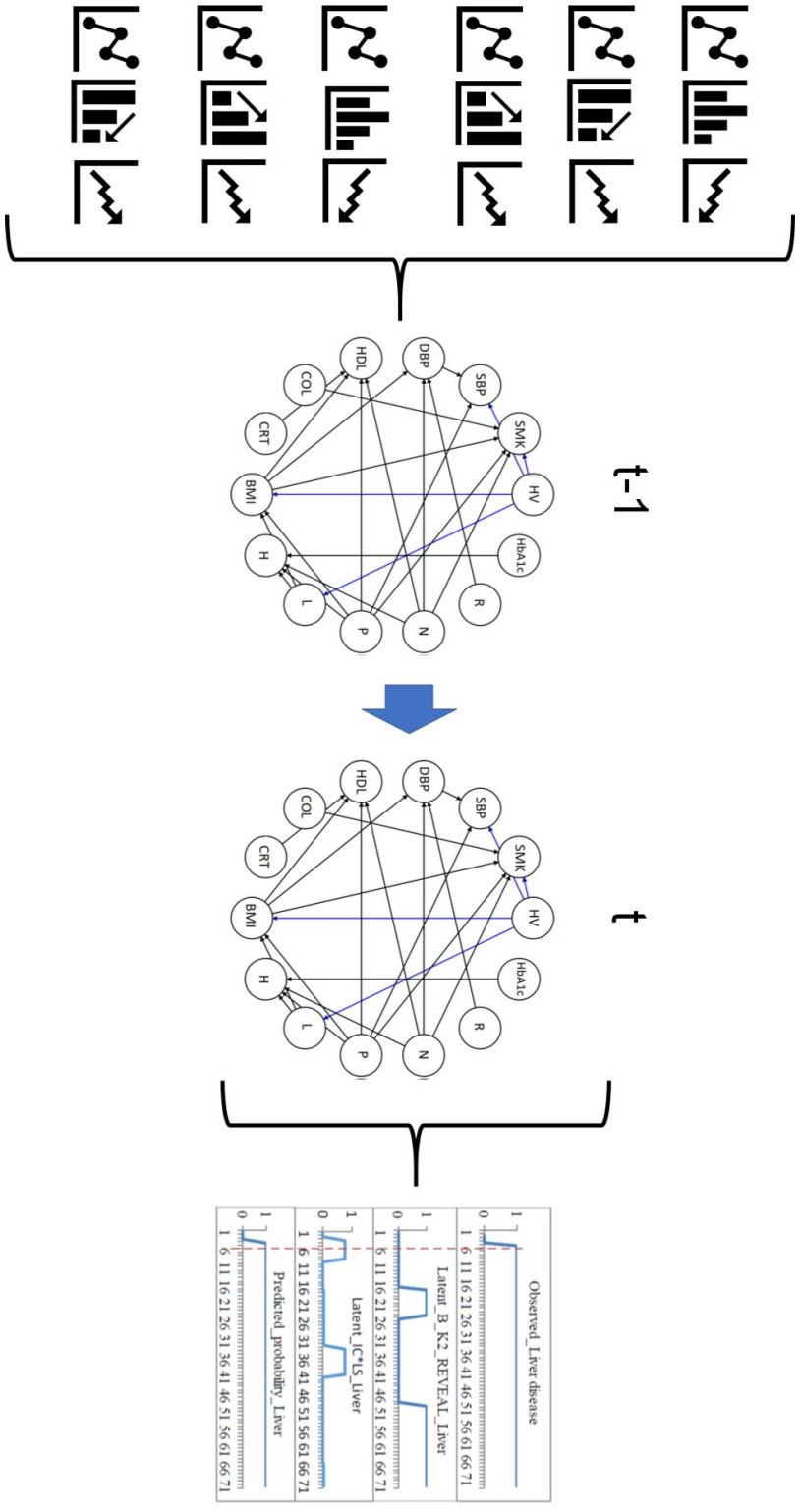

Figure 5.10: A DBN Latent Model: from the top, in the middle, and bottom demonstrate the patients history, the inferred latent variable probabilities, the prediction, respectively.

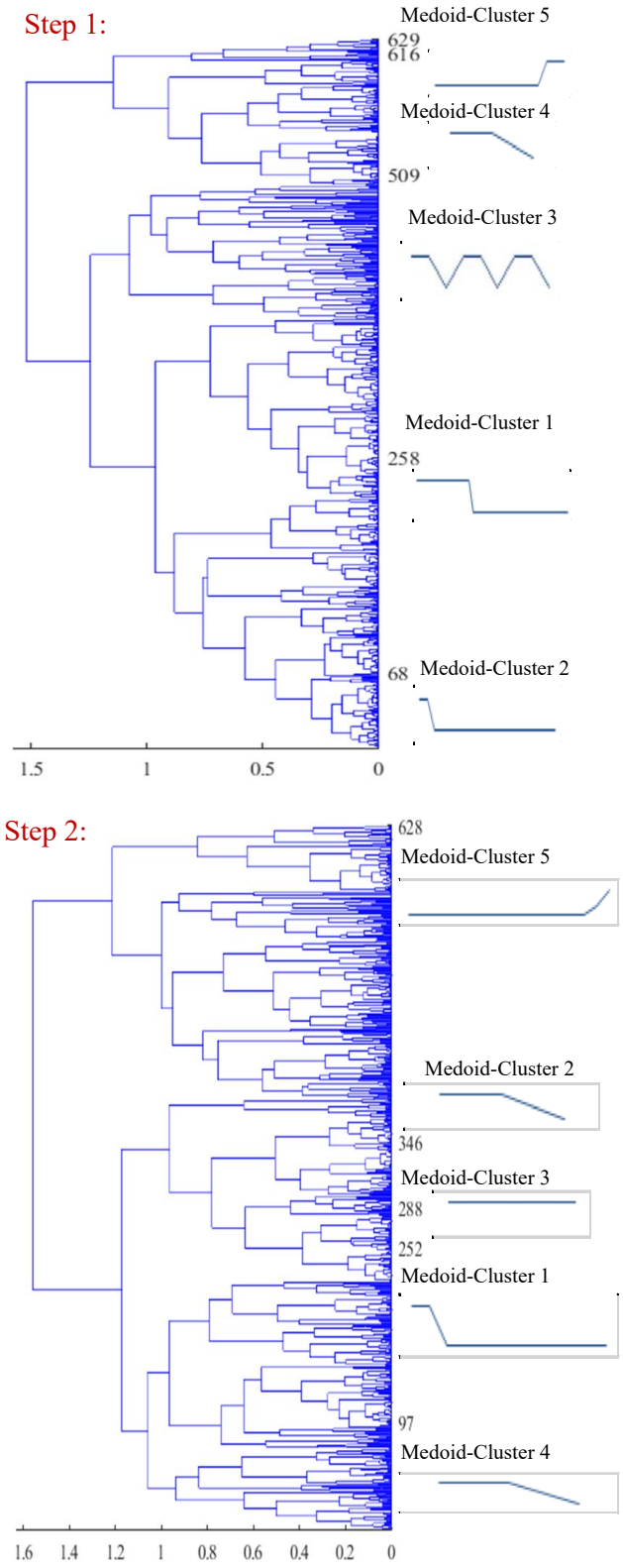

Figure 5.11: Temporal phenotypes (The First Hidden Clusters "Profiles") in hierarchical clustering. Deprograms of Hierarchical clustering (complete) for the first and second hidden variable with the DTW distance metric. The x-axis represents is a measure of closeness of either individual data points or clusters, while y-axis is representing patient IDs as data points.

noted as cluster "Profiles". Dendograms of hierarchical clustering in Figure 5.11 were shown for each sub-group of the patients. This Figure demonstrated the Medoid-clusters at the first and second learned hidden variables.

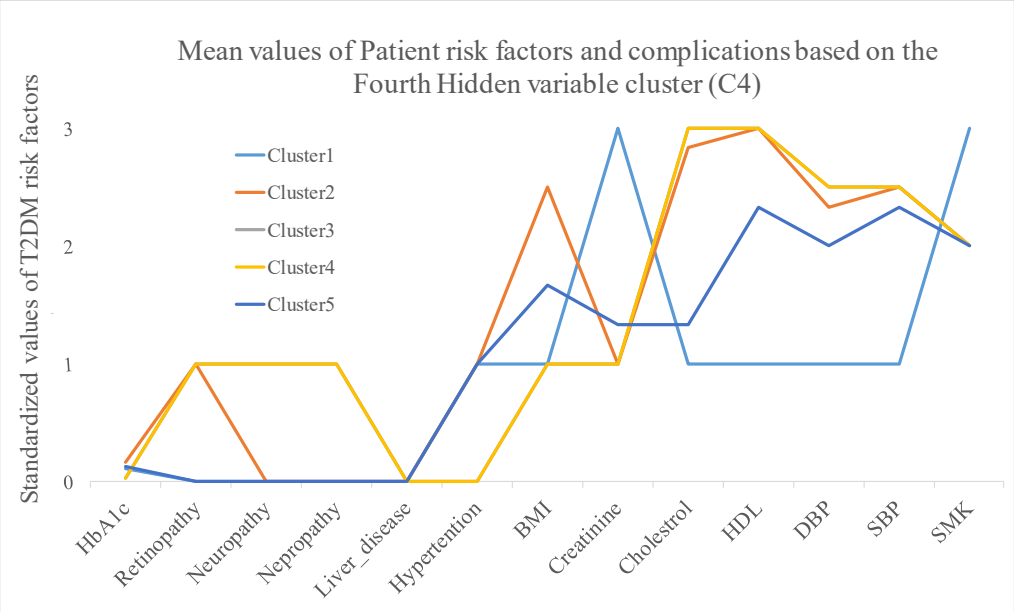

Figure 5.12: Cluster Profile on mean values of patient risk factors and complications. Patients clustered using the fourth hidden variable obtained from the fourth step of the enhanced stepwise IC*LS algorithm (C4).

In Figure 5.11, these profiles captured quite different behaviours: one was fluctuating between the higher state and lower state of the first hidden variable (Cluster 3 in step 1), involved a switch-like behaviour, one involved a general decreasing trend (cluster4 in all two steps), and another was flat-lining (Cluster 3 in step 2). Considering the associated mean values of the clinical variables for each cluster, it seemed that the data had generated clearly separated cohorts of patients. Figure 5.12-C4 revealed impressive results for each T2DM risk factor in terms of the type of patients in a cluster. For example, patients with high BMI, low HDL, and low SBP are represented in yellow (Cluster 1), whilst patients in Cluster 3 (with the flat-lining hidden cluster profile) generally had low BMI and high HDL. Cluster 4, with the decreasing hidden cluster profile, also reported much higher BMI values amongst the patients and very low HDL and DBP. As a result, for each patient in Cluster 4, the Mediod-

cluster represented an informative pattern shown in Figure 5.11, Step 1, Mediod-Cluster 4, which was identified based on the hidden variable and phenotypic discovery approach using DBNs and IC*LS algorithm in the enhanced stepwise algorithm. This also represented the overall patients' patterns of risk factors over time (for Cluster4 profile showing yellow line in Figure 5.12).

5.5 Summary

This chapter proposed the IC*LS approach as an enhanced version of the stepwise IC* approach with more robust stopping points to reduce uninformative hidden variables. It also revealed how these hidden variables could improve prediction performance with a study using confidence intervals. Furthermore, it clustered patients based upon the discovered hidden variables and used the Medoid hidden variable profile of each cluster to characterise the temporal phenotype of that set of patients. The proposed methodology in this chapter can be combined using pattern mining approaches to validate the target hidden variables and enhance the understanding of the sub-types of the disease based upon the developing disease complications. This is the subject of next chapter, which explores how the discovered latent variables interact amongst themselves and with clinical variables by using inference techniques on different complications. It also discusses how significant subgroups of patients can improve the prediction performance for a sequence of complications.

Chapter 6

Personalised Patients in Precision Medicine Using Explainable Latent Model

6.1 Introduction

The results of the descriptive experiments discussed in Chapter 5 showed the possibility of identifying different subgroups of patients based on the temporal phenotype. Nevertheless, the techniques used in these investigations were not validated for interpreting each subgroup to enhance the prediction of the associated complications. Moreover, in type 2 diabetes literature no attempt has been made to quantify the association between complications in the prediction performance, while they can be numerous and interact in complex non-linear ways throughout the disease process. This chapter proposes a hybrid approach that includes Temporal Association Rules to identify frequent co-occurrences of complications over time, and Temporal Pattern Clustering to build meaningful subgroups. These methods can also be combined for a better understanding of an informative temporal phenotype (which will be referred to as the "temporal phenotype" for the remained of this chapter) as well as underlying patterns of complications associated with the patients. The obtained clusters of the rules are compared to

groups of the latent phenotypes extracted as reported in Chapter 5 (using the DTW distance). Finally, several validation strategies involving the Jaccard Index and Bayesian analysis are employed where inference is more transparent.

Overall, this chapter provides two types of strategies in data mining as follows: Section 2 discusses descriptive strategies where Temporal Association Rules, time-series clustering and Association Rule Mining were combined to build the hybrid approach. Section 3 documents the result obtained from Section 4 to evaluate the subgroups by using the clustering comparison methodology approach. Section 5 turns to predictive strategies where the prediction accuracy of the associated complications was tested on the DBNs framework before concluding in Section 6.

6.2 Data Mining Techniques: Personalising Patients in Precision Medicine

So far, the previous chapters have either focused on predictive strategies in order to improve accuracy or descriptive analysis to ease explainability in understanding the underlying models (and latent variables). Now, to achieve both goals, this section suggests various data mining techniques based upon an integration of descriptive and predictive analysis. In Figure 6.1, a multiple-stage process has been taken to find explainable subgroups of the patients and to interpret the latent variable using the proposed methodology personalising diabetic patients in precision medicine. The overall methodology is labelled in Figure 6.1 and explained as follows:

1. Data pre-processing and data discretisation approaches are employed to generate the original T2DM dataset (DS).

2. For each patient an informative pattern (as temporal phenotype) is identified based on the latent variable discovery approach using DBNs and IC*LS algorithm in the enhanced stepwise algorithm (which was explained in chapter 5).

3. The DTW method is used to calculate dissimilarities between the discovered temporal phenotypes. It captures the complexities/homogeneity of the risk factors of the disease

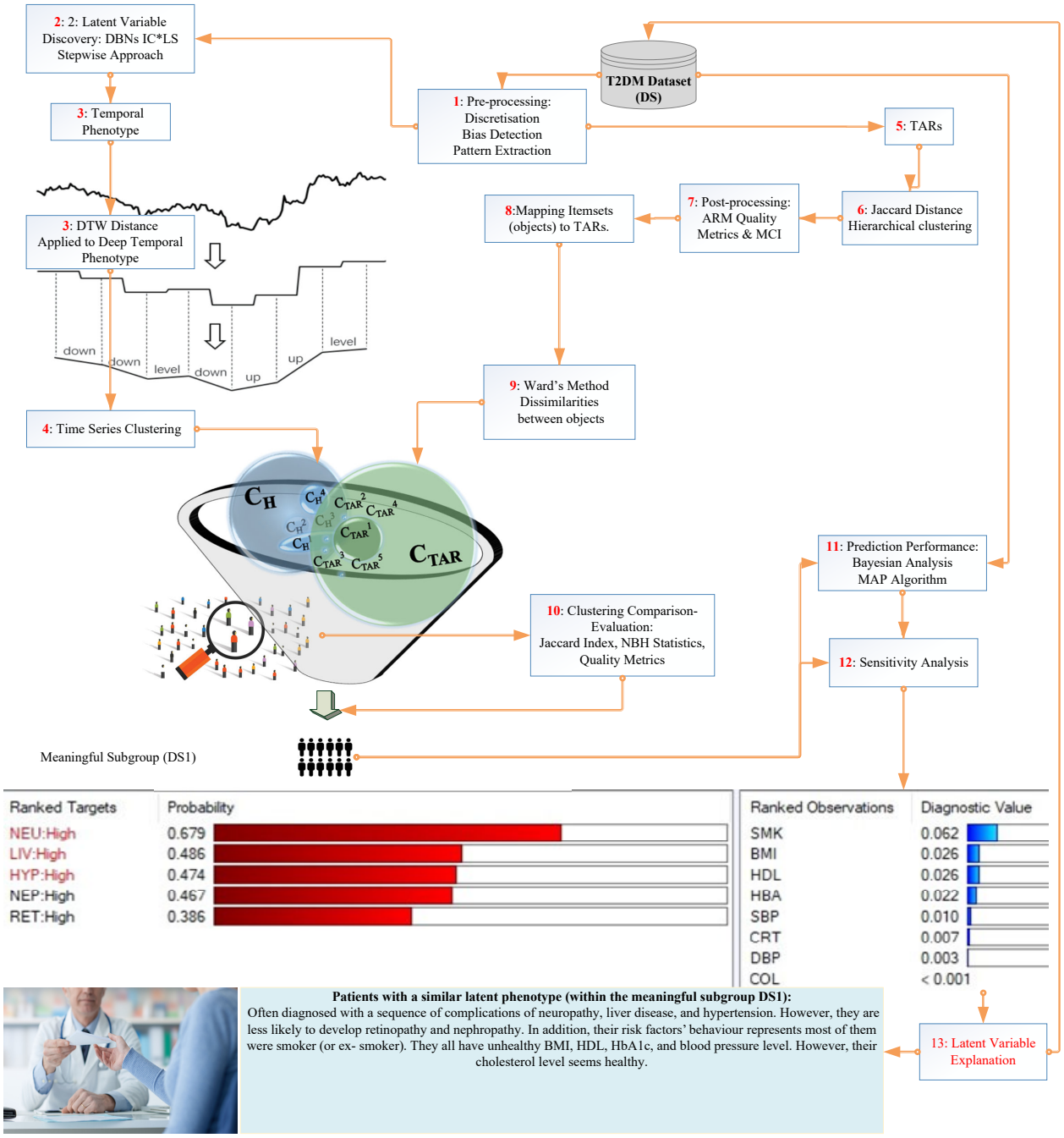

Figure 6.1: The proposed hybrid methodology to find explainable subgroups of patients by personalising diabetic patients in precision medicine.

as well as the associated complications over time.

4. Time-series clustering based on DTW distance is applied on the data to stratify patients into four clusters considering their temporal phenotypes (which is known as "H cluster" demonstrated in Figure 6.4).

5. The multiple binary complications, as items from the pre-processed dataset DS, are extracted and mined to retrieve the temporal patterns of items. TARs are applied on the obtained patterns from DS and generate database R in Table 6.1.

6. The Jaccard index is applied to the rules in R to measure the distance between the itemsets. Hierarchical Agglomerative Clustering groups sub-rules and generates rules in R.

7. A post-processing ARM approach, which is called Minimum Coverage Itemsets (MCI), is utilised for pruning the rules and investigating the most reasonable distances to obtain meaningful clustering outcomes. Thus, the proposed Algorithm 1 MCI generates least itemsets (which is known as "objects") in dataset D covering the most important/interesting/significant rules from R, in Table D, unique objects identify each of resulted itemsets of the applied MCI.

8. All rules in R are mapped to the relevant objects/itemsets in D based on the implications of the antecedents and consequents. The pattern mining and sequence discovery are performed to explain and highlight the potential usefulness of identifying patterns of T2DM complications which are called "complications-rules". These patterns of complications-rules are considered as itemsets (a basket of complications) in TARs.

9. By using Hierarchical clustering and Ward's Method, objects are grouped into five groups. Subgroups of patients with a common pattern of the complications-rules are identified (which is called "TAR clusters").

10. Comparison and validation strategies such as Jaccard Index and Normal Approximation for the Binomial Approximation of the Hypergeometric distribution (NBH) are employed

to explain the most significant subgroup through the integration of TARs with time-series clustering.

11. The most meaningful subgroup of patients is found from the intersection of H and TAR clusters. Prediction performance of the discovered meaningful subgroup (DS1) as a subset is compared to DS.

12. Sensitivity analysis is applied to DS1 and assessed its prediction performance comparing to DS. Bayesian statistics is used to test the explainability of the meaningful subgroup.

13. The outcome of the latest model for predicting and stratifying T2DM patients, with a focus on DS1, is explained with the corresponding pattern of complications-rules and temporal phenotypes. The associated complications-rules are mined to assess the occurrence likelihood of binary complications concerning the rest of the complications associated with the patients. For example, to find out whether the increasing prevalence of HYP has been accompanied by an increase in NEU or NEP by LIV. To understand how the temporal phenotypes help to group patients, a combination of the TARs mining and time series clustering is performed in the next section.

To understand how the latent phenotype helps to group patients, a combination of the TARs mining and time series clustering is performed in the next section. An outline of the MCI algorithm is provided in in Algorithm 1 MCI.

6.3 Descriptive Strategies: Personalising Patients using a Hybrid Type Methodology

To understand how the latent phenotype helps to group patients, the next stage has further extended the patient models discussed in the previous chapter by using hybrid methods consisting of Temporal Association Rules (TARs), Association Rule Mining (ARM) and Pattern Clustering. The first stage describes TARs to extract the underlying relationships among the complications, which have been utilised according to the needs of patient personalisation.

6.3.1 Temporal Associations Rules and Sequence Discovery of Complication Patterns

Temporal Association Rules are employed to explain and highlight the potential usefulness of identifying patterns of T2DM complications in identifying patients groups. For retrieving the conditional rules (conditional statement representing an ordering pattern) among the complications using TARs, which can be thought of as a basket in the shopping problem introduced in Chapter 3. This stage extends these previous concepts in more detail as it will be used in Algorithm 1 MCI. Based on the nature of the clinical records in the T2DM case study, criteria for selecting the items were as follows: there is some assumption described below. Itemsets of $\Im$ is a transaction that represents a pattern of all associated complications over a patient time series (from the first recorded visit to the last visit). In particular, $\Im$ in TARs is shown by $\{antecedent \implies consequent\}$ referring to a sequence of complications co-occurrence or visits ($\{antecedent\ ,\ consequent\}$) which is a representation of $\{consequent\}$ occurrence captured in a patient visits/time series followed by the corresponding $\{antecedent\}$. The consequent itemsets may consist of more than one item for each rule. In the process of pruning the rules to pick the most interesting one, the main priority in a predictive model for the decision making is based on consequents.

In terms of explaining temporal notation, every two itemsets with a similar ordering pattern of the complications (co-occurrences) are treated equivalent, and any redundant complication in their intersection is being ignored. An empty antecedent ($\{\}$) and two empty antecedents ($\{\}\{\}$) are equivalent. These notations represent a patient/transaction with no complication during the first two visits. Thus, time gaps among the two complications are being ignored. A symbol of "," represents a logical "AND" between two itemsets of $\Im_i$ of $\{HYP, RET\}$ and $\Im_j$ of $\{HYP, RET\}$ indicate a complications-rules for those patients who have developed HYP before RET during their visits $\Im_i \equiv \Im_j$. However, according to temporal abstraction rules, the ordering pattern/ sequence of the complications co-occurrence are important where $\{HYP, RET\}$ and $\{RET, HYP\}$ did not resemble each other. A symbol of "|" (representing a logical "OR" gate) between to items in particular itemsets indicates any of two items (RET and NEU) or their combination or none of them ($\{\}$) could occur. In addition, item-

sets of $\Im$ by having a "|" among their items can be a subset of another itemsets by having a ",", e.g., $(\{NEU|RET\} \subseteq \{NEU, RET\}$. For example, two itemsets of $\{NEU, RET, RET\}$ and $\{NEU, RET\}$ are assumed to be equivalent, whereas $\{NEU|RET\}$ can be different from $\{NEU, RET\}$ (indicating NEU must be developed before RET). Thus, similar to the first assumption, the repetition of the complications is ignored, but without consideration of ordering rules. Based on these assumptions and in order to find the most frequent complications-rules in DS, in the next section, a mixed methodology based on TARs and ARM is utilised to enhance the methodological approach taken in Chapter 5.

Table 6.1: Database R of the associated rules with the complications generated using TARs.

Rule	Antecedent Consequent	Object ID ($\Im$)	Support	Confidence	Lift
1	{ } $\Longrightarrow$ {HYP,RET}	3,14,23,27,28,33,38,41	≥ 0.001	≥ 0.001	1.00
2	{ } $\Longrightarrow$ {RET,HYP}	3,14,23,27,28,33,38,41	0.01	0.01	1.00
3	{ } $\Longrightarrow$ {NEU,HYP}	5,13,21,26,28,31,38,41	0.01	0.01	1.00
4	{ } $\Longrightarrow$ {LIV,HYP}	6,24,30,31,33,35,36,37,39,40	0.02	0.02	1.00
5	{ } $\Longrightarrow$ {NEP,HYP}	2,13,14,20,26,27,30,38,41	0.03	0.03	1.00
6	{ } $\Longrightarrow$ { }{ }	9	0.02	0.02	1.00
7	{ } $\Longrightarrow$ {NEP}	2,11,35-38,41	0.11	0.11	1.00
8	{ } $\Longrightarrow$ {NEU}	5,7,37-41	0.16	0.16	1.00
9	{ } $\Longrightarrow$ {RET}	3,4,32-34,38,41	0.15	0.15	1.00
10	{ } $\Longrightarrow$ {LIV}	6,12,18,19,22,24,29,30-37,39,40	0.15	0.15	1.00
11	{ } $\Longrightarrow$ {HYP}	2-6,10,30-33,38,41	0.86	0.86	1.00
12	{NEU,HYP} $\Longrightarrow$ {NEU}	13,26,38,41	0.01	0.27	1.71
13	{NEU} $\Longrightarrow$ {NEP,HYP}	13,26,38,41	0.01	0.05	1.71
14	{NEP,HYP} $\Longrightarrow$ {RET}	14,27,38,41	0.01	0.27	1.79
15	{RET} $\Longrightarrow$ {NEP,HYP}	14,27,38,41	0.01	0.05	1.79
16	{ }{ } $\Longrightarrow$ {RET}	3,4	0.01	0.22	1.46
17	{RET} $\Longrightarrow$ { }{ }	3,4	0.01	0.03	1.46

Table 6.1: Database R of the associated rules with the complications generated using TARs.

Rule	Antecedent	Consequent	Object ID (ℑ)	Support	Confidence	Lift
18	{ }{ }	{HYP}	2-6,33,38,41	0.02	0.78	0.90
19	{HYP}	{ }{ }	2-6	0.02	0.02	0.90
20	{NEP}	{NEU}	13,16,25,26,29,38,41	0.02	0.19	1.17
21	{NEU}	{NEU}	13,16,25,26,29,38,41	0.02	0.13	1.17
22	{NEP}	{RET}	14,17,25,27,32,38,41	0.02	0.14	0.92
23	{RET}	{NEP}	14,17,25,27,32,38,41	0.02	0.10	0.92
24	{NEP}	{LIV}	18,29,30,32,36,37	0.04	0.37	2.49
25	{LIV}	{NEP}	18,29,30,32,36,37	0.04	0.27	2.49
26	{NEP}	{HYP}	2,13,14,20,26,27,30,38,41	0.10	0.93	1.08
27	{HYP}	{NEP}	2,13,14,20,26,27,30,38,41	0.10	0.12	1.08
28	{NEU}	{RET}	15,25,28,34,38,39,40,41	0.04	0.25	1.64
29	{RET}	{NEU}	15,25,28,34,38,39,40,41	0.04	0.26	1.64
30	{NEU}	{LIV}	19,29,31,34,35,37,39,40	0.02	0.11	0.73
31	{LIV}	{NEU}	19,29,31,34,35,37,39,40	0.02	0.12	0.73
32	{NEU}	{HYP}	5,13,21,26,28,31,35,37-41	0.12	0.78	0.91
33	{HYP}	{NEU}	5,13,21,26,28,31,35,37-41	0.12	0.14	0.91
34	{RET}	{LIV}	22,32,34,36,39,40	0.03	0.20	1.31
35	{LIV}	{RET}	22,32,34,36,39,40	0.03	0.20	1.31
36	{RET}	{HYP}	3,14,23,27,28,33,36,38,41	0.12	0.79	0.91
37	{HYP}	{RET}	3,14,23,27,28,33,36,38,41	0.12	0.14	0.91
38	{LIV}	{HYP}	6,24,30,31,33,35,36,37,39,40	0.14	0.92	1.06
39	{HYP}	{LIV}	6,24,30,31,33,35,36,37,39,40	0.14	0.16	1.06
40	{{NEP,HYP},NEU}	{RET}	28,38,41	0.01	1.00	6.57

Table 6.1: Database R of the associated rules with the complications generated using TARs.

Rule	Antecedent Consequent	Object ID (ℑ)	Support	Confidence	Lift
41	{{NEP,HYP},RET} $\Longrightarrow$ {NEU}	28,38,41	0.01	1.00	6.27
42	{NEU,RET} $\Longrightarrow$ {NEP,HYP}	28,38,41	0.01	0.19	6.84
43	{NEP,NEU} $\Longrightarrow$ {RET}	25,38,41	0.01	0.25	1.64
44	{NEP,RET} $\Longrightarrow$ {NEU}	25,38,41	0.01	0.33	2.09
45	{NEU,RET} $\Longrightarrow$ {NEP}	25,38,41	0.01	0.13	1.17
46	{NEP,NEU} $\Longrightarrow$ {LIV}	29,35,37	0.01	0.25	1.67
47	{LIV,NEP} $\Longrightarrow$ {NEU}	29,35,37	0.01	0.13	0.78
48	{LIV,NEU} $\Longrightarrow$ {NEP}	29,35,37	0.01	0.29	2.66
49	{NEP,NEU} $\Longrightarrow$ {HYP}	26,35,37,38,41	0.02	0.88	1.01
50	{HYP,NEP} $\Longrightarrow$ {NEU}	26,35,37,38,41	0.02	0.18	1.10
51	{HYP,NEU} $\Longrightarrow$ {NEP}	26,35,37,38,41	0.02	0.14	1.31
52	{NEP,RET} $\Longrightarrow$ {LIV}	32,36	≥ 0.001	0.17	1.11
53	{LIV,NEP} $\Longrightarrow$ {RET}	32,36	≥ 0.001	0.06	0.41
54	{LIV,RET} $\Longrightarrow$ {NEP}	32,36	≥ 0.001	0.08	0.78
55	{NEP,RET} $\Longrightarrow$ {HYP}	27,36,38,41	0.01	0.67	0.77
56	{HYP,NEP} $\Longrightarrow$ {RET}	27,36,38,41	0.01	0.10	0.66
57	{HYP,RET} $\Longrightarrow$ {NEP}	27,36,38,41	0.01	0.08	0.78
58	{LIV,NEP} $\Longrightarrow$ {HYP}	30,35,36,37	0.04	0.94	1.09
59	{HYP,NEP} $\Longrightarrow$ {LIV}	30,35,36,37	0.04	0.38	2.51
60	{HYP,LIV} $\Longrightarrow$ {NEP}	30,35,36,37	0.04	0.27	2.54
61	{NEU,RET} $\Longrightarrow$ {LIV}	34,39,40	0.01	0.13	0.84
62	{LIV,NEU} $\Longrightarrow$ {RET}	34,39,40	0.01	0.29	1.88
63	{LIV,RET} $\Longrightarrow$ {NEU}	34,39,40	0.01	0.17	1.04
64	{NEU,RET} $\Longrightarrow$ {HYP}	28,38,39,40,41	0.03	0.75	0.87
65	{HYP,NEU} $\Longrightarrow$ {RET}	28,38,39,40,41	0.03	0.24	1.58

Table 6.1: Database R of the associated rules with the complications generated using TARs.

Rule	Antecedent Consequent	Object ID (ℑ)	Support	Confidence	Lift
66	{HYP,RET} $\Longrightarrow$ {NEU}	28,38,39,40,41	0.03	0.25	1.57
67	{LIV,NEU} $\Longrightarrow$ {HYP}	31,35,37,39,40	0.02	0.86	0.99
68	{HYP,NEU} $\Longrightarrow$ {LIV}	31,35,37,39,40	0.02	0.12	0.80
69	{HYP,LIV} $\Longrightarrow$ {NEU}	31,35,37,39,40	0.02	0.11	0.68
70	{LIV,RET} $\Longrightarrow$ {HYP}	33,36,39,40	0.03	1.00	1.16
71	{HYP,RET} $\Longrightarrow$ {LIV}	33,36,39,40	0.03	0.25	1.67
72	{HYP,LIV} $\Longrightarrow$ {RET}	33,36,39,40	0.03	0.22	1.43
73	{NEP,NEU,RET} $\Longrightarrow$ {HYP}	38,41	≥ 0.001	0.50	0.58
74	{HYP,NEP,NEU} $\Longrightarrow$ {RET}	38,41	≥ 0.001	0.14	0.94
75	{HYP,NEP,RET} $\Longrightarrow$ {NEU}	38,41	≥ 0.001	0.25	1.57
76	{HYP,NEU,RET} $\Longrightarrow$ {NEP}	38,41	≥ 0.001	0.08	0.78
77	{LIV,NEP,NEU} $\Longrightarrow$ {HYP}	37	0.01	1.00	1.16
78	{HYP,NEP,NEU} $\Longrightarrow$ {LIV}	37	0.01	0.29	1.91
79	{HYP,LIV,NEP} $\Longrightarrow$ {NEU}	37	0.01	0.13	0.84
80	{HYP,LIV,NEU} $\Longrightarrow$ {NEP}	37	0.01	0.33	3.11
81	{LIV,NEP,RET} $\Longrightarrow$ {HYP}	36	≥ 0.001	1.00	1.16
82	{HYP,NEP,RET} $\Longrightarrow$ {LIV}	36	≥ 0.001	0.25	1.67
83	{HYP,LIV,NEP} $\Longrightarrow$ {RET}	36	≥ 0.001	0.07	0.44
84	{HYP,LIV,RET} $\Longrightarrow$ {NEP}	36	≥ 0.001	0.08	0.78
85	{LIV,NEU,RET} $\Longrightarrow$ {HYP}	39	0.01	1.00	1.16
86	{HYP,NEU,RET} $\Longrightarrow$ {LIV}	39	0.01	0.17	1.11
87	{HYP,LIV,NEU} $\Longrightarrow$ {RET}	39	0.01	0.33	2.19

6.3.2 Association Rule Mining and Quality Metrics

Association Rule Mining (ARM), TARs and pattern clustering approaches are combined in the hybrid methodology during the process of pruning/analysing the rules to pick the most interesting complications-rules. ARM involves the generation of itemsets in TARs applied to the sets of T2DM complications to discover all combinations/sequences/sets of items (which are known as itemsets). As a result, in Table 6.1, the frequent complications-rules included in at least a significant number of patients are known as the frequent/interesting itemsets. ARM and the quality metrics are applied to the transactions of sub-rules to cluster the interesting itemsets.

In the T2DM dataset, support as a metric is regarded as an explicit constraint to identify the outliers. It is assumed to be a set of patients (representing transactions or baskets of items) containing the itemsets. These constraints can be based on the frequency of itemsets (which is referred to support) and whether their appearance is more significant than a pre-defined minimum threshold (which for the sake of simplicity is assumed to be 0.001). The minimum constraints must be assigned at a low level as the complications-rules with pre-defined constraints vary from patient to patient.

In addition, confidence calculates the probability of occurrence of $\{consequent\}$ given $\{antecedent\}$ is present. Based on the nature of clinical data, a confidence constraint of 25% is chosen to generate interesting rules. This is because, in the small-sized dataset with the appearance of bias, it is essential to ascertain that the frequent items have not been affecting the associations of other items rather than HYP. Therefore, only the effect of four out of five complications are being considered, and at least one-fourth of the complications should be measured for confidence. For example, the confidence of a rule shown in Equation 6.1 is identified by the proportion of transactions with the most interesting/important relationships.

$$support = (\mathbb{C}(\pi)_i \cup \mathbb{C}(\pi)_j) > \sigma, confidence = (\mathbb{C}(\pi)_i \Rightarrow \mathbb{C}(\pi)_j) > \delta \tag{6.1}$$

In the above equation, parameters such as σ and δ are the minimum support ($\sigma <= 0.001$) and confidence ($\delta <= 25\%$), respectively. The support metric for itemsets of $\mathbb{C}(\pi)_i *$

($support(\mathbb{C}(\pi)_i)$ is defined as the proportion of transactions in the dataset containing $RHS(\mathbb{C}(\pi)_i)$. In particular, an association of $\partial(\mathbb{C}(\pi)_i) \Rightarrow \partial(\mathbb{C}(\pi)_j)$ has a support of $P(\mathbb{C}(\pi)_i \mathbb{C}(\pi)_j)$. Confidence measures the strength of the association rules in which patients that had a complication in $\mathbb{C}(\pi)_i$ also might develop another complications in $\mathbb{C}(\pi)_j$.

For example, in T2DM case study if confidence is given for HYP, LIV implying RET, it can represent the likelihood of developing HYP, LIV and also RET over the likelihood of developing only HYP and LIV. A rule of $\{RET, HYP, NEU, RET\}$ implying LIV, which is calculated based on confidence, reveals how likely a patient develops both itemsets of $\{RET, HYP\}, NEU, RET$ and LIV.

In order to find the most interesting itemsets, support ensures that all sub-rules of the frequent itemsets are also frequent; hence no superset of infrequent itemsets can be considered as frequent. Confidence is very sensitive to the frequency of the consequent. It has been reported that consequents with higher support produced higher confidence even though there was no association among the antecedent and consequent. Nevertheless, the already mentioned metrics are not able to filter complications-rules based on the different dependencies among the rules, while another metric like lift can measure the independence between $\mathbb{C}(\pi)_i$ and $\mathbb{C}(\pi)_j$, as is shown in Equation 8. Lift is the ratio of confidence to a baseline probability of $\{consequent\}$ occurring. It assesses the probability of developing both HYP and LIV that is associated with the likelihood of developing RET. Lift is the deviation of the whole rule's support from the expected support under independence given both sides of the rule's support. Higher lift values indicate strong associations. For example, lift of 1 represents $\mathbb{C}(\pi)_i$ and $\mathbb{C}(\pi)_j$ are independent, as shown in Equation 6.3.

$$lift = \frac{P(\mathbb{C}(\pi)_i \ \cap \ \mathbb{C}(\pi)_j)}{P(\mathbb{C}(\pi)_i) \times P(\mathbb{C}(\pi)_j)} \tag{6.2}$$

$$lift(\mathbb{C}(\pi)_i \implies \mathbb{C}(\pi)_j) = support(\mathbb{C}(\pi)_i \cup \mathbb{C}(\pi)_j) = support(\mathbb{C}(\pi)_i) \times support(\mathbb{C}(\pi)_j) \tag{6.3}$$

As a result, two rules of $\{\{NEP, HYP\}, NEU\} \implies \{RET\}$ **and** $\{\{NEP, HYP\}, RET\} \implies$

$\{NEU\}$ **are considered as the most interesting rules based on the evidence showing that they had the highest confidence and lift among all itemsets.**

Having considered support, confidence and lift, it might be not be ideal and useful to effectively filter out unimportant complications-rules as they are performing with the existence of bias in DS with a small number of patients. Although the minimal constraints have been applied to these quality metrics, the outcome of TARs contained too many redundant sub-rules, in both of antecedents and consequents. This is because the total number of associated extracted sub-rules from DS is 174, without taking into account an empty set for consequent and antecedent. Moreover, the combination of these sub-rules increased the database size exponentially based on the number of items.

Therefore, a clustering approach based on the Jaccard distance has been discussed in the next stage to filter out the infrequent itemsets and measure distances between itemsets. Accordingly, Agglomerative hierarchical clustering is employed to cluster antecedents and then map them to appropriate antecedents.

6.3.3 Agglomerative hierarchical clustering and Jaccard distance

Previously, TARs have been used on the temporal co-occurrence pattern of complications to obtain their associated patterns and relatively the sub-rules/itemsets. Here, the itemsets are grouped by using the Agglomerative hierarchical clustering approach. Thus, Jaccard distance was utilised, where $d_{i,j}$ shows the difference between two itemsets $\Im_i$ and $\Im_j$ calculated the number of similar sub-rules between them over all their unique sub-rules. For comparing two different sequences of the complications (i and j) in the hierarchical clustering of the itemsets of $\Im_i$ and $\Im_j$, Jaccard distance ($d_{i,j}$) is calculated based on Jaccard Index ($Jaccard(\Im_i, \Im_j)$), as seen in Equation 6.4.

$$d_{i,j} = 1 - Jaccard(\Im_i, \Im_j), \text{where } Jaccard(\Im_i, \Im_j) = \frac{|\Im_i \cap \Im_j|}{|\Im_i \cup \Im_j|} \tag{6.4}$$

Thus far, metrics such as support, confidence and lift were used to identify the most interesting rules. Altogether, the discovered rules were generated to database R (which were

represented in Table 6.1) consisting of 174 sub-rules (87 antecedents and 87 consequents). However, as shown in the previous phase, there may still be many uninteresting/uninformative rules remaining, which can be challenging to be interpreted due to the complex nature of the associated complications. In the next stage, a minimum number of aggregated sub-rules in R are retrieved to produce the most interesting rules. Then, the identified sequence of complications is mined to extract the useful rules and detect the most common ordering pattern of the complications by using MCI. The outcome rules and the overall process in the hybrid type methods are represented in Figure 6.2.

6.3.4 Interesting Itemsets in Complications-Rules Using Minimal Coverage Itemsets Algorithm

In order to eliminate a number of infrequent rules, the Minimum Coverage of Itemsets (MCI) algorithm Algorithm 1 MCI was proposed to discover the minimum coverage of rules, which was a variation of the methodology conducted by Liu and co-authors to enhance k-means clustering in [76].

The proposed MCI procedure to discover the most interesting itemsets (which were called objects/clustering data points) was illustrated below:

MCI procedure starts from step 1, by initialising the variables, parameters (from step 2 to 10) to obtain the most interesting itemsets (step 31) temporal pattern of complications. Then, in step 11 till step 22 it identifies the meaningful sub-rules by filtering out the minimal quality metrics and using TARs mining. The outcome of this generated database R. In steps 22 to 25, Hierarchical Clustering is outlined as follows:

- The distance between these points should be measured. Save the findings in a matrix of distance.
- Check via the distance matrix to locate the two clusters/objects that are the most close.

Algorithm 1 Algorithm 6.1 MCI

procedure $MCI(R)$

▷ Initialisation:

$\sigma \leftarrow 0.001$

$\delta \leftarrow 0.25$

$DS \leftarrow$ original T2DM dataset

$m \leftarrow$ number of sub-rules $\leftarrow 87$ (antecedent and consequent)

$\chi \leftarrow$ {RET,NEU,NEP,LIV,HYP}$=\sum_{i=1}^{5} \chi_i$

Require: $\chi_i \subseteq$ PowerSet χ ▷ Set of all combinations of binary complications

$MCI(\chi_i) \leftarrow \emptyset$

$R \leftarrow$ Apriori(DS, σ, δ)

OverlapRate$\leftarrow \emptyset$

$\Im \leftarrow \emptyset$ ▷ End of Initialisation ▷ Generating Database R:

for $i \leftarrow 1, m$ **do** ▷ antecedent of R

for $j \leftarrow 1, m$ **do** ▷ consequent of R

$R_{i,j} \leftarrow \{\chi_i \implies \chi_j\}$ ▷ R of the form $(\{\chi_i\}, \{\chi_j\})$.

$R \leftarrow R_{i,j} \cup R$ ▷ $R \leftarrow \sum_{i=1,j=1}^{m=87} R_{i,j}$

end for

end for ▷ Pruning R using Quality Metrics in TARs:

$k \leftarrow \sum_{i=1,j=1}^{174}$ antecedent and consequents of $R_{i,j}$

for $k \leftarrow 1, 174$ **do**

if $support(R_k) \geq \sigma$ AND $confidence(R_k) \geq \delta$ **then**

$\Im(R_k) \leftarrow \{LHS(R_k)\} \cap \{RHS(R_k)\}$

$D \leftarrow D \cup \Im(R_k)$

Require: $lift(R_k) \geq$ MAX$(lift(R))$

end if

▷ Clustering antecedents and consequents in R using Jaccard dissimilarities

Require: $Jaccard(\Im_i, \Im_j) \leftarrow \frac{|\Im_i \cap \Im_j|}{|\Im_i \cup \Im_j|}$

$d_{i,j} \leftarrow 1 - Jaccard(\Im_i, \Im_j)$

return Filtered sub-rules

▷ Creating Dataset D of interesting objects

$l \leftarrow \sum_{i=1,j=1}^{174} R_{i,j}$ ▷ antecedent and consequents of the filtered sub-rules

OverlapRate$\leftarrow$Count $\{LHS(R_l)\} \cap \{RHS(R_l)\}$

if OverlapRate $\leq$MIN$(\{LHS(R_l) \implies RHS(R_l)\} \cap R)$ **then**

$MCI(R) \leftarrow MCI(R) \cup \{LHS(R_l) \implies RHS(R_l)\}$

end if

end for

return $MCI(R)$

Require: $C_{TARs} \leftarrow MCI(R)$

end procedure

Table 6.2: The frequent itemsets are generated in dataset D based on the rules in generated using TARs.

Objects ID	Objects(interesting itemsets)
1	{ }
2	{NEP,HYP}
3	{HYP,RET}
4	{RET,HYP}
5	{NEU,HYP}
6	{LIV,HYP}
7	{NEU}
8	{RET}
9	{ }{ }
10	{HYP}
11	{NEP}
12	{LIV}
13	{{NEP,HYP}|NEU}
14	{{NEP,HYP}|RET}
15	{NEU|RET}
16	{NEU}
17	{NEP|RET}
18	{LIV|NEP}
19	{LIV|NEU}
20	{HYP|NEP}
21	{HYP|NEU}
22	{RET}
23	{HYP|RET}
24	{HYP|LIV}
25	{NEP|NEU|RET}
26	{HYP|NEP|NEU}
27	{HYP|NEP|RET}
28	{HYP|NEU|RET}
29	{LIV|NEP|NEU}
30	{HYP|LIV|NEP}
31	{HYP|LIV|NEU}
32	{LIV|NEP|RET}
33	{HYP|LIV|RET}
34	{LIV|NEU|RET}
35	{LIV|HYP|NEU}
36	{HYP|LIV|RET|NEP}
37	{HYP|LIV|NEU|NEP}
38	{NEU|RET|NEP|HYP}
39	{NEU|HYP|RET|LIV}
40	{LIV|HYP|NEU|RET}
41	{NEP,HYP}RET,NEU}

- Join the two groups to create a cluster of at least 2 items.
- The matrix can be modified by measuring the distances between such a group and other such groups.
- Repeat phase 2 until any case is in a group.

Furthermore, from step 25 to 31 it provides the actual MCI steps needed to take to remove uninteresting rule. Therefore it returns the most interesting itemsets as objects of Wards' methods to be grouped into meaningful subgroup of patients.

As can be seen in Figure 6.2 in the left-hand side, temporal patterns of the complications co-occurrences were retrieved from DS. The database was mined to include the temporal relationships in the complications and their associated sub-rules by using TARs and ARM. The clustering method allocated an itemsets/sub-rules to a cluster in such a way that the itemsets in the same subgroup coincided with each of the other subgroups, based upon the Jaccard Distance. Then, MCI mapped these clusters of the sub-rules in database R to find the related objects of the relevant associated rules. Once all of the objects are identified, they were matched to the rules in Table 6.1. The minimal coverage itemsets generated the related objects in D of the relevant associated rules in R based on their uniqueness and lower overlap rate among one another while covering the most frequent/interesting rules. By choosing these specific objects instead of the 87 rules, a minimum overlap among the data points was produced, which could not be achieved using only lift. Thus, the distances among the objects represented higher quality data points in the clustering with less repetition of unimportant rules.

More explanations and examples have been provided in Appendix B A.1. In the next stage, the patients are grouped based on the dissimilarities among objects.

6.3.5 Combined Methodology of TARs, ARM and Pattern Clustering

This stage introduces the unsupervised techniques used to find clusters of associated complications-rules in the forms of objects by employing a hybrid type methodology. This methodology, which consisted of TARs, MCI, ARM and pattern clustering methods, is further utilised to validate the H clusters. Therefore, patients that have been diagnosed with a similar occurring pattern

of complications over time (corresponding frequent itemsets) are gathered in one cluster. In the next stage, the distance among the objects is calculated where the pattern clustering approach discussed, and the patients grouped based on the dissimilarities among objects.

6.3.6 Pattern Clustering to Obtain an Optimum Number of TAR Clusters

In this stage of the combined methodology, Ward's Minimum Variance Clustering Method [117] is performed to obtain fewer clusters (which was called "TAR cluster"). Although MCI helped to achieve the most frequent itemsets as objects, using these objects in cluster analysis has been subjected to considerable criticism. T2DM data due to the issue of the variety of ordering patterns of complications-rules in the sparse dataset, the scarce number of patients and the temporal complications-rules could have a different degree of relevance at each cluster of patients. Among a set of m itemsets/objects, there might be overall of $\frac{m(m-1)}{2}$ distances that could be used to cluster the objects. It seemed to be still some clustering errors with having too many objects and relative distances to cluster the small-sized dataset with only 368 patients whilst needed to be grouped by 41 objects and relatively up to 820 distances.

Agglomerative clustering initialised the sum of squares by a zero, it then merged two similar clusters [18] and repeats this process until only one cluster remained and the sum squares have been increased. Despite this in this stage, Ward's method is used in order to minimise the clustering error before stratifying patients to the corresponding cluster's object. The dissimilarity metric in Ward's method $(1 - |correlation|)$ is calculated based on unrelated patients based on their objects in each cluster. It also helps to generate every possible combination of clusters at each step of clustering by minimising the sum of squares based on the total inter-cluster (within-cluster) variance. As a result, a pair of clusters has led to less growth in total inter-cluster variance after merging. In Ward's hierarchical cluster analysis, 41 data point (clusters of objects) are combined to produce a new cluster containing all objects based on the assumption that the sum of squares for the objects should be as small as possible. The patients are allocated to the relevant TAR clusters (C_{TAR}), where each cluster shared a similar complications sequence (co-occurrence patterns of complications). In Table 6.4, the elements of TAR

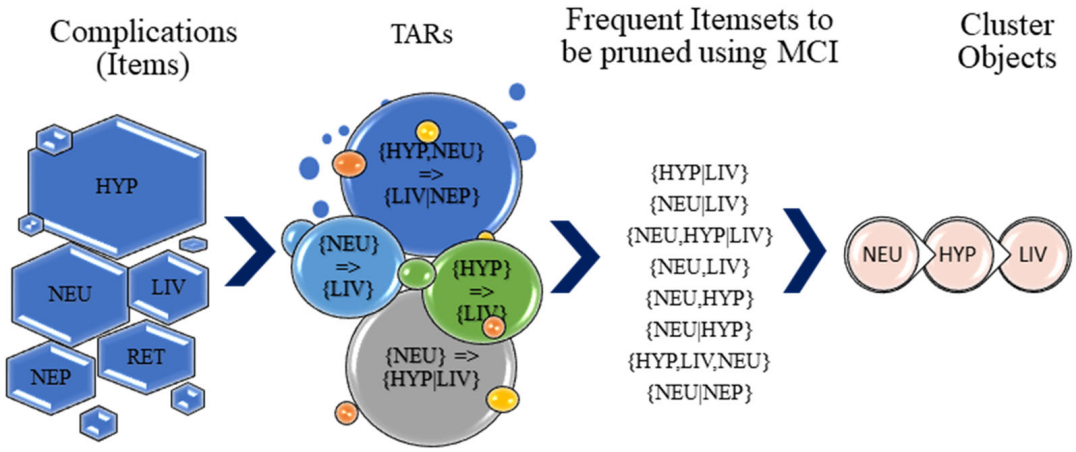

Figure 6.2: The proposed complication pattern mining methodology by using ARM and MCI to obtain the interesting itemsets as clustering objects.

Table 6.3: Clusters of the frequent itemsets identified by groups of Objects in the associated interesting itemsets from Table 6.1-6.2.

TAR clusters	Elements of cluster (Interesting itemsets/Objects)
C^1_{TAR}	10,13,2,20,21,24,26,30,31,5,6,38
C^2_{TAR}	11,12,16,18,19,29,7,9
C^3_{TAR}	14,23,27,28,3,33,38,4,41
C^4_{TAR}	15,17,22,25,32,34,8
C^5_{TAR}	35,36,37,39,40

clusters are represented as the most interesting/common itemsets amongst the corresponded patients, which in the next section are used to give meaning to the H clusters.

6.4 Clustering Comparison and Validation Strategies

In this section, an attempt has been made to ascertain the usefulness/trustworthiness of the TAR cluster in understanding the underlying disease as well as being a reliable source to validate the temporal phenotype subgroups. In order to achieve these goals, two validation strategies were employed and defined as follows:

Internal Validation Strategy In internal validation, the groups of objects and their distances were assessed, then their uninformative and rare objects in D mined to be compared to four H clusters effectively. This was achieved by pruning objects in which both high lift and confidence score were selected, and the least frequent itemsets were also ignored. The internal validation strategy tested the validity of the TAR clusters through the use of the knowledge contained within the given database of complications-rules.

External Validation Strategy In the external validation technique, H clusters were assessed based upon another data source (TAR clusters). Jaccard Index was also applied to calculate the proportion of the overlapped patients for each pair of the temporal phenotype and TAR clusters. Although the Jaccard Index seemed to be useful to measure the overlap between H and TAR clusters, the resulting value might not able to indicate the likelihood of the observed overlap. Therefore, the probable score of the random overlap was modelled using a binomial distribution in Normal Approximation for the Binomial Approximation of the Hypergeometric distribution (NBH) metric, which was introduced by Swift et al. [116] and calculated from Equations 6.5. In Equations 6.5, n was assumed to be the number of patients in the union of C_i and C_j. If both n and npq were large, the binomial distribution could be approximated by a normal distribution.

$$P(\text{observing } x \text{ from group } j) = \binom{k_j}{x} p^x \, q^{k_j} \;,\; x = Jaccard(C_H^j \;,\; C_{TAR}^i) \tag{6.5}$$

$$\text{,where } n = |C_H \cup C_{TAR}| \quad , s_i = |C^i_{TAR}| \quad , k_j = |C^j_H| \quad , \quad p_i = \frac{s_i}{n} \text{and} \quad q = 1 - p.$$

The NBH was utilised to evaluate the probability of observing an overlap between each pair of clusters from C_H and C_{TAR}. Thus, obtaining a very low NBH probability represented a possibly considerable overlap between two clusters from different data sources. A low value (probability) indicated that the chance of observing a given overlap was very low, especially by a random chance. For example, C^i_{TAR} of size s_i, where i indicates the temporal phenotype cluster's number, compared to C^j_H of size k_j, where j indicates the TAR cluster's number.

6.5 Experimental Results in the Patient Personalisation

In this section, TAR clusters were validated and compared to the temporal phenotype clusters to understand whether the temporal phenotype could reduce uncertainty, which was caused by the complex relationships among the temporal complications. In Table 6.3, the most frequent and interesting itemsets (ordering pattern of complications) were identified by corresponding object in Table 6.2. In order to quantify a distance between two heterogeneous complications-rules, one solution could be to use cluster rules based on their features (support, confidence and lift). However, these measures could only capture the interactions of sub-rules on the dataset only characterising one single rule. Agglomerative hierarchical clustering was employed in order to build homogeneous groups of sub-rules. Then, MCI analysed sub-rules (antecedents and consequents) as input and produced the minimum coverage itemsets (41 objects found) as output in Table 6.2 The distances among the frequent itemsets were aggregated for two patients within a cluster by using Jaccard Distance, which was applied to the group of the object associated with the corresponding pattern. Furthermore, more in-depth analysis of the correlation/causation between rules were proposed by using Ward's method. This hierarchical clustering analysis obtained TAR clusters based on the dissimilarities found among the itemsets (1 - $|correlation|$). Association rules were grouped according to the descriptors (itemsets or objects), as seen in Table 6.3. In the next stage, more metrics were employed to validate the H clusters.

Table 6.4: Proportion of patients with the complication co-occurrence pattern for C_{TAR} and C_H. On the right-hand, there are comparison results of the complication rates occurring in each cluster.

TAR Cluster	RET	NEU	NEP	LIV	HYP	Interesting Itemsets
C^1_{TAR}	0	15	10	16	100	{HYP}{LIV,NEU}
C^2_{TAR}	0	80	10	40	0	{NEU,LIV}
C^3_{TAR}	96	25	8	13	90	{RET,HYP},{NEU}{LIV}
C^4_{TAR}	67	0	33	17	50	{RET,HYP,NEP,LIV}
C^5_{TAR}	30	40	40	60	100	{HYP,LIV}{NEP,NEU,RET}
C^1_H	7	11	8	13	61	{HYP,LIV,NEU}
C^2_H	13	10	6	7	63	{HYP,RET,NEU}
C^3_H	4	16	11	13	56	{HYP,NEU,LIV,NEP}
C^4_H	12	13	6	11	57	{HYP,NEU,RET,LIV}

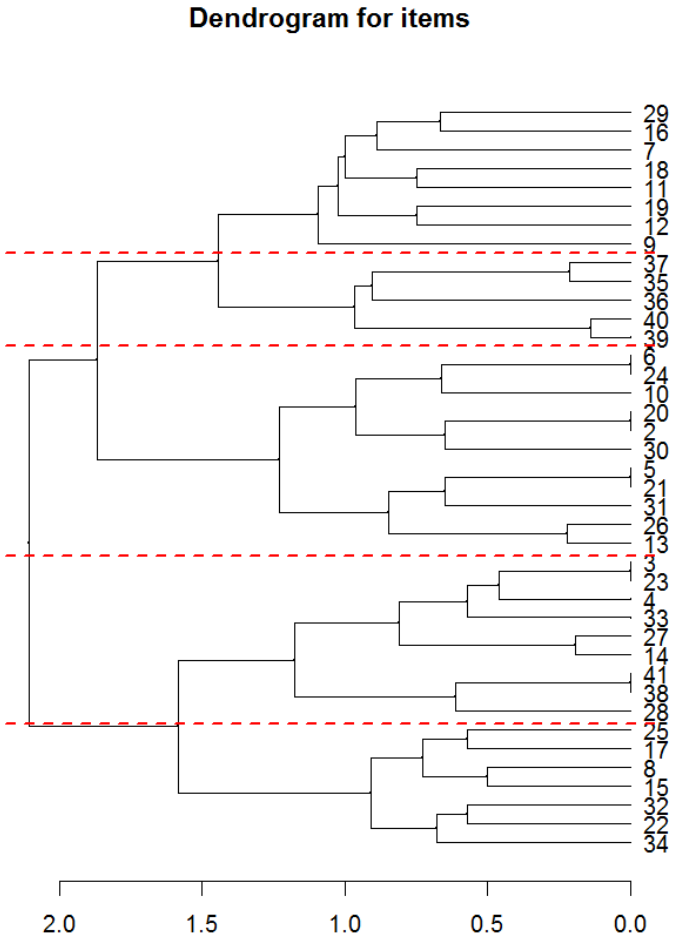

Figure 6.3: Hierarchical Clustering applied on the objects (interesting itemsets in Table 6.2). X-axis and Y-axis illustrate Jaccard Distance among objects and objects id, respectively. The red lines split the objects into five clusters.

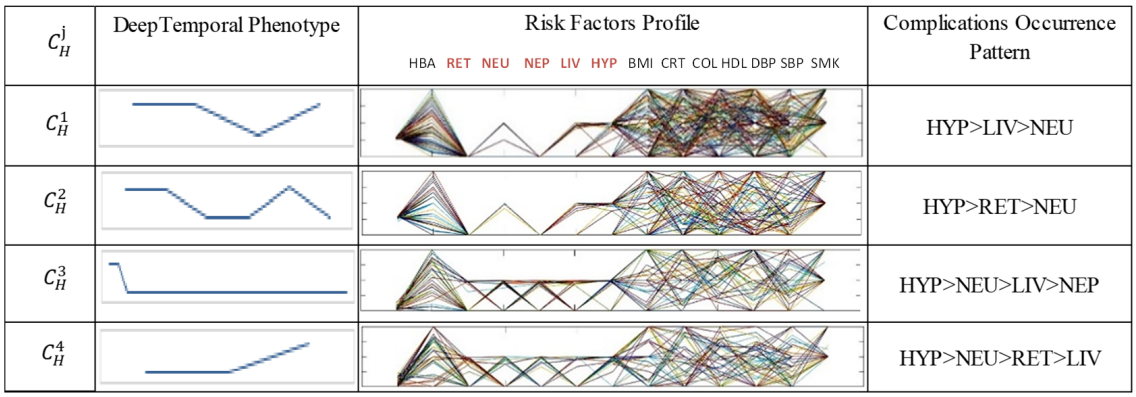

Figure 6.4: The discovered Temporal Phenotype for C_H, the corresponding risk factor profiles, and the most frequent ordering pattern of the complications (labelled in red).

6.5.1 Discovered Clusters

So far T2DM patients were stratified based on two different clustering groups, including C_{TAR} and C_H. The hybrid technique has obtained the initial clusters of the temporal association rules. In Figure 6.3 showed a dendrogram of the TAR clusters based upon the objects. Five TAR clusters (C_{TAR} = { C^1_{TAR}, C^2_{TAR}, C^3_{TAR}, C^4_{TAR}, C^5_{TAR}}) were obtained by using Ward's method, which established the adaptable number of C_{TAR} to be compared to four latent phenotype clusters. The optimal number of clusters was also validated by using the Elbow Method [145]. The time series clustering method identified the H clusters (in Figure 6.4, there were four T2DM patient clusters as the discovered hidden variable C_H = {C^1_H, C^2_H, C^3_H, C^4_H}) based on the DTW distance. As it was discussed previously in Chapter 5, each cluster had a unique profile of the latent variable and risk factors; thus, it could have various ordering patterns of complications. A symbol of ">", in the right-hand column in Figure 6.4 indicated how the most frequent complications could be prioritised based on their number of occurrences. For example, it demonstrated a complication in the left-hand side occurred before the complication in the right-hand side with a higher occurrence rate. In the next stage, comparisons between these two clusters were made by using unrelated rules on their associated complications.

Table 6.5: Probabilities of the Jaccard Similarity, Overlapped Rate (Đ), and NBH across C_H and C_{TAR}.

	C^1_{TAR}		C^2_{TAR}		C^3_{TAR}		C^4_{TAR}		C^5_{TAR}	
	NBH	Đ	NBH	Đ	NBH	Đ	NBH	Đ	NBH	Đ
C^1_H	<0.001	90%	0.580	45%	<0.001	4%	0.480	38%	0.490	40%
C^2_H	0.064	66%	0.072	0%	0.290	16%	0.440	13%	0.092	0%
C^3_H	0.032	60%	0.630	9%	<0.001	28%	0.610	13%	<0.001	40%
C^4_H	<0.001	55%	0.045	45%	<0.001	52%	0.170	38%	0.530	20%

Table 6.6: Prediction performance of T2DM complications for each dataset assessed by using causal inference.

Complication	DS	DS1	Low	High	Evidence (E)	P(MAP|E)	P(E)	P(MAP,E)
NEU	✓		✓		HYP,LIV	0.57	0.23	0.13
NEU		✓		✓	HYP,LIV	0.83	0.29	0.24
NEU	✓		✓		HYP,LIV,RET	0.57	0.23	0.13
NEU		✓		✓	HYP,LIV,RET	0.85	0.03	0.02
RET	✓		✓		HYP,LIV	0.71	0.23	0.16
RET	✓			✓	HYP,LIV	0.87	0.29	0.27
NEP		✓		✓	HYP,LIV	0.76	0.29	0.22
NEP		✓	✓		HYP,LIV,RET,NEU	0.76	0.02	0.02
NEP	✓			✓	HYP,LIV,RET,NEU	0.86	0.03	0.02
SMK		✓	✓		NEP	0.33	0.49	0.16
SMK		✓		✓	NEP	0.99	0.49	0.50

Table 6.7: Overall prediction accuracy of T2DM complications for patients in DS is compared to DS1.

Target Complication	Accuracy in DS	Accuracy DS1
NEP	0.81	0.93
LIV	0.77	0.88
HYP	0.91	0.99
NEU	0.76	0.81
RET	0.81	0.79
All Complications	0.81	0.88

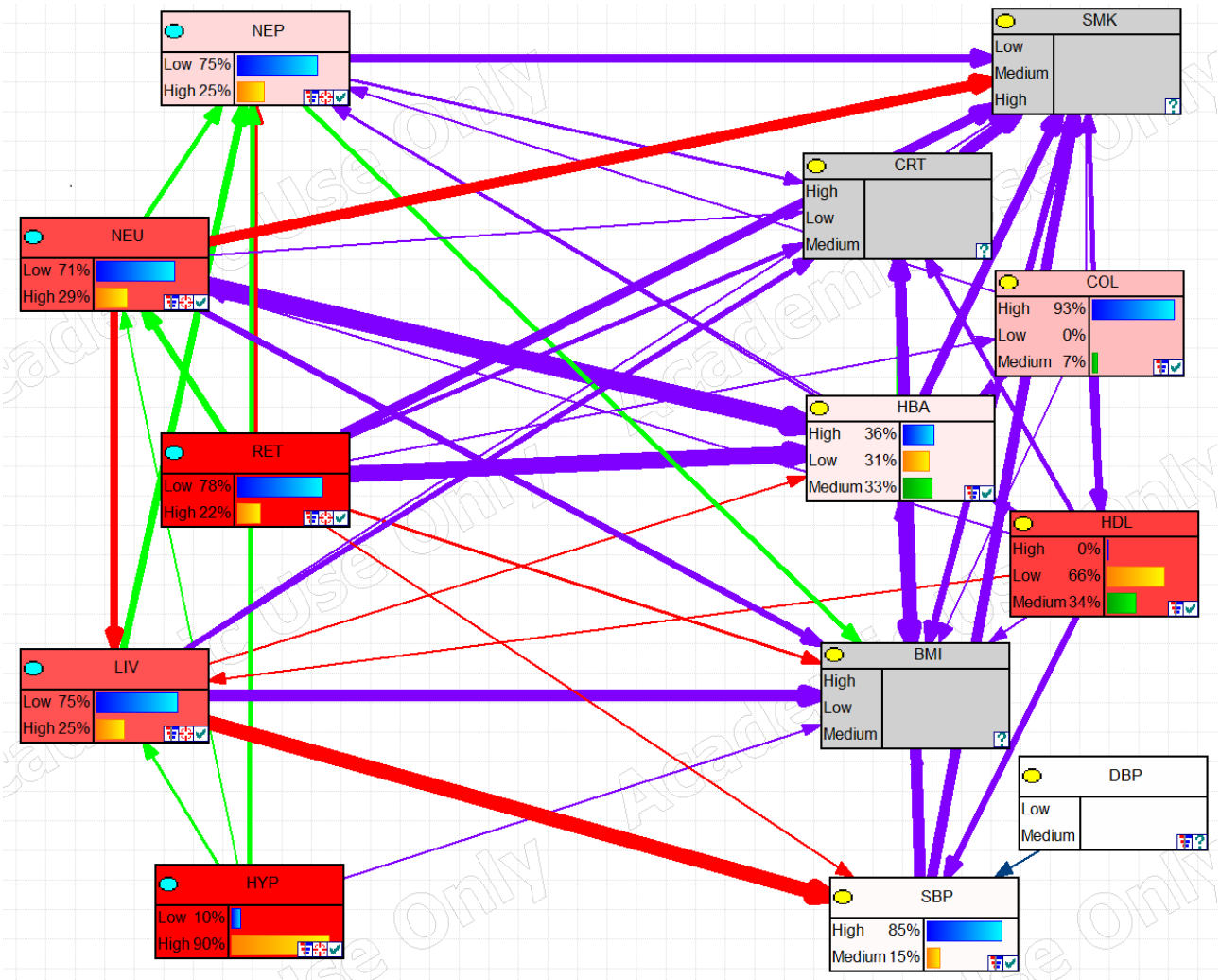

Figure 6.5: An influence diagram to represent Bayesian Structure applied to DS.

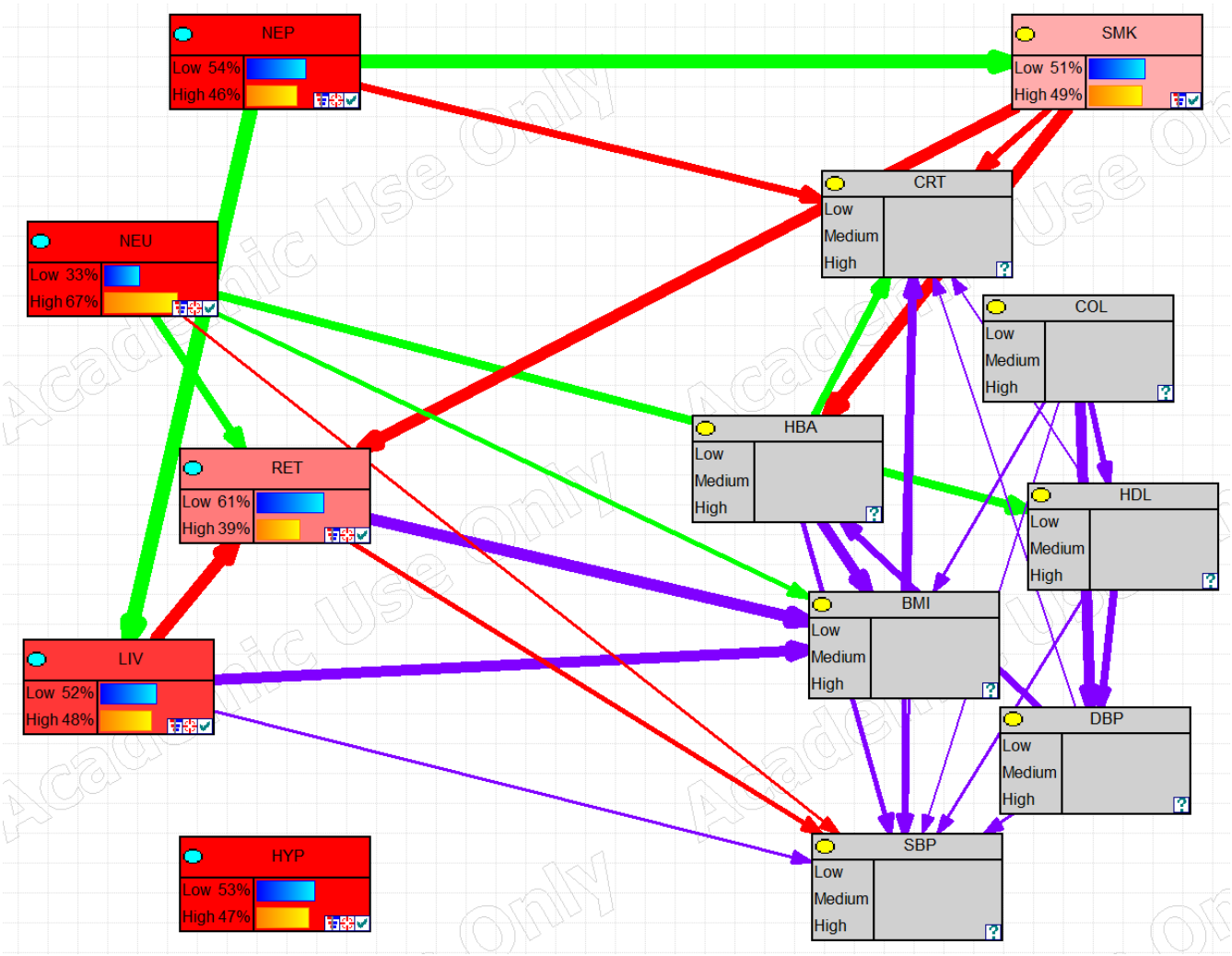

Figure 6.6: An influence diagram to represent Bayesian Structure applied to the subgroup of patients in DS1.

6.5.2 Clustering Comparison and Findings Validation

As the previous stage stated that the temporal phenotype clusters could be analysed by applying several validation strategies, while it has been compared to the TAR clusters. These strategies ensured a more appropriate decision for discovering the most meaningful subgroup of patients as well as explaining the behaviour of the temporal phenotype. The external validation strategy assessed the similarities among subgroups of patients within C_H, whereas they were clustered based upon different data sources (C_{TAR}).

In Table 6.5, the Jaccard Dissimilarities were calculated to differentiate between two patients, one selected from C_H and another patient from C_{TAR}. In addition to this metric, the overlap rate was calculated. An overlapping pair of patients could be detected if an occurring pattern of complications in any of C_H found in the right-hand column in Figure 6.3 resembled interesting itemsets belonged to C_{TAR}.

Alternatively, if two complications-rules have not been shared between two patients, it could be assumed that these patients did not belong to both clusters. For example, the intersection of C_H^4 and C_{TAR}^3 (in the right-hand column of Table 6.5) revealed a significant number of patients (with an overlap of $>50\%$) revealed a significant number of patients shared a similar complications co-occurrence pattern C_H^4 with the complications pattern of $\{HYP, NEU, RET, LIV\}$ and C_{TAR}^3 with the ordering pattern of $\{RET, HYP\}, NEU, LIV$ have also coincided. The intersection of C_{TAR}^3 and C_H^4 showed that they greatly resembled each other, and it revealed an important link between the two clustering methods.

Overall, it is believed that there was a strong link between C_H^1 and C_{TAR}^1 where both clusters were sharing a similar complications co-occurrence pattern of $\{HYP, LIV, NEU\}$. Patients within C_{TAR}^3 were more likely to develop RET, HYP, NEU and LIV with the occurrence percentages of 96, 90, 25, and 13, respectively. Similarly, C_H^4 was more likely to develop $\{RET, HYP\}, NEU$ whilst LIV were not likely to be developed in patients within the corresponding cluster (which was seen in Table 6.4), revealing a significant as well as a meaningful relationship between those two clusters (C_H^4 and C_{TAR}^3). Furthermore, a C_{TAR}^i pattern, e.g., $\{RET, HYP\}, \{NEU\}, \{LIV\}$ revealed that $\{RET, HYP\}$ was more likely to be seen than NEU, and NEU was more likely to be developed compared to LIV and the rest

of complications were not likely to be developed in patients within the corresponding cluster C^3_{TAR} (as seen in Table 6.4).

Additionally, as shown in Table 6.5, $C^2_H \cap C^4_{TAR}$ with the lowest NBH probability of <7.9E-90 and second highest overlapped number of patients of 25 per cent revealed a significant and meaningful relationship between those two clusters (C^2_H and C^4_{TAR}). In this chapter, the dissimilarities (distances) between clusters are analysed as the interestingness to filter discovered rules, which was optimised after filtering out uninteresting rules effectively. These results will attract a domain expert to choose interesting patterns from the remaining small set of rules. For instance, the itemsets consisting of similar items are uninteresting, despite the fact that the frequent itemsets with different items are interesting. Figure 6.3 represents a dendrogram of the TAR clusters based upon Table 6.5, $C^2_H \cap C^4_{TAR}$ with the lowest NBH probability of <7.9E-90 but with the second-highest percentage for the overlapped patients (25 percent) revealed a strong relationship between C^2_H and C^4_{TAR}.

In the next section, the prediction results are analysed to investigate the differentiation of DS1 and DS in terms of how accurate the hybrid complications prediction is in the personalised dataset compared to the raw dataset.

As can be seen in Table 6.4, for the patients within C^4_{TAR} the chances of having RET, HYP and NEP were approximated by percentages of 67, 50 and 33, respectively. Similarly, the chance of having a consequence of RET, HYP, and NEP for patients in C^2_H was high (see evidence in Table 6.4).

Association rules are grouped according to the descriptors (itemsets or objects), as seen in Table 6.3. Whereas, they are not grouped according to their coverage, as explained in Algorithm 1 MCI. Each of patients within DS that have been diagnosed with the a similar occurring pattern of complications (the corresponding frequent itemsets) are gathered in one cluster. The distances among the frequent itemsets are aggregated for two patients within a cluster by using Jaccard distance, which are applied to the group of the object associated with the corresponding pattern.

6.5.3 The Meaningful Subgroup of the Personalised Patients

This section attempted to investigate how the similarities between the C^i_{TAR} and C^j_H could be validated and give meaning to the temporal phenotype. Figure 6.4 represented patients in C^1_H, with a decreasing and an increasing pattern in their deep temporal phenotype, shared similar trajectories over the observed risk factor profiles. Almost 90 per cent of patients within C^1_H was found in C^1_{TAR}. More importantly, it was significantly validated from a statistical point of view as the likelihood of randomly observing this overlap was very low with an NBH probability of <0.001 as shown in Table 6.5. Thus, there was sufficient evidence to suggest that nearly all patients belonged to a similar TAR cluster (C^1_{TAR}). It also appeared that the most frequent Ordering Pattern of Complications of HYP, LIV and NEU belonged to C^1_H matched $\{HYP, LIV, NEU\}$ belonged to C^1_{TAR}.

More importantly, having considered that patients within C^1_H and (C^1_{TAR}) were selected from two different data sources, not only statistically validated the clusters but also revealed the meaningfulness of the temporal phenotype. Therefore, patients in the intersection of C^i_{TAR} and C^j_H ($C^i_{TAR} \cap C^j_H$) with the highest similarities among other clusters might represent a link between their latent phenotype and complications-rules. The most significant intersection of the TARs and latent phenotype clusters ($C^1_{TAR} \cap C^1_H$) was considered as the most informative (meaningful) subgroup and thought as DS1 (see Figure 6.1).

6.6 The predictive Strategy: Improving the Prediction Performance of the Complications Sequence

In this section, the prediction performance of the underlying patterns of complications for these patients within DS1 (which discovered using the descriptive strategy) was analysed and compared to all patients belonged to DS. It also suggested that DS1 (by personalising patients) could be considered as a dataset with less uncertainty than DS. In order to describe the inference problem in this thesis, the causal relationships seemed to be a reliable option to represent static and dynamic correlations between T2DM risk factors. The causal inference has a greater focus on distinguishing causes from other associations than on uncovering detailed

temporal relationships. Therefore, in this section, several predictive strategies in order to test whether the descriptive approaches have contributed to improving the prediction performance of the ordering patterns of complications. Therefore, in the following stages, several techniques have been discussed, such as Bayesian analysis, Optimal Posterior Likelihood Algorithm and Sensitivity analysis.

6.6.1 Bayesian Analysis and Optimal Posterior Likelihood Algorithm

Previously, it was suggested that an appropriate ordering pattern of complications could be predicted for each patient visit based on prior knowledge as well as current risk factors and complications. To better predict the ordering patterns of complications, here, an approach was utilised to not only approximate the posterior likelihood of the complication co-occurrences but to optimise the Bayesian parameters simultaneously. The dangers of these microvascular conditions can be significantly reduced by eliminating the chance of developing further complications. For this purpose, the posterior likelihood of the developing complications is approximated by using a "Maximum A posteriori Probability" (MAP) algorithm [58, 60] which converged toward the set of parameters. In T2DM patients' model, MAP employed an iterative strategy to investigate the maximum probability for the parameters. In addition, an optimisation procedure was required to produce optimal posterior results along with the evidence. Therefore, the simulated annealing algorithm [51] was aggregated to the stochastic simulation of the Hidden Markov Chain, which relied on data augmentation in the same way as the EM algorithm. It could be concluded that the choice of clinical decision with the highest expected gain seemed to be an optimal option which has been often chosen by clinicians. Therefore, the following experimental results have provided a possible solution to reduce human mistake by computing an expected utility as a likelihood of each decision alternative.

6.6.2 Experimental Findings and Overall Prediction Accuracy

Table 6.7 demonstrated a significant enhancement in the prediction accuracy of the hybrid complications by comparing DS1 against DS. This comparison was based on an optimal posterior likelihood of a High or Low clinical level. For example, the optimal posterior likelihoods

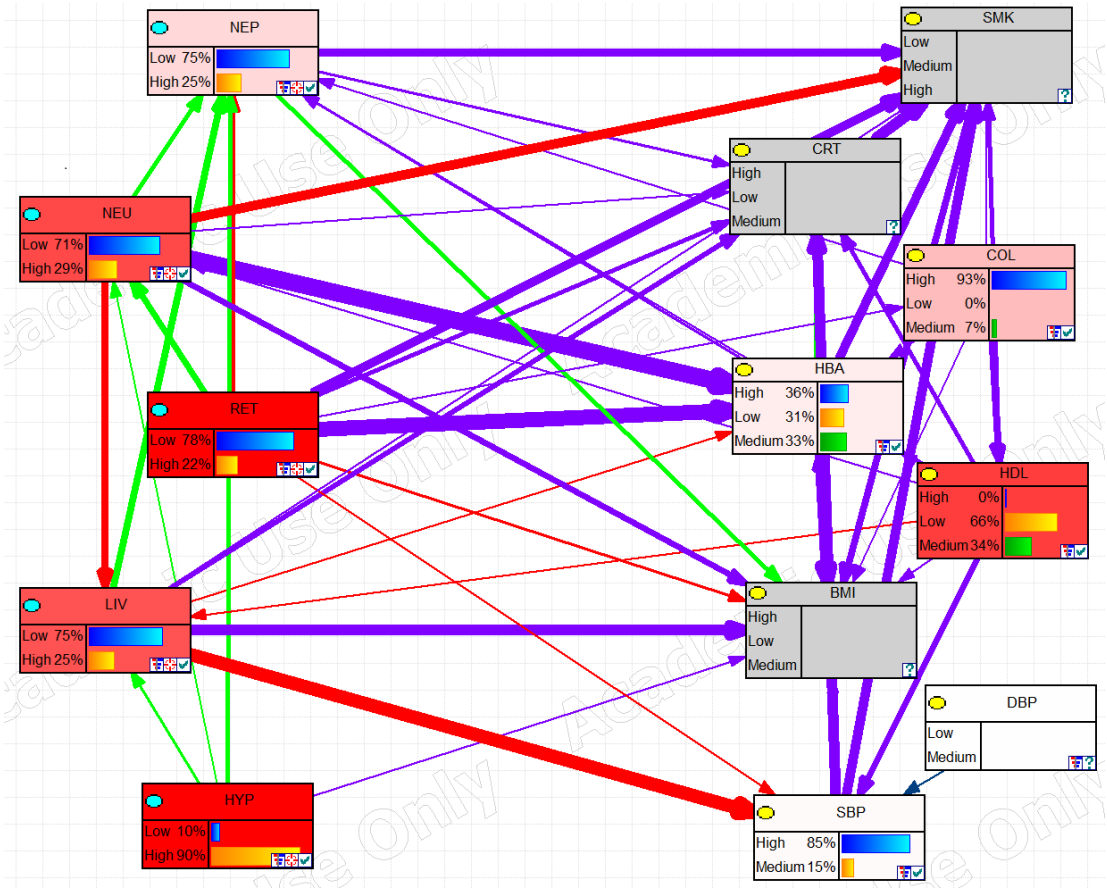

Figure 6.7: An influence diagram to represent Bayesian Structure applied to DS.

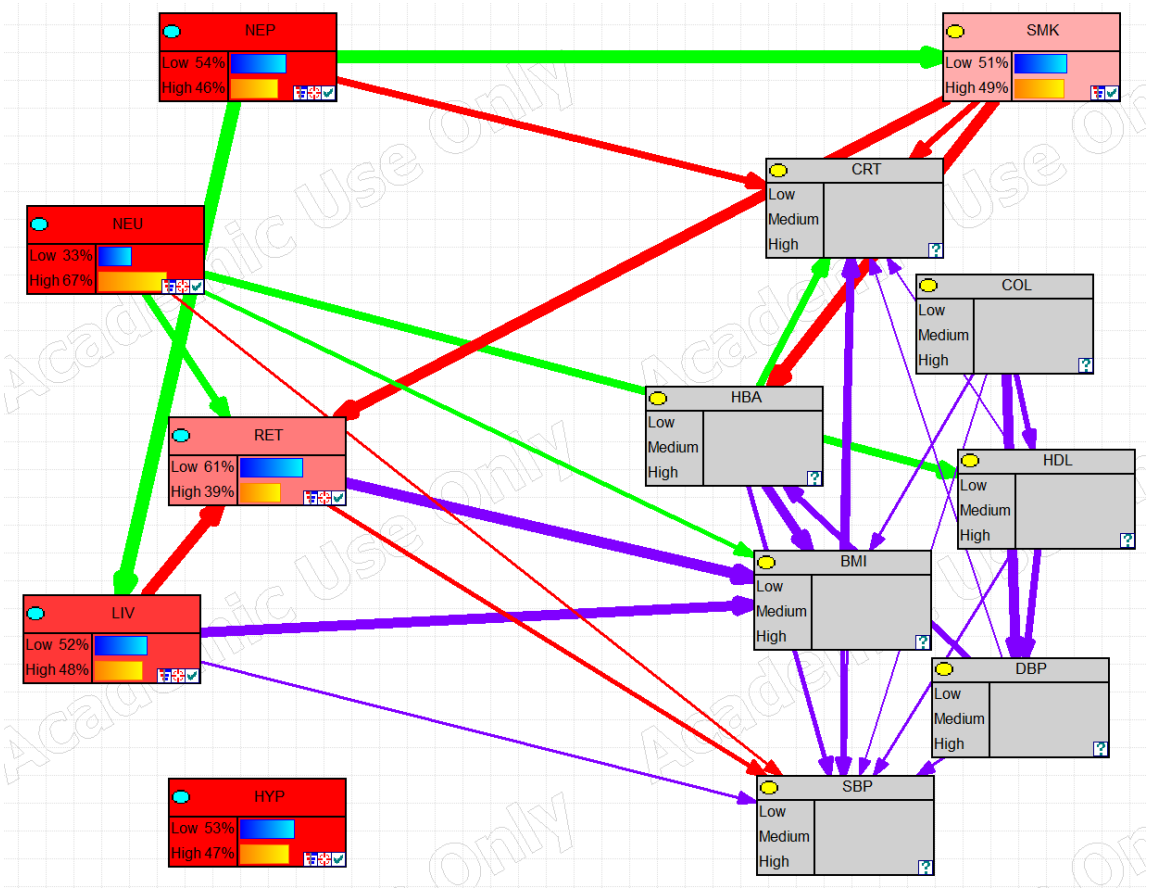

Figure 6.8: An influence diagram to represent Bayesian Structure applied to the subgroup of patients in DS1.

of developing RET, NEU, LIV and SMK were compared between DS1 and DS, with a consideration of prior complications (e.g., those patients who have already developed complications such as HYP and LIV). Thus, the prediction accuracy of HYP, LIV and NEU in DS1 were approximated as 0.99, 0.88 and 0.81, respectively. This suggested that the overall prediction accuracy across all complications in DS1 was 0.88 compared to a lower overall accuracy of 0.81 in DS, which is shown in Table 6.7. Similarly, the prediction accuracy of HYP, LIV, and NEU in DS were not significant compared to DS1 by 0.90, 0.77, and 0.76, respectively. Having shown these extensive findings, the prediction accuracy of complications the meaningful subgroup of patients in DS1 and relatively, their ordering patterns have been significantly improved while compared to DS. It can be challenging to predict a target complication without considering its associated complications and decide whether a diagnostic test result will be positive or negative. Therefore, sensitivity analysis has been applied to the Bayesian structures of DS and DS1, which was discussed below.

6.7 Summary

This chapter has explored various descriptive and predictive strategies that enable personalised patient analysis. In the descriptive study, two different subgroups of patients were identified from two clustering methods using two sources of data. Firstly, subgroups of patients were obtained based on the temporal phenotypes, which was introduced as the deep temporal phenotype in the previous chapter. Secondly, another subgroup of patients was extracted by using the hybrid methods applied to the temporal association rules of complications. The validation and clustering comparison strategies were applied to these two types of clusters, and the underlying patterns of complications were explored for specific patient clusters. The semantics of the subgroups provided by a combination of the temporal phenotypes and the TARs revealed interesting clinical characteristics. It seemed to be evident that not only the prediction performance of the underlying complications-rules was significantly improved, but also it eased the explainability of the latent variables. To conclude, the next chapter discusses the novelty of the contributions of the work in the previous chapters, along with a discussion of their main

limitations while providing recommendations and directions for future research.

Chapter 7

Conclusion

This Chapter first lays out the objectives and re-states the contributions of this thesis by providing solutions for disease prediction. Furthermore, it discusses the generalisability of the models for the broader domain and anticipates potential criticisms of the research. It then recommends some insight into important future practice to diagnose complex diseases such as diabetes and explain so-called black box systems.

7.1 Objectives and Contributions

The analysis of complex patient models in the literature in Chapter 2 has revealed that mortality often occurs due to complications caused by the disease and not the disease itself. This is because the impact of associated long-term complications as well as unmeasured variables, are underestimated. The main aim of this study was to predict these complications that are associated with diabetes and interpret the computational outcomes using innovative, analytical, and methodological approaches motivated by the clinical need for transparent longitudinal data modelling:

7.1.1 Utilising appropriate data mining (supervised and unsupervised learning) approaches to modelling disease

Chapter 2 showed that appropriate data mining techniques could be employed to model complex interactions among the complications and uncover the underlying pattern of hidden/observed risk factors. In Chapter 3, probabilistic, statistical approaches and time series predictive models were employed for the early prediction of the complications from the diabetic patients' follow up visits at the IRCCS Istituto clinic scientific (ICS) Maugeri of Pavia, Italy. The proposed approach effectively integrated Bayesian methods with latent variables identification.

7.1.2 Addressing data imbalance

In this study, DBN models were extended and adjusted to handle highly unbalanced time series clinical data by proposing the following techniques:

- Time Series Bootstrapping: In Chapter 3, the imbalanced data was re-balanced by utilising Time-Series Bootstrapping (TS Bootstrapping), which was adapted to the standard latent model to fit T2DM longitudinal data. The TS Bootstrapping technique then was integrated with the enhanced stepwise approach in Chapter 5. The finding obtained in Section 3.7 and Section 5.4.3 showed that the re-balancing approach for unequal time series along with hidden variable discovery led to an improvement in the predictive performance over standard probabilistic models, especially those with no hidden variables.

- Pair-sampling strategy: In Chapter 4, a pair-sampling strategy was employed to effectively address imbalanced data before learning the structure of hidden variables in which stratified T2DM patients into two types of cases, positive and negative. The obtained re-balanced data not only revealed an enhancement in the prediction results but also it reduced the bias and uncertainty to learn data.

7.1.3 Modelling complex interactions among both observed disease risk factors/complication and unmeasured effects using a targeted hidden variable approach:

- Discovering a hidden variable and finding its precise location within the DBN structure: The AI methodology in Chapter 3 expanded the current probabilistic AI models to predict disease when addressing variability in complicated patient results. K2 and REVEAL algorithms were used to learn the network structure in the standard latent model. Bayesian modelling was chosen to discover a hidden variable effect on the developing patterns of complications. In Chapter 4 to have a better understanding of the latent effects and to show how the complications could be predicted accurately, an intuitive stepwise method was proposed by the employment of Induction Causation (IC*), and then with the addition of a discovered latent variable at each step (to be considered as prior knowledge in the following step). This helped to discover and understand the semantics of the hidden variables and target their precise location within the network structure of risk factors.

- Obtaining an optimal number of hidden variables in a stepwise approach to avoid creating overly complex models that risk overfitting and becoming "black box" in nature: As shown in Chapter 4, multiple hidden variables were discovered to take a deeper look at latent variables, and better understand the relationship among disease risk factors and the latent factors. For example, a latent variable may represent a subcohort of patients who are at higher risk of certain complications while other groups of patients are not.

- Incorporating a combination of the IC* algorithm and Mutual Information to understand the impact of the latent variables: The enhanced method in Chapter 5 reduced the number of uninformative hidden variables by measuring the mutual information of links between disease risk factors and obtained a better predictive performance (to results in Chapter 3 and 4).

7.1.4 Personalising and handling the variability of progression in patients (via a temporal phenotype)

This was performed throughout this thesis (with main focus in Chapter 5-6) to demonstrate that a significant change in the prediction outcomes could be obtained by applying the latent model to each patient subgroup on a stand-alone basis. The following methodologies were utilised:

- The characterisation of temporal phenotypes from discovered hidden variables: In Chapter 5, the most influential latent variable was discovered to derive the temporal phenotype, which extracted the most descriptive data point within the centre of discovered subgroups of patients.

- Using a combination of time-series clustering with dynamic time warping and the Jaccard index to group patients: Subgroups of patients were extracted based on two data sources: In Chapter 5, one subgroup of patients was characterised based on temporal phenotypes by using Dynamic Time Warping. In addition, time-series pattern clustering was also used to organise patients on the basis of their corresponding rules associated with the complications (illustrated in Section 6.5.1).

- The discovery of latent temporal phenotypes was combined with Temporal Association Rule Mining to find similar subgroups of patients that aids explanation: In Chapter 6, Temporal Associations Rules and pattern discovery were designed to assess the subgroups of patients mentioned in Chapter 5 based on temporal phenotypes. This was shown by the creation of phenotypes that differentiated distinct patient groups over the period of disease progression by defining the most important subgroup of patients.

- This thesis rejected the claim raised in [65] for the proposed clustering approach with respect to T2DM dataset. A research conducted in [65], STS (Subsequence Time Series) clusters are required to conform according to a certain phenotypically impossible restriction in every data, and as such, the groups identified by any clustering method are inherently meaningless. The promising results in this thesis contrasted this claim

by testing the meaningful subgroup and showing its usefulness. As a result of these, it questioned the clustering approaches by employing the comparison techniques to ensure the worthiness of the discovered clusters found based on temporal phenotype, from both qualitative and quantitative points of view. To this end it employed several evaluation strategies categorised in Chapter 6 under Internal validation and External validation sections. For example, it assessed the obtained clusters from different data sources by calculating the proportion of the overlapped patients between any pairs of those clusters based on Jaccard Index. More importantly, as Jaccard index could not indicate the likelihood of the observed overlap, it calculated NBH measure to statistically test whether the outcomes are reasonable of random cluster centres. The obtained meaningful patient groups in Table 6.5 the first cluster identified based on temporal phenotype as well as the fist cluster discovered under TARs clustering approach, Jaccard index and NBH metric were 90 percent and bellow 0.001, respectively. From these results, it seemed to be evident that high overlap rate is not random and hence these clusters were meaningful as the clustering outcome was not independent of the patients data.

7.2 Evaluation

The generalisability of the results presented in this thesis is subject to certain limitations as follows:

This research was conducted to explain and discover the unmeasured factors with a few patients and relatively few features. Thus, this thesis focused on time-series complex clinical dataset like T2DM, which was a small-sized dataset with an unequal number of patient's follow up visits (which is common in clinical data). This study was specific to T2DM concept and Bayesian modelling; hence, one fundamental criticism could be the bias towards this dataset and whether the method could be developed in other fields of clinical data in the future. This was because the proposed methodologies were suitable for a specific type of longitudinal data which potentially involved latent variables. Nevertheless, it is very likely that unmeasured effects are common to many if not most clinical datasets.

A limitation of the proposed stepwise approach (Chapter 4-5) could be the stopping rule to the approach, and in some cases, an additional uncontrolled factor could be the possibility of overfitting so that accuracy starts to drop. Classification accuracy could be monitored and used as a stopping condition (i.e., if it drops significantly). There is also room for further progress in determining the optimal number of latent variables using Partial Least Squares (PLS). Due to the difficulties in explanation and constraints of the latent factors in the black box model, there is a need to seek more advice from clinicians in interpreting latent variable and their cause and effect relationship toward other T2DM risk factors and complications as well as the disease prediction process.

7.3 Future Work

In order to help overcome the limitations discussed in the previous section, the following recommendations are suggested:

7.3.1 Development and extension of the proposed explanatory model

The originality of the proposed thesis consisted in its innovative, analytical, and methodological strategies to predict and explain complex clinical data to improve patients' quality of life. A natural progression of this work for a better generasability should involve extending the latent DBNs model with more hidden variables to capture a greater variety of unmeasured factors to characterise critical changes and produce interesting findings that account more for better explainability and predictability. In addition, to address the limitation related to the small-sized dataset, this work could be extended to further investigation and experimentation into clinical impacts and environmental factors, such as family history, pollution, and glucose. More research also might be conducted to monitor disease progression effectively and detect the underlying patterns of complications, which could provide clinicians with a better understanding of the obtained findings. For example, a greater focus on phenotype discovery could enable assessment of the long-term effects of the temporal phenotype on the patient, which might be done by following qualitative approaches to support the obtained findings from the

biomedical literature.

7.3.2 Development of alternative models to explain complex disease progression

Further research is also needed to develop other models which can explain the complex data. For example, one alternative solution to better interpret clinical data can be decision trees or schematic models where the reasoning becomes more straightforward and thus more explainable in comparison to black box approaches. Nonetheless, with the utilisation of decision trees, it may be difficult to distinguish complex and time-based clinical data on its own. This is because in black box AI models it can be challenging to determine, from temporal clinical data alone, what is triggering the visible patterns to separate the underlying causes into meaningful causes that can help patient stratification, disease prediction and a deeper understanding of the disease process.

One approach to deal with the above issues can be the use of AI techniques such as Deep Learning, which has become ubiquitous to provide a high-performance prediction. Although this approach often provides an early and accurate prediction of disease, understanding its mechanisms has become a significant concern worldwide when the goal is to gain clinicians and patients trust. The reason behind this is due to overuse of hidden variables and lack of explainability, which can cause sources of complexity and uncertainty in the patient model. Overall, as observed from prior studies and mentioned in Chapter 1, there has been a balance that needs to be made between the accuracy of complex/deep models such as Deep Learning and the interpretability of models (e.g., decision trees) that aims to model data in a more human-like way such as AI expert systems. There have been a few attempts to make this balance in disease progression. For example, Google's AI Doctor was proposed, which is designed to reproduce current problem-solving methods (e.g., the detection of cancers) [40, 68, 84, 134]. Furthermore, it demonstrated how an explanation of such methods could be used to make further predictions, which are generated by local classifiers from conventional image classification networks to a more focused clinical application.

Nevertheless, the concepts of Google's explanatory power are subject to some limitations

as follows:

1. there is still a lack of comprehension of what hidden layer and artificial neurons will offer in determining the underlying causes of disease.

2. there have been many challenges to understanding and designing numerous hidden layers on a meta-level, which are required for more in-depth modelling.

3. selection bias is another potential concern because Google's method fails to consider the different categories of these hidden layers to determine the underlying causes of disease.

4. to understand how nodes become active, a detailed analysis of these hidden layers is needed; however, these methods not only overlooks the hidden layers but also ignores pre-existing knowledge of these layers. For example, it often does not make any attempt to quantify the association among these nodes.

5. it fails to draw a conclusion based on the structural nature of each neuron in each network. Although these nodes also are triggering as groups of interconnected neurons, they are limited at the same time and space.

These are the main weaknesses of the black box models as the stratification of a network for the categories of interconnected neurons would make its configurations even more abstractable.

In spite of the fact that the black box models are provided with sufficient data, from a clinician's perspective, Deep Learning seems to be overconfident. In addition, in the world that it is possible to fully allocate decision making to computer systems, confidence in AI systems will be hard to achieve. The reason behind this is that it represents a completely different health knowledge that can be generated without user intervention that needs to be understood by clinicians and patients to facilitate transparency. It is also due to several obstacles that arise in interpreting the findings, such as the scale of big data, complex interactions, and high-dimensional internal state.

Thus, the existing explanatory Deep Learning approaches would need to be adapted for further sophisticated longitudinal modelling strategy (rather than with a multivariate distribution) and should be simplified in several aspects. For example, if an object is detected,

an image detection machine can only focus on specific attributes including shape, colour and texture of the image, and then reduce the predictions to a mathematical method by checking the classification error and then background diffusion to improve the practices. In the future work, one approach that can be applied to a small-sized dataset like T2DM can be the use of Bayesian Neural Networks, which will deal with uncertainties in data and model structure by exploiting the advantages of both Neural Networks and Bayesian modelling. To conclude, AI can improve current methods of medical diagnosis in terms of interpretability while needing further evaluation to be trusted by both patients and practitioners.

7.3.3 Outstanding works and other types of data

The generalisability of the findings obtained in this study might be tested on other data with potentially non-stationary, complex, and incomplete data. For instance, the DBNs model and bootstrapping approaches could be used in educational data to predict student drop out and to extract knowledge to students' development, progression, engagement, and learning. Furthermore, the pre-processing approaches and statistical analysis have been applied to COVID-19 data at London North West NHS Trust. A similar patient model to this thesis was mainly employed, which primarily concentrated on helping NHS staff in their understanding of how COVID-19 spread and how they could be better prepared. [1]

7.4 Clinical and Computational Implications

This section summarises the clinical implications and shows how the obtained experimental findings in the previous chapters and their significance have led to developing explanatory AI models. This study offered several valuable insights into the prediction challenges in diabetes and similar diseases and explained how they could be tackled.

The results for an early prediction of T2DM complications confirmed that the proposed latent model can be employed as an indicator even outside of the FSM for problematic circumstances, as the data collection strategy and need for this research was discussed in MOSAIC

[1] Note that this is ongoing / very recent work and so not part of the thesis.

project [22]. Given the attributes of the dataset, there is a potential for developing the model for predicting the development of a disorder, while comparing its performance with that obtained by applying the well-established and commonly used UKPDS risk engine.

Throughout this thesis, appropriate machine learning techniques were conducted to model complex interactions among the complications, risk factors and unmeasured factors. For instance, the use of probabilistic graphical models provided a significant improvement in the accuracy of predictive models while reducing uncertainty in disease management. Having adopted DBNs to learn hidden risk factors and effectively understand the AI black box model was the key contribution of this research. The temporal phenotype was identified to represent the overall patterns of disease risk factors for each patient based on the discovered hidden variables over time. The descriptive analytics, in Section 5.4, provided valuable insights into the hidden variable effects on stratifying patients into different sub-groups, whether or not they developed the same complications. These findings also explained the influence of the latent variable on the bootstrapped data (as illustrated in the discussion section in Chapter 5). In Chapter 6, phenotype discovery was utilised to categorise and investigate meaningful subgroups of patients based on how an individual matches historical data. The hybrid type methods in discovering meaningful subgroups and explaining temporal phenotype also led to a better understanding of clinical data as well as aiding to interpret the unmeasured factors while demonstrating their risks.

Furthermore, to construct meaningful explanations of patients' subgroups in a precise prediction, several computational approaches inspired the detailed observations of this study as follows: The hidden variables were learned as a set of random variables in a DBNs structure and a graphical model of the probability distribution over the disease complications and risk factors. In Chapter 5, the parameters of the Bayesian model were legitimately influenced by the interactions among the risk factors over time. The hidden variables were extracted incrementally by utilising the enhanced stepwise IC*LS based on Induction Causation algorithm and Link Strength. A precise estimate of the uncertainty related to parameter estimation is essential to avoid misleading inference. This uncertainty is typically outlined by a confidence interval, which is professed to incorporate the actual parameter value with a predefined

likelihood.

To manage the uncertainty in the prediction, the experimental findings and their significance should be tested statistically and confined to Confidence Interval results derived from a randomly selected subset of T2DM patients. Thus, the effect of adding a hidden variable at each step of the enhanced stepwise was assessed on the bootstrapped T2DM patients in predicting a common complication of T2DM (e.g., retinopathy and liver disease). Furthermore, by looking at how these different structures perform within a DBN for predicting the appearance of complications, prediction performance was assessed in Chapter 5. The obtained findings illustrated that there was to be a general trend to improvement in accuracy as more hidden variables were added, but this improvement levelled out after a few steps (at the fifth step in the results were shown in Chapter 5). By effectively adding the discovered hidden variables to evidence has proved contribution in determining the most realistic structure of disease risk factors. The 95% confidence interval result demonstrated with high confidence that the IC*LS methodology resulted in a highly significant improvement in the classification accuracy, sensitivity and precision compared to the standard approaches in Bayesian modelling such as K2 and REVEAL algorithm as well as no latent variable approaches (as shown in Chapter 5).

The enhanced stepwise approach method along with phenotypic discovery uncovered the impact of the hidden variables on other risk factors/complications, which characterised the temporal phenotype to capture the most influential hidden variables (as demonstrated in Section 5.3 and 5.4). Furthermore, to construct meaningful explanations of patients' subgroups in a precise prediction, phenotypic discovery uncovered the impact of the hidden variables on other risk factors/complications. The quantitative analyses, in Chapter 5, provided valuable insights into the hidden variable effects on stratifying patients into different sub-groups, whether they developed the same complications. The use of clinician interpretation from a clinical point of view yielded insight into interpreting the latent states (looking at the associated distributions of complications). This further illustrated how phenotypes, that distinguished various patient populations, have been developed over the course of the disease experience (which was shown in Section 6.3).

Chapter 6 suggested that the latent variable explanation could be strongly connected to the

phenomenon of interpretability of the outcome considering correlation and causation among the complex interactions of complications/risk factors. This was because a patient must switch to another medication if more complications develop (which might be followed/caused by more complications). In Chapter 6, other subgroups of patients were identified based on time series clustering, Jaccard distance and then validated using a wide verity of comparison and statistical approaches such as quality metrics in Association Rule Mining, Jaccard Index, Normal Approximation for the Binomial Approximation of the Hypergeometric distribution (NBH) metric, causal inference and Bayesian modelling (which was illustrated in Section 6.4).

7.4.1 Main Findings

Based on these clinical implications and computational outcomes, this study certainly added to our understanding of the complex AI model in time-series analysis of clinical data. It has demonstrated the potential for the progress of the complex disease by utilising Bayesian modelling in an incomplete dataset, stratifying individuals to have personalised patients in precision medicine.

One purpose of this study was to assess the extent to which the latent variable predicted these comorbidities. The single most striking observation to emerge from the data comparison was the improvement of accuracy and sensitivity by adding latent variables. A clear benefit of anticipating latent variable in the prediction of comorbidities could be identified in this analysis. Furthermore, this study predicted whether the specific comorbidities, will be developed at the next visit or not. It also compared the result of applying imbalanced bootstrapped algorithms and making re-balanced dataset with imbalance dataset. In comparison, it was obvious that by using the algorithm, the accuracy would be increased significantly.

7.4.2 Clinical Recommendations

The key goal of this research was to model the variability of progression from person to person, identify sub-categories of disease within a cohort, and explicitly model the time-varying nature of disease by searching for parameters, structures and locations of change within the time-series, simultaneously. Overall, this thesis created new models to predict the onset of microvascular

complications, such as retinopathy and nephropathy. The experiments in which presented in this work were interpreted after being applied a Bayesian modelling approach, different ways of handling missing data (with or without imputation), different combinations of predictors (with or without lipid-related variables), and different ways to manage the class unbalance problem.

7.4.2.1 Hidden Variables Influence on Clinical Risk factors:

According to the clinical evidence in Diabetes literature and the experimental results obtained so far in Chapter 4 and Chapter 5, it is now possible to show how the targeted use of latent variables improves prediction accuracy, specificity, and sensitivity over standard approaches. This was achieved by using the cause and effect relationship among complications which described the association of an eventful complication. By observing the structure of the links with respect to the hidden variables on their Markov blanket, which can be seen in Figure 5.9. This showed that there was a strong relationship between *Hidden Variable 2* at the third step of the enhance IC*LS and T2DM key risk factors (e.g., HbA1c, BMI, liver disease, and smoking). Additionally, in the first step of the stepwise IC*LS shown in the second left DAG of Figure 5.9, the initial hidden variable (*Hidden variable 1*) is weakly linked to a small number of clinical factors, notably HbA1c, Liver disease and creatinine. However, as subsequent hidden variables are added at the second step, this structure changes. The second hidden variable (*Hidden variable 2*) is linked stronger to HbA1c by 35.9 and also is connected to more risk factors including *Hidden variable 1* (seen as the third left DAG in Figure 5.9).

In the third step as seen as the second right DAG in Figure 5.9, *Hidden variable 3* is closely connected to HbA1c by 63.2, and there is a 0 scored link between *Hidden Variable 3* and *Hidden Variable 2*. This is while *Hidden Variable 2* is strongly connected to *Hidden Variable 1*, which is scored 44.4. Thus, *Hidden Variable 3* (which is closely connected to HbA1c) seems to be irrelevant and independent of *Hidden Variable 2* (that is linked to nephropathy, liver disease and hypertension). At this step, there is a strong relationship between *Hidden Variable 2*, retinopathy, liver disease, DBP, SBP and smoking. Having considered these hidden variable results obtained from Figure 5.9, Diabetes literature (see evidence in [121]), and advice of clinician expert in diabetes, it was suggested that the presence of HbA1c was associated with

an increased incidence of nephropathy, while HbA1c emerged as an independent risk factor for developing retinopathy. Additionally, nephropathy and liver disease were independently associated with an increased incidence of hypertension in T2DM patients (see evidence in [120]).

7.4.3 Sensitivity Analysis and Cause and Effect Relationships

The proposed computational methodology and interpretation of the underlying patterns of complications were based on correlation and causation (both positive and negative). It was suggested that the joint implications of correlation and causality could provide more detailed questions about disease progression, considering a better understanding of the predictability of the complex models. This has been achieved by statistically testing multiple hypotheses (considering each causal relationship as a hypothesis). For example, Chapter 6 demonstrated why some symptoms of the illness were closely associated with other complications, and whether some of these complications have been developed in particular groups of patients but not in others.

To assess the obtained subgroup in terms of have achieved a higher performance he cause and effect relationships between complications were investigated by using influence diagrams demonstrated in Figures 6.7,6.8). It has been reported that the prevalence of microvascular and macrovascular complications could be caused by the associated complications [135]. Nearly 95% of T2DM patients who have developed a complication were also at increased risk of developing other complications. It appeared to be a considerable number of patients who have been more prone to develop complications once they were tested positive for another complication.

The prediction performance of a target complication was assessed when the clinical class as evidence were set to either the highest risk or the lowest risk level. For example, in Table 6.6, if HYP and LIV class values were set to their high clinical level, the probability of 0.83 for NEU ($P(NEU|\{HYP, LIV\})$ was higher than RET of 0.96 ($P(RET|\{HYP, LIV\})$. It seemed evident that if patients in DS1 have developed HYP and LIV, there were most likely to develop NEU compared to a higher probability of RET. It seemed that once a patient in DS1 has been tested positive for HYP and LIV, the likelihood of developing NEU was increased to 0.84.

Alternatively, with a hypothesis tested the patients in DS with a low probability (0.57) for developing NEU, if they have already developed HYP and LIV. Again, in Table 6.6, if the posterior likelihood of LIV in DS1 has risen to above 0.88, the growth of damaged eye cells (developing RET) dropped by 0.96. An observation can be ensured, the risk of developing RET for patients in DS1 appeared to be affected negatively, while LIV was occurred, which demonstrated by a thick red arrow in Figure 6.8. To emphasise consistency and reliability of the meaningful subgroup, in Figure 6.7, a thicker red-coloured arrow pointing to RET from LIV in DS1 compared to the no influence arrow in DS, as illustrated. In Figure 6.7, a thick purple edge from NEP to SMK revealed the development of NEP were followed/caused by SMK. Additionally, in Figure 6.8, positive causation was represented by a green edge from NEP to SMK. Hence, it could conceivably be hypothesised that once a patient has been diagnosed with NEP, the probability of being a smoker could have significantly been increased from 0.33 to 0.99 by comparing P(SMK|NEP) values between DS and DS1 (see Table 6.6). It was also suggested that if the patients in DS1 were at a low clinical level of NEP with the optimal likelihood posterior of 0.86, a higher risk of developing HYP, LIV, RET and NEU could be less likely. However, that posterior with the same evidence/prior knowledge but at the high clinical level was equal to 0.76 (see Table 6.6). This contrast provided some support for the conceptual premise that was often significant in at least one group of patients, which was the most interesting subgroup.

Taken together, these results suggest that there is an association between latent variables and mixtures of microvascular complications in the prognosis process of comorbidities.

7.4.4 Conclusion

One purpose of this study was to assess the extent to which the latent variable predicted these comorbidities. A clear benefit of anticipating latent variable in the prediction of comorbidities could be identified in this analysis. The single most striking observation to emerge from the data comparison was the improvement of accuracy and sensitivity by adding latent variables.

Furthermore, this study predicted whether the specific comorbidities, will be developed at the next visit or not. It also compared the result of applying imbalanced bootstrapped

algorithms and making re-balanced dataset with imbalance dataset. In comparison, it was obvious that by using the adapted algorithms in this study, the accuracy of the predictive model would be increased significantly.

7.4.4.1 MOSAIC Tool

The Data is belonged to the MOSAIC European Union project retrieved from MOSAIC website [22]. This work is mainly presented to provide the risk of complications, which will be included in MOSAIC instrument. Adopting DBNs to learn hidden risk factors and understanding the AI black box model effectively was the key contribution of this research. It aided to gain insight into it by understanding the unmeasured factor and discuss their dangers. The mosaic tool is exploited as an instrument to identify potentially critical behaviours that might need closer control to be considered in the analysis of clinical data from the FSM hospital dataset (Body Mass Index, glycated haemoglobin, lipid profile, smoking habit). The MOSAIC instrument, and the outcomes of the proposed predictive model can be further extracted further to justify the software's effectiveness [38].

In terms of values for showing high risk of T2DM complications and risk factors (at a higher clinical level) which characterises the training results, the probabilities determined by the Bayesian Statistics is discretised and binariased for the risk factors and complications, respectively. Assessment of performance is based on the sensitivity and specificity. In particular, the model is tested using the accuracy of prediction as a percentage of the correct prognosis of the specific comorbidities. Whereas, prognostications were made of the true positive (TP), actual negative (TN), false positives (FP) and false negatives (FN), models were assessed throughout this thesis.

Data accessibility: Due to the data protection policy data is not publicly available. The final models presented can be embedded into the MOSAIC instrument, and the outcomes extracted further justify the software's effectiveness [22]. The findings extracted were encouraging and indicated that this model can be included in the MOSAIC instrument to provide the risk of complication. The results showed that an early prediction of T2DM complications

confirmed that the proposed latent model can be employed as an indicator even outside of the FSM for problematic circumstances. Given the attributes of the dataset, there is a potential for developing the model for predicting the development of a disorder, while comparing its performance with that obtained by applying the well-established and commonly used UKPDS risk engine.

Appendix A

A.1 Extra Information

This Appendix is designed as a supplementary materials for this thesis. It first intends to explain and clarify the concept of Link Strength measure, which was discussed previously in Chapter 5, especially Section 5.2. Then, it discusses the Association Rule Mining, quality metrics and pattern mining approaches used in Chapter 6 in more details.

Link Strength methodology

To find optimum number of hidden variable, the enhanced stepwise IC*LS (explained previously in Section 5.2 employed local and global sensitivity analysis [67] that consisted of Mutual Information (MI) and Link Strength (LS). The Link Strength methodology finds a structure for locating latent variables within a Bayesian structure. We exploited the following measures to handle the uncertainty in the discovered model:

- Entropy introduced in [109] to measure the uncertainty in a single node and shown below:

$$U(X) = \sum_{x_i} P(x_i) \log_2 \frac{1}{P(x_i)}. \tag{1}$$

- Mutual Information is a way of inferring links in data and measuring the connection strength [109] [97]. The MI between node X and Y is uncertainties in Y that is decreased

by knowing the state of X (or vice verse):

$$MI(X,Y) = U(Y) - U(Y|X), \tag{2}$$

where $U(Y|X)$ is calculated by averaging $U(Y|x_i)$ over all possible states x_i of X, taking $P(x_i)$ into account in which:

$$MI(X,Y) = \sum_{x,y} P(x,y) \log_2(\frac{P(x,y)}{P(x)P(y)}). \tag{3}$$

- The Link Strength [64] measure enables us to observe the impact of each discovered edge. Moreover, the percentage points of uncertainty reduction in Y are utilised by knowing the state of X if the states of all other parent variables are known. There are two types of LS in measuring uncertainties: True Average Link Strength (LSTA), and Blind Average Link Strength (LSBA).

- The LSTA calculates LS based on the average over the parent states using their actual joint probability. For a node with only one parent, MI Percentage and the LSTA Percentage yields the same value. LSTA of the edge $X \rightarrow Y$ is defined as the MI of (X,Y) conditioned on all other parents of Y, which shown as:

$$LSTA(X \rightarrow Y) \text{ ,X requires P(For all parents of Y)} \tag{4}$$

$$= MI(X,Y|Z) = U(Y|Z) - U(Y|X,Z) \tag{5}$$

$U(Y|X,Z)$ is the average over the states of all parents, and $U(Y|Z)$ is the average over all other parents.

- The LSBA is derived from the LSTA but ignores the actual frequency of occurrence of the parent states. Thus, in the LSBA measure, all parents are assumed to be independent

of each other and uniformly distributed.

$$LSBA(X \rightarrow Y) = \text{requires no inference at all.} \tag{6}$$

The same probabilities as the corresponding absolute measure above are converted to each percentage measure. For removing all uncertainty, we require deterministic functions, in which the state of a child is completely known if the states of all of its parents are known. Representing all parents from Y in $MI(X, Y|Z)$ in Equation(5) essentially blocks all information flow through the other parents, Z. According to [45], we are confident that there are no other indirect open links between Y and X through descendants of Y, once all different parents are instantiated then there is a direct link from X to Y.

Theorem: Consider a BN (G, P) consisting of a DAG (G) and a joint probability (P). Let $X \rightarrow Y$ be an edge in G and denote the set of all other parents of Y as Z. Let $G\%$ be the modified DAG generated by deleting edge $X \rightarrow Y$ in G. Then X and Y are conditionally independent given Z in BN $(G\%, P\%)$ for any joint probability $P\%$. As indicated by the LSTA, most links are quite strong, can be classified as significant, except for those with LSTA of less or equal to zero (removed from the final structure). The LS measure may be useful in the context of constraint-based structure learning algorithms to derive hypotheses of a system's primary causal pathways. In addition, it can be used to evaluate the quality of the structure learning algorithms. Currently, structure learning algorithms are evaluated by counting the number of incorrect arrows when identifying known systems. It may be more appropriate to weight those counts by the LS of the incorrect arrows. By setting the value of the LSTA greater than 20 percent, though, overfitting in the DAG can be reduced.

Quality Metrics

This Section provides an example to help understand the MCI algorithm and pattern mining methods introduced in Section 6.3.2. The support measure of itemsets $\mathbb{C}(\pi)_i * (supp(\mathbb{C}(\pi)_i)$ is defined as the proportion of transactions in the dataset containing $RHS(\mathbb{C}(\pi)_i)$. In particular,

an association rule of $\partial(\mathbb{C}(\pi)_i) \Rightarrow \partial(\mathbb{C}(\pi)_j)$ has a support of $P(\mathbb{C}(\pi)_i \mathbb{C}(\pi)_j)$. The confidence measure of a rule identifies the proportion of transactions with the most interesting/important relationships. In addition, the confidence of a rule is defined as confidence $(\mathbb{C}(\pi)_i \implies \mathbb{C}(\pi)_j \equiv support(\mathbb{C}(\pi)_i \cup \mathbb{C}(\pi)_j) \equiv support(\mathbb{C}(\pi)_j)$ which satisfies Equation 7.

$$support(\mathbb{C}(\pi)_i \cup \mathbb{C}(\pi)_j) > \sigma, confidence(\mathbb{C}(\pi)_i \Rightarrow \mathbb{C}(\pi)_j) > \delta, lift(\frac{P(\mathbb{C}(\pi)_i \ \cap \ \mathbb{C}(\pi)_j)}{P(\mathbb{C}(\pi)_i) \times P(\mathbb{C}(\pi)_j)}), \tag{7}$$

Parameters such as σ and δ are the minimum support and confidence, respectively. Instead of using accuracy, efficiency is an appropriate way to evaluate association rules [94]. To obtain the frequents itemsets, first TARs are filtered by using minimum support ($\sigma <= 0.001$) and confidence ($\delta <= 25\%$). However, they are not able to filter complications-rules based on the different dependencies among the rules. For this purpose, a measurement of independence of $\mathbb{C}(\pi)_i$ and $\mathbb{C}(\pi)_j$ which is known as lift. Lift is the deviation of the whole rule support from the expected support under independence given both sides of the rule support. Higher lift values indicate strong associations. Lift of 1 represents $\mathbb{C}(\pi)_i$ and $\mathbb{C}(\pi)_j$ are independent as shown in Equation 8.

$$lift(\mathbb{C}(\pi)_i \implies \mathbb{C}(\pi)_j) = support(\mathbb{C}(\pi)_i \cup \mathbb{C}(\pi)_j) = support(\mathbb{C}(\pi)_i) \times support(\mathbb{C}(\pi)_j) \tag{8}$$

For example, the probability of developing both HYP and LIV is associated with the likelihood of developing RET. Confidence of HYP, LIV implying RET is given as the likelihood of developing HYP, LIV and also RET over the likelihood of developing only HYP and LIV (see Equation 9).

$$confidence(\{HYP, LIV\} \implies \{RET\}) = \frac{support(\{HYP, LIV, RET\})}{support(\{HYP, LIV\})} \tag{9}$$

The confidence measures whether $\{RET, HYP, NEU, RET\}$ implies LIV. This reveals that how likely a given patient develops $\{RET, HYP\}, NEU, RET$ and LIV. In order to find the most interesting itemsets, support ensures that all sub-rules of the frequent itemsets are also

frequent, hence no superset of infrequent itemsets can be frequent. Confidence is very sensitive to the frequency of the consequent. It has been reported that consequents with higher support will produce higher confidence even though there is no association among the antecedent and consequent. Thus, it might not be useful in performing effectively with the existence of bias in dataset DS with a having small number of patients and relatively complications. Confidence measures the strength of the association rules in which the patients that have complication $\mathbb{C}(\pi)_i$ also developed $\mathbb{C}(\pi)_j$ together. There is a number of choices for selecting the filtering measures [55] such as lift, leverage and coverage. Where Lift($\mathbb{C}(\pi)_i \implies \mathbb{C}(\pi)_j$) = confidence($\mathbb{C}(\pi)_i \implies \mathbb{C}(\pi)_j$) $\times$ support(χ_j), leverage($\mathbb{C}(\pi)_i \implies \mathbb{C}(\pi)_j$) = support($\mathbb{C}(\pi)_i \implies \mathbb{C}(\pi)_j$) - (support($\chi_i$) $\times$ support(χ_j)), coverage($\mathbb{C}(\pi)_i \implies \mathbb{C}(\pi)_j$) = support($\mathbb{C}(\pi)_i$). In T2DM dataset, there is a strong association (indicated by the highest lift) among the complications, which shows the likelihood of the complication being developed relative to its general developing rate, given that the patient developed other complications. For instance, the conditional probability of developing both HYP and LIV in are associated with the likelihood of the patient developing RET. There is a strong association (indicated by the highest lift) among the complications, which shows the likelihood of the complication being developed relative to its general developing rate, given that the patient developed other complications. For example, the conditional probability of a patient developing both HYP and LIV is associated with the likelihood of the patient developing RET. Whereas Coverage filters the rules mostly based on their antecedents. This opposite the present paper preferences where the consequents (the complications occur in the future visits) have been considered as the most revealing itemsets in the decision making and prediction process. Similar to lift, conviction metric assesses the likelihood of the appearance of an antecedent in which the corresponding consequent is not likely to occur.

Overall, a question still remains to answer whether it could be possible to trust these metrics by the user-defined thresholds. In particular, there are many challenges to find the most interesting rules [42] only by relying on TARs. Nevertheless, most of the previously mentioned metrics in this study are mainly depended on the support and frequency. In a small-sized dataset like DS, where there is a different imbalance ratio for each item (complication), bias,

and latent factors, it may not be beneficial if is only trust on the obtained itemsets resulted by using support, confidence, and lift.

Moreover, there are some itemsets which are called frequent itemsets while their occurrence exceeds the threshold in the database. In order to generate interesting rules, one could come across many frequent itemsets with minimal confidence. In the other words, by applying a rigid constraint with having bias in data, the final itemsets can be identified as interesting itemsets wrongly. This is because interestingness is only based on the association of HYP with the items, not the relationships among the items themselves. An item like HYP with a high occurrence rate can affect the way how other items are associated with each other. To avoid the above issue in a small-sized dataset, this thesis tended to discover all types of associations regardless of effect of bias (e.g., HYP) and focus mostly on the relaxed or flexible filtering metrics.

It does not seem to be possible to only rely on lift as it may not be trustworthy enough and unable to perform effectively with the existence of bias in the incomplete data. Lift suffers from having non-fixed range of variables. It only assesses the dependency and correlation of the items without taking into consideration the importance of the cause and effect relationships among antecedents and consequents. Similar to the issue related to support and confident, lift is susceptible to infrequent items with a relatively low probability complications-rules that can be ranked wrongly as the most interesting itemsets. Although having a very low or minimal constraints to be applied on the quality metrics, it does not eliminate the above issue which is caused by generating all possible permutations of complications for all transactions as an non-optimal option. This is because, Tables 6.1 contains many different antecedents and consequents which increase the database size exponentially based on the number of items. It also leads to generating large number of uninteresting distances among many small rules despite the previously chosen optimal/minimal threshold for support and confidence. In this situation, neither clustering nor ARM methodology perform effectively and can be even worse and problematic in a sparse dataset (such as T2DM). In conclusion, for making a better decision, the uninteresting rules needs to be reduced at another level which is addressed by using pattern clustering.

Table A.1: Database r_1 of the associated rules with the complications generated using TARs.

Rule	Antecedent		Consequent	Itemsets/Objects in D	Support	Confidence	Lift
7	{ }	$\Longrightarrow$	{NEP}	2,11,13,14,16,17,18,20,25 ,29,30,32,35,36,37,38,41	0.11	0.11	1.00
8	{ }	$\Longrightarrow$	{NEU}	5,7,13,15,16,19,21,25,26,28 ,29,31,34,35,37,38,39,40,41	0.16	0.16	1.00
9	{ }	$\Longrightarrow$	{RET}	3,4,8,14,15,17,22,23,25,27,28 ,32,33,34,38,41	0.15	0.15	1.00
10	{ }	$\Longrightarrow$	{LIV}	6,12,18,19,22,24,29,30,31,32 ,33,34,35,36,37,39,40	0.15	0.15	1.00
11	{ }	$\Longrightarrow$	{HYP}	2,3,4,5,6,10,13,14,20,21,23 ,24,26,27,28,30,31,33,38,41	0.86	0.86	1.00
12	{NEU|HYP}	$\Longrightarrow$	{NEU}	13,26,38,41	0.01	0.27	1.71
39	{HYP}	$\Longrightarrow$	{LIV}	6,24,30,31,33,35,36,37,39,40	0.14	0.16	1.06
40	{{NEP|HYP}|NEU}	$\Longrightarrow$	{RET}	28,38,41	0.01	1.00	6.57
42	{NEU|RET}	$\Longrightarrow$	{NEP|HYP}	28,38,41	0.01	0.19	6.84
53	{LIV|NEP}	$\Longrightarrow$	{RET}	32,36	≥ 0.001	0.06	0.41
62	{LIV|NEU}	$\Longrightarrow$	{RET}	34,39,40	0.01	0.29	1.88
80	{HYP|LIV|NEU}	$\Longrightarrow$	{NEP}	37	0.01	0.33	3.11

Table A.2: A subset of Database r_2 of the associated rules with the complications generated using TARs.

Rule	Antecedent		Consequent	Itemsets/Objects in D	Support	Confidence	Lift
10	{ }	$\Longrightarrow$	{LIV}	6,12,18,19,22,24,29,30,31,32 ,33,34,35,36,37,39,40	0.15	0.15	1.00
11	{ }	$\Longrightarrow$	{HYP}	2,3,4,5,6,10,13,14,20,21,23 ,24,26,27,28,30,31,33,38,41	0.86	0.86	1.00
16	{ }{ }	$\Longrightarrow$	{RET}	3,4,8,14,15,17,22,23,25,27 ,28,32,33,34,38,41	0.01	0.22	1.46
18	{ }{ }	$\Longrightarrow$	{HYP}	2,3,4,5,6,10,13,14,20,21,23 ,24,26,27,28,30,31,33,38,41	0.02	0.78	0.90
20	{NEP}	$\Longrightarrow$	{NEU}	13,16,25,26,29,38,41	0.02	0.19	1.17
25	{LIV}	$\Longrightarrow$	{NEP}	18,29,30,32,36,37	0.04	0.27	2.49
39	{HYP}	$\Longrightarrow$	{LIV}	6,24,30,31,33,35,36,37,39,40	0.14	0.16	1.06
40	{{NEP|HYP}|NEU}	$\Longrightarrow$	{RET}	28,38,41	0.01	1.00	6.57
42	{NEU|RET}	$\Longrightarrow$	{NEP,HYP}	28,38,41	0.01	0.19	6.84
48	{LIV|NEU}	$\Longrightarrow$	{NEP}	29,35,37	0.01	0.29	2.66
60	{HYP|LIV}	$\Longrightarrow$	{NEP}	30,35,36,37	0.04	0.27	2.54
69	{HYP|LIV}	$\Longrightarrow$	{NEU}	31,35,37,39,40	0.02	0.11	0.68
80	{HYP|LIV|NEU}	$\Longrightarrow$	{NEP}	37	0.01	0.33	3.11

Table A.3: The power set (MCI) obtained based on the MCI algorithm of the most interesting rules in MCI representing two subsets (r_1 and r_2) of the intersected associated rules with the complications.

Rule	LHS		RHS	Objects	Support	Confidence
7	{ }	$\Longrightarrow$	{NEP}	2,11,13,14,16,17,18,20,25 ,29,30,32,35,36,37,38,41	0.11	0.11
8	{ }	$\Longrightarrow$	{NEU}	5,7,13,15,16,19,21,25,26,28 ,29,31,34,35,37,38,39,40,41	0.16	0.16
9	{ }	$\Longrightarrow$	{RET}	3,4,8,14,15,17,22,23,25,27,28 ,32,33,34,38,41	0.15	0.15
10	{ }	$\Longrightarrow$	{LIV}	6,12,18,19,22,24,29,30,31,32 ,33,34,35,36,37,39,40	0.15	0.15
11	{ }	$\Longrightarrow$	{HYP}	2,3,4,5,6,10,13,14,20,21,23 ,24,26,27,28,30,31,33,38,41	0.86	0.86
12	{{NEU|HYP}}	$\Longrightarrow$	{NEU}	13,26,38,41	0.01	0.27
11	{ }	$\Longrightarrow$	{HYP}	2,3,4,5,6,10,13,14,20,21,23 ,24,26,27,28,30,31,33,38,41	0.86	0.86
42	{NEU|RET}	$\Longrightarrow$	{NEP,HYP}	28,38,41	0.01	0.19
60	{HYP|LIV}	$\Longrightarrow$	{NEP}	30,35,36,37	0.04	0.27
62	{LIV|NEU}	$\Longrightarrow$	{RET}	34,39,40	0.01	0.29
80	{HYP|LIV|NEU}	$\Longrightarrow$	{NEP}	37	0.01	0.33

Minimal Coverage Itemsets Algorithm

The MCI algorithm in Section A.1 ascertained whether patients in each cluster were developing a similar pattern of complications. This was used to discover the most frequent itemsets which were also unique for the corresponding cluster as well as a different pattern from other patients within another cluster. For instance, once HYP have happened before the appearance of LIV, RET and (NEU or NEP or no complication), there could be a co-occurrence pattern of $\{HYP, \{LIV, RET\}, \{NEU|NEP\}\}$. To illustrate this, two subsets of rules were selected with maximum lift and reasonable support and confidence (meeting the constraints defined in Section) as was shown in Tables A.1,A.2:

$$r_1 = \{R_7, R_8, R_9, R_{10}, R_{12}, R_{27}, R_{39}, R_{40}, R_{42}, R_{53}, R_{62}, R_{80}\},$$

$$r_2 = \{R_{10}, R_{11}, R_{16}, R_{18}, R_{20}, R_{25}, R_{39}, R_{40}, R_{42}, R_{48}, R_{60}, R_{69}, R_{80}\}$$

The union of objects in these subsets was met the most items in D. It is then aimed to find out whether the rules set have covered the optimal/minimal number of the associated Objects. This illustrated an ideal itemsets MCI of the intersection of r_1 and r_2 which is defined as MCI $= \{R_{10}, R_{11}, R_{42}, R_{60}, R_{62}\}$ (as was shown in Table A.3. These itemsets are generated based upon the intersection of objects in MCI representing a unique/minimum coverage set of Items in D and seen in Figure 6.2.

A.2 Research Ethics approval and consent to participate

All participants of focus groups consented to study participation. The studies described in the evaluation activities were approved by the biomedical research ethics committee from the Ethics Committee at Istituti Clinico Scientifici Maugeri.

Glossary

Accuracy An analysis results from a classification model generated by splitting the labelled data and comparing the predicted class membership generated by the classification model with the actual class membership provided by the observational data; sum of true positive and true negative cases over all population.

AUC Area Under receiver operating characteristic Curve

BMI Body Mass Index

Confusion matrix A table generated to show true labels derived from the input data that are compared to the predicted labels from a classifier

classification a supervised machine learning approach to assign observation into different pre-defined classes.

clustering an unsupervised machine learning technique to assign unknown observational samples into categories.

COL Cholesterol

complications-rules rules between complications

CRT Creatinine

DAG Directed Acyclic Graph

DBN Dynamic Bayesian Network

DBP Diastolic Blood Pressure

EM Expectation-Maximization

H Hidden / latent variable

HBA Glycated Haemoglobin/H2A1c

HDL High-Density Lipoprotein

HMM Hidden Markov Model

HYP Hypertension

IC Induction Causation for observed variables

IC* Induction Causation for observed variables and latent variables

Jaccard Jaccard Index

LS Link Strength

LHS Left Hand Side

LSTA True Average Link Strength

LSBA Blind Average Link Strength

MAP Maximum A posteriori Probability

MCI Minimum Coverage Itemsets

MCMC Markov Chain Monte Carlo

NEP Nephropathy

NEU Neuropathy

NHS National Health Service

PLS Partial Least Squares

REVEAL Reverse Engineering Algorithm

REP Retinopathy

RHS Right Hand Side

ROC Receiver Operating Characteristic curve

SBP Systolic Blood Pressure

SD Standard Deviation

SMK Smoking habit

SSM State Space Model

T2DM Type 2 Diabetes Mellitus

TAR Temporal Association Rules

TS Time-Series

TS Bootstrapping Time-Series Bootstrapping approach

WHO World Health Organisation

Bibliography

[1] A. I. Adler, I. M. Stratton, H. A. W. Neil, J. S. Yudkin, D. R. Matthews, C. A. Cull, A. D. Wright, R. C. Turner, and R. R. Holman. Association of systolic blood pressure with macrovascular and microvascular complications of type 2 diabetes (ukpds 36): prospective observational study. *Bmj*, 321(7258):412–419, 2000.

[2] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah. Time-series clustering–a decade review. *Information Systems*, 53:16–38, 2015.

[3] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[4] J. M. Ale and G. H. Rossi. An approach to discovering temporal association rules. In *Proceedings of the 2000 ACM symposium on Applied computing-Volume 1*, pages 294–300, 2000.

[5] A. Z. Ali, M. Hossain, R. Pugh, et al. Diabetes, obesity and hypertension in urban and rural people of bedouin origin in the united arab emirates. *The Journal of tropical medicine and hygiene*, 98(6):407–415, 1995.

[6] J. F. Allen et al. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154, 1984.

[7] F. Altiparmak, H. Ferhatosmanoglu, S. Erdal, and D. C. Trost. Information mining over heterogeneous and high-dimensional time-series data in clinical trials databases. *IEEE Transactions on Information Technology in Biomedicine*, 10(2):254–263, 2006.

[8] H. Amirkhani, M. Rahmati, P. J. Lucas, and A. Hommersom. Exploiting experts' knowledge for structure learning of bayesian networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2154–2170, 2017.

[9] L. Bellamy, J.-P. Casas, A. D. Hingorani, and D. Williams. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *The Lancet*, 373(9677):1773–1779, 2009.

[10] R. Bellazzi. Big data and biomedical informatics: a challenging opportunity. *Yearbook of medical informatics*, 9(1):8, 2014.

[11] R. Bellazzi, A. Dagliati, L. Sacchi, and D. Segagni. Big data technologies: new opportunities for diabetes management. *Journal of diabetes science and technology*, 9(5):1119–1125, 2015.

[12] R. Bellazzi, F. Ferrazzi, and L. Sacchi. Predictive data mining in clinical medicine: a focus on selected methods and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):416–430, 2011.

[13] R. Bellazzi, L. Sacchi, and S. Concaro. Methods and tools for mining multivariate temporal data in clinical and biomedical applications. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5629–5632. IEEE, 2009.

[14] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97, 2008.

[15] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[16] M. R. Berthold, C. Borgelt, F. Höppner, and F. Klawonn. *Guide to intelligent data analysis: how to intelligently make sense of real data.* Springer Science & Business Media, 2010.

[17] J. A. Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

[18] R. K. Blashfield. Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, 83(3):377, 1976.

[19] X. Boyen, N. Friedman, and D. Koller. Discovering the hidden structure of complex dynamic systems. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 91–100. Morgan Kaufmann Publishers Inc., 1999.

[20] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 33–42. Morgan Kaufmann Publishers Inc., 1998.

[21] W. L. Buntine. Bayesian back-propagation. *Complex systems*, 5(6):603–643, 1991.

[22] J. Cancela, G. Fico, M. T. Arredondo, A. G. Paton, and A. Guillen. Mosaic: Models and simulation techniques for discovering diabetes influence factors. *Transactions of Japanese Society for Medical and Biological Engineering*, 51(Supplement):R–162, 2013.

[23] O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Science & Business Media, 2006.

[24] S. Ceccon, D. Garway-Heath, D. Crabb, and A. Tucker. The dynamic stage bayesian network: identifying and modelling key stages in a temporal process. *Advances in Intelligent Data Analysis X*, pages 101–112, 2011.

[25] S. Ceccon, D. F. Garway-Heath, D. P. Crabb, and A. Tucker. Exploring early glaucoma and the visual field test: Classification and clustering using bayesian networks. *IEEE journal of biomedical and health informatics*, 18(3):1008–1014, 2014.

[26] D. Chicharro and S. Panzeri. Algorithms of causal inference for the analysis of effective connectivity among brain regions. *Frontiers in neuroinformatics*, 8:64, 2014.

[27] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

[28] I.-G. Choi, J. Kwon, and S.-H. Kim. Local feature frequency profile: a method to measure structural similarity in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3797–3802, 2004.

[29] L. co Todorovski, B. Cestnik, and M. Kline. Qualitative clustering of short time-series: A case study of firms reputation data. *IDDM-2002*, 141, 2002.

[30] S. Colagiuri. Glycated haemoglobin (hba1c) for the diagnosis of diabetes mellitus–practical implications. *Diabetes research and clinical practice*, 93(3):312, 2011.

[31] E. R. F. Collaboration et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *The Lancet*, 375(9733):2215–2222, 2010.

[32] E. A. Colosimo, M. A. Fausto, M. A. Freitas, and J. A. Pinto. Practical modeling strategies for unbalanced longitudinal data analysis. *Journal of Applied Statistics*, 39(9):2005–2013, 2012.

[33] S. Concaro, L. Sacchi, C. Cerra, P. Fratino, and R. Bellazzi. Mining health care administrative data with temporal association rules on hybrid events. *Methods of information in medicine*, 50(02):166–179, 2011.

[34] G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[35] M. Cusick, A. D. Meleth, E. Agron, M. R. Fisher, G. F. Reed, G. L. Knatterud, F. B. Barton, M. D. Davis, F. L. Ferris, E. Y. Chew, et al. Associations of mortality and diabetes complications in patients with type 1 and type 2 diabetes: early treatment diabetic retinopathy study report no. 27. *Diabetes Care*, 28(3):617–625, 2005.

[36] A. Dagliati, A. Malovini, P. Decata, G. Cogni, M. Teliti, L. Sacchi, C. Cerra, L. Chiovato, and R. Bellazzi. Hierarchical bayesian logistic regression to forecast metabolic control in type 2 dm patients. In *AMIA Annual Symposium Proceedings*, volume 2016, page 470. American Medical Informatics Association, 2016.

[37] A. Dagliati, A. Marinoni, C. Cerra, P. Decata, L. Chiovato, P. Gamba, and R. Bellazzi. Integration of administrative, clinical, and environmental data to support the management of type 2 diabetes mellitus: From satellites to clinical care. *Journal of diabetes science and technology*, 10(1):19–26, 2016.

[38] A. Dagliati, L. Sacchi, M. Bucalo, D. Segagni, K. Zarkogianni, A. M. Millana, J. Cancela, F. Sambo, G. Fico, M. T. M. Barreira, et al. A data gathering framework to collect type 2 diabetes patients data. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 244–247. IEEE, 2014.

[39] R. Daly, Q. Shen, and S. Aitken. Learning bayesian networks: approaches and issues. *The knowledge engineering review*, 26(2):99–157, 2011.

[40] G. Das, K.-I. Lin, H. Mannila, G. Renganathan, and P. Smyth. Rule discovery from time series. In *KDD*, volume 98, pages 16–22, 1998.

[41] U. Diabetes. Diabetes: facts and stats. *Diabetes UK*, 3:1–21, 2014.

[42] Y. Djenouri, Y. Gheraibia, M. Mehdi, A. Bendjoudi, and N. Nouali-Taboudjemat. An efficient measure for evaluating association rules. In *2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pages 406–410. IEEE, 2014.

[43] A. Doddi, S. Marathe, D. C. Ravi, and S. Torney. Discovery of association rules in medical data. *Medical informatics and the Internet in medicine*, 26(1):25–33, 2001.

[44] P. J. Dyck, K. Kratz, J. Karnes, W. J. Litchy, R. Klein, J. Pach, D. Wilson, P. O'brien, and L. Melton. The prevalence by staged severity of various types of diabetic neuropathy, retinopathy, and nephropathy in a population-based cohort: the rochester diabetic neuropathy study. *Neurology*, 43(4):817–817, 1993.

[45] I. Ebert-Uphoff. Measuring connection strengths and link strengths in discrete bayesian networks. Technical report, Georgia Institute of Technology, 2007.

[46] G. Elidan, N. Lotner, N. Friedman, and D. Koller. Discovering hidden variables: A structure-based approach. In *Advances in Neural Information Processing Systems*, pages 479–485, 2001.

[47] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.

[48] M. J. Fowler. Microvascular and macrovascular complications of diabetes. *Clinical diabetes*, 26(2):77–82, 2008.

[49] N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 139–147. Morgan Kaufmann Publishers Inc., 1998.

[50] T. F. Gharib, H. Nassar, M. Taha, and A. Abraham. An efficient algorithm for incremental mining of temporal association rules. *Data & Knowledge Engineering*, 69(8):800–815, 2010.

[51] V. Granville, M. Krivanek, and J.-P. Rasson. Simulated annealing: A proof of convergence. *IEEE transactions on pattern analysis and machine intelligence*, 16(6):652–656, 1994.

[52] U. P. D. S. Group et al. Uk prospective diabetes study 16: overview of 6 years' therapy of type ii diabetes: a progressive disease. *Diabetes*, 44(11):1249–1258, 1995.

[53] M. Grzegorczyk and D. Husmeier. Non-stationary continuous dynamic bayesian networks. In *Advances in Neural Information Processing Systems*, pages 682–690, 2009.

[54] Y. Guo, G. Bai, and Y. Hu. Using bayes network for prediction of type-2 diabetes. In *2012 International Conference for Internet Technology and Secured Transactions*, pages 471–472. IEEE, 2012.

[55] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[56] D. J. Hand. Intelligent data analysis: Issues and opportunities. In *International Symposium on Intelligent Data Analysis*, pages 1–14. Springer, 1997.

[57] M. F. Harris and N. A. Zwar. Care of patients with chronic disease: the challenge for general practice. *Medical Journal of Australia*, 187(2):104–107, 2007.

[58] G. T. Herman. Application of maximum entropy and bayesian optimization methods to image reconstruction from projections. In *Maximum-Entropy and Bayesian Methods in Inverse Problems*, pages 319–338. Springer, 1985.

[59] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[60] H. Hurwitz Jr. Entropy reduction in bayesian analysis of measurements. *Physical Review A*, 12(2):698, 1975.

[61] S. E. Inzucchi and R. S. Sherwin. The prevention of type 2 diabetes mellitus. *Endocrinology and Metabolism Clinics*, 34(1):199–219, 2005.

[62] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[63] H. Jen-Wei and C. M. S. Dai Bi-Ru. Twain: Two-end association miner with precise frequent exhibition periods. *ACM Transactions on Knowledge Discovery from Data*, 8(2):800–815, 2007.

[64] N. Jitnah. *Using Mutual Information for Approximate Evalutation of Bayesian Networks*. Monash University, 1999.

[65] E. Keogh and J. Lin. Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and information systems*, 8(2):154–177, 2005.

[66] M. Khashei, M. Bijari, and G. A. R. Ardali. Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (anns). *Neurocomputing*, 72(4-6):956–967, 2009.

[67] J. Khoo, T.-L. Tay, J.-P. Foo, E. Tan, S.-B. Soh, R. Chen, V. Au, B. J.-M. Ng, and L.-W. Cho. Sensitivity of a1c to diagnose diabetes is decreased in high-risk older southeast asians. *Journal of Diabetes and its Complications*, 26(2):99–101, 2012.

[68] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677, 2018.

[69] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138, 2019.

[70] C.-H. Lee, M.-S. Chen, and C.-R. Lin. Progressive partition miner: an efficient algorithm for mining general temporal association rules. *IEEE Transactions on Knowledge and Data Engineering*, (4):1004–1017, 2003.

[71] Y. Li, S. Swift, and A. Tucker. Modelling and analysing the dynamics of disease progression from cross-sectional studies. *Journal of biomedical informatics*, 46(2):266–274, 2013.

[72] G. Liang. An effective method for imbalanced time series classification: Hybrid sampling. In *Australasian Joint Conference on Artificial Intelligence*, pages 374–385. Springer, 2013.

[73] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. 1998.

[74] L. Litwak, S.-Y. Goh, Z. Hussein, R. Malek, V. Prusty, and M. E. Khamseh. Prevalence of diabetes complications in people with type 2 diabetes mellitus and its association with baseline characteristics in the multinational a 1 chieve study. *Diabetology and metabolic syndrome*, 5(1):57, 2013.

[75] C. Liu, W. Ding, Y. Hu, X. Xia, B. Zhang, J. Liu, and D. Doermann. Circulant binary convolutional networks for object recognition. *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[76] G. Liu, S. Huang, C. Lu, and Y. Du. An improved k-means algorithm based on association rules. *International Journal of Computer Theory and Engineering*, 6(2):146, 2014.

[77] A. Lloyd, W. Sawyer, and P. Hopkinson. Impact of long-term complications on quality of life in patients with type 2 diabetes not using insulin. *Value in Health*, 4(5):392–400, 2001.

[78] P. J. Lucas, L. C. Van der Gaag, and A. Abu-Hanna. Bayesian networks in biomedicine and health-care, 2004.

[79] S. Mani, Y. Chen, T. Elasy, W. Clayton, and J. Denny. Type 2 diabetes risk forecasting from emr data using machine learning. In *AMIA annual symposium proceedings*, volume 2012, page 606. American Medical Informatics Association, 2012.

[80] S. Mani and G. F. Cooper. Causal discovery using a bayesian local causal discovery algorithm. In *Medinfo*, pages 731–735, 2004.

[81] S. Marini, E. Trifoglio, N. Barbarini, F. Sambo, B. Di Camillo, A. Malovini, M. Manfrini, C. Cobelli, and R. Bellazzi. A dynamic bayesian network model for long-term simulation of clinical complications in type 1 diabetes. *Journal of biomedical informatics*, 57:369–376, 2015.

[82] J. Martin and K. VanLehn. Discrete factor analysis: Learning hidden variables in bayesian networks. Technical report, Technical report, Department of Computer Science, University of Pittsburgh, 1995.

[83] J. Mennis and J. W. Liu. Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Transactions in GIS*, 9(1):5–17, 2005.

[84] A. I. Messinger, G. Luo, and R. R. Deterding. The doctor will see you now: How machine learning and artificial intelligence can extend our understanding and treatment of asthma. *Journal of Allergy and Clinical Immunology*, 145(2):476–478, 2020.

[85] N. Moniz, P. Branco, and L. Torgo. Resampling strategies for imbalanced time series. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 282–291. IEEE, 2016.

[86] R. Moskovitch and Y. Shahar. Medical temporal-knowledge discovery via temporal abstraction. In *AMIA annual symposium proceedings*, volume 2009, page 452. American Medical Informatics Association, 2009.

[87] E. Mueller, S. Maxion-Bergemann, D. Gultyaev, S. Walzer, N. Freemantle, C. Mathieu, B. Bolinder, R. Gerber, M. Kvasz, and R. Bergemann. Development and validation of the economic assessment of glycemic control and long-term effects of diabetes (eagle) model. *Diabetes technology and therapeutics*, 8(2):219–236, 2006.

[88] K. R. Munana. Long-term complications of diabetes mellitus, part i: Retinopathy, nephropathy, neuropathy. *Veterinary Clinics: Small Animal Practice*, 25(3):715–730, 1995.

[89] K. P. Murphy and S. Russell. Dynamic bayesian networks: representation, inference and learning. 2002.

[90] R. M. Neal. Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems 5,[NIPS Conference]*, pages 475–482. Morgan Kaufmann Publishers Inc., 1992.

[91] R. M. Neal. Connectionist learning of belief networks. *Artificial intelligence*, 56(1):71–113, 1992.

[92] D. of Health. Five year forward view, 2014.

[93] C. Ordonez, C. A. Santana, and L. De Braal. Discovering interesting association rules in medical data. In *ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, pages 78–85. Citeseer, 2000.

[94] A. Parvez, S. Qamar, and S. Q. A. Rizvi. Techniques of data mining in healthcare: a review. *International Journal of Computer Applications*, 120(15), 2015.

[95] J. Pearl. Probabilistic reasoning in intelligent systems. 1988. *San Mateo, CA: Kaufmann*, 23:33–34.

[96] J. Pearl. *Causality*. Cambridge university press, 2009.

[97] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[98] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini. Meaningful explanations of black box ai decision systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9780–9784, 2019.

[99] L. Peelen, N. F. de Keizer, E. de Jonge, R.-J. Bosman, A. Abu-Hanna, and N. Peek. Using hierarchical dynamic bayesian networks to investigate dynamics of organ failure in patients in the intensive care unit. *Journal of biomedical informatics*, 43(2):273–286, 2010.

[100] M. Plasse, N. Niang, G. Saporta, A. Villeminot, and L. Leblond. Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Computational Statistics and Data Analysis*, 52(1):596–613, 2007.

[101] A. Ramachandran, C. Snehalatha, K. Satyavani, E. Latha, R. Sasikala, and V. Vijay. Prevalence of vascular complications and their risk factors in type 2 diabetes. *The Journal of the Association of Physicians of India*, 47(12):1152–1156, 1999.

[102] R. Raman, A. Gupta, S. Krishna, V. Kulothungan, and T. Sharma. Prevalence and risk factors for diabetic microvascular complications in newly diagnosed type ii diabetes

mellitus. sankara nethralaya diabetic retinopathy epidemiology and molecular genetic study (sn-dreams, report 27). *Journal of Diabetes and its Complications*, 26(2):123–128, 2012.

[103] F. Rijmen, E. H. Ip, S. Rapp, and E. G. Shaw. Qualitative longitudinal analysis of symptoms in patients with primary and metastatic brain tumours. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(3):739–753, 2008.

[104] J. W. Robinson and A. J. Hartemink. Learning non-stationary dynamic bayesian networks. *Journal of Machine Learning Research*, 11(Dec):3647–3680, 2010.

[105] S. Russell, J. Binder, D. Koller, and K. Kanazawa. Local learning in probabilistic networks with hidden variables. In *IJCAI*, volume 95, pages 1146–1152, 1995.

[106] L. Sacchi, C. Larizza, C. Combi, and R. Bellazzi. Data mining with temporal abstractions: learning rules from time series. *Data Mining and Knowledge Discovery*, 15(2):217–247, 2007.

[107] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, and D. C. M. Saade. A bayesian network decision model for supporting the diagnosis of dementia, alzheimer disease and mild cognitive impairment. *Computers in biology and medicine*, 51:140–158, 2014.

[108] M. B. Sesen, T. Kadir, R.-B. Alcantara, J. Fox, and M. Brady. Survival prediction and treatment recommendation with bayesian techniques in lung cancer. In *AMIA Annual Symposium Proceedings*, volume 2012, page 838. American Medical Informatics Association, 2012.

[109] C. E. Shannon, W. Weaver, and A. W. Burks. The mathematical theory of communication. 1951.

[110] R. Silva and Z. Ghahramani. The hidden life of latent variables: Bayesian learning with mixed graph models. *Journal of Machine Learning Research*, 10(Jun):1187–1238, 2009.

[111] L. SIMAR. An invitation to the bootstrap: Panacea for statistical inference? *Institut de Statistique, Universite Catholique de Louvain, Louvain*, 2008.

[112] G. Sparacino, A. Facchinetti, A. Maran, and C. Cobelli. Continuous glucose monitoring time series and hypo/hyperglycemia prevention: requirements, methods, open problems. *Current diabetes reviews*, 4(3):181–192, 2008.

[113] C. SPEARMAN. " general intelligence," objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

[114] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

[115] D. J. Stekhoven and P. Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[116] S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, and P. Kellam. Consensus clustering and functional interpretation of gene-expression data. *Genome biology*, 5(11):R94, 2004.

[117] G. J. Szekely and M. L. Rizzo. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of classification*, 22(2), 2005.

[118] M. Talih and N. Hengartner. Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):321–341, 2005.

[119] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

[120] G. Targher, L. Bertolini, S. Rodella, R. Tessari, L. Zenari, G. Lippi, and G. Arcaro. Nonalcoholic fatty liver disease is independently associated with an increased incidence of cardiovascular events in type 2 diabetic patients. *Diabetes care*, 30(8):2119–2121, 2007.

[121] 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. *Diabetes and Vascular Disease Research*, 15(5):424–432, 2018.

[122] P. Thuluvath and D. Triger. Autonomic neuropathy and chronic liver disease. *QJM: An International Journal of Medicine*, 72(2):737–747, 1989.

[123] N. Tishby, E. Levin, and S. A. Solla. Consistent inference of probabilities in layered networks: Predictions and generalization. In *International Joint Conference on Neural Networks*, volume 2, pages 403–409, 1989.

[124] K. G. Tolman, V. Fonseca, A. Dalpiaz, and M. H. Tan. Spectrum of liver disease in type 2 diabetes and management of patients with diabetes and liver disease. *Diabetes care*, 30(3):734–743, 2007.

[125] N. Trifonova, A. Kenny, D. Maxwell, D. Duplisea, J. Fernandes, and A. Tucker. Spatio-temporal bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. *Ecological Informatics*, 30:142–158, 2015.

[126] A. Tucker, X. Liu, and D. Garway-Heath. Spatial operators for evolving dynamic bayesian networks from spatio-temporal data. In *Genetic and Evolutionary Computation—GECCO 2003*, pages 205–205. Springer, 2003.

[127] A. Tucker, V. Vinciotti, X. Liu, and D. Garway-Heath. A spatio-temporal bayesian network classifier for understanding visual field deterioration. *Artificial Intelligence in Medicine*, 34(2):163 – 177, 2005.

[128] R. Turner, R. Holman, D. Matthews, S. Oakes, P. Bassett, I. Stratton, C. Cull, S. Manley, and V. Frighi. Uk prospective diabetes study (ukpds). viii. study design, progress and performance. *Diabetologia*, 34(12):877–890, 1991.

[129] R. Turner, H. Millns, H. Neil, I. Stratton, S. Manley, D. Matthews, and R. Holman. Risk factors for coronary artery disease in non-insulin dependent diabetes mellitus: United kingdom prospective diabetes study (ukpds: 23). *Bmj*, 316(7134):823–828, 1998.

[130] M. Van der Heijden, M. Velikova, and P. J. Lucas. Learning bayesian networks for clinical time series analysis. *Journal of biomedical informatics*, 48:94–105, 2014.

[131] M. A. Van Gerven, B. G. Taal, and P. J. Lucas. Dynamic bayesian networks as prognostic models for clinical patient management. *Journal of biomedical informatics*, 41(4):515–529, 2008.

[132] T. Wang and Q. Lin. Hybrid predictive model: When an interpretable model collaborates with a black-box model. *arXiv preprint arXiv:1905.04241*, 2019.

[133] W. Wang, J. Yang, and R. Muntz. Tar: Temporal association rules on evolving numerical attributes. In *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 283–292. IEEE, 2001.

[134] J. Wiener. Will ai make me a better doctor?(conference presentation). In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, page 113141I. International Society for Optics and Photonics, 2020.

[135] R. Williams, L. Van Gaal, and C. Lucioni. Assessing the impact of complications on the costs of type ii diabetes. *Diabetologia*, 45(1):S13–S17, 2002.

[136] F. Wittig. Learning bayesian networks with hidden variables for user modeling. In *UM99 User Modeling*, pages 343–344. Springer, 1999.

[137] Q. Yang and X. Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(04):597–604, 2006.

[138] K.-H. Yoon, J.-H. Lee, J.-W. Kim, J. H. Cho, Y.-H. Choi, S.-H. Ko, P. Zimmet, and H.-Y. Son. Epidemic obesity and type 2 diabetes in asia. *The Lancet*, 368(9548):1681–1688, 2006.

[139] L. Yousefi, M. Al-Luhaybi, L. Sacchi, L. Chiovato, and A. Tucker. Identifying latent variables in dynamic bayesian networks with bootstrapping applied to type 2 diabetes complication prediction. 2020.

[140] L. Yousefi, L. Saachi, R. Bellazzi, L. Chiovato, and A. Tucker. Predicting comorbidities using resampling and dynamic bayesian networks with latent variables. In *Computer-*

Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium on, pages 205–206. IEEE, 2017.

[141] L. Yousefi, S. Swift, M. Arzoky, L. Saachi, L. Chiovato, and A. Tucker. Opening the black box: Discovering and explaining hidden variables in type 2 diabetic patient modelling. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1040–1044. IEEE, 2018.

[142] L. Yousefi, S. Swift, M. Arzoky, L. Saachi, L. Chiovato, and A. Tucker. Opening the black box: Personalizing type 2 diabetes patients based on their latent phenotype and temporal associated complication rules. *Computational Intelligence*, 2020.

[143] L. Yousefi, S. Swift, M. Arzoky, L. Sacchi, L. Chiovato, and A. Tucker. Opening the black box: Exploring temporal pattern of type 2 diabetes complications in patient clustering using association rules and hidden variable discovery. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 198–203. IEEE, 2019.

[144] L. Yousefi, A. Tucker, M. Al-luhaybi, L. Saachi, R. Bellazzi, and L. Chiovato. Predicting disease complications using a stepwise hidden variable approach for learning dynamic bayesian networks. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 106–111. IEEE, 2018.

[145] A. E. Zambelli. A data-driven approach to estimating the number of clusters in hierarchical clustering. *F1000Research*, 5, 2016.

[146] X. Zhang, K. B. Korb, A. E. Nicholson, and S. Mascaro. Latent variable discovery using dependency patterns. *arXiv preprint arXiv:1607.06617*, 2016.

[147] Q. Zhao and S. S. Bhowmick. Association rule mining: A survey. *Nanyang Technological University, Singapore*, page 135, 2003.

[148] L. M. Zintgraf, T. S. Cohen, and M. Welling. A new method to visualize deep neural networks. *arXiv preprint arXiv:1603.02518*, 2016.