Identifying Latent Variables in Dynamic Bayesian Networks with Bootstrapping Applied to Type 2 Diabetes Complication Prediction

Leila Yousefi [a,*], Allan Tucker [a]
[a] *Department of Computer Science, Brunel University London, United Kingdom*
E-mails: Leila.yousefi2@Brunel.ac.uk, Allan.tucker@brunel.ac.uk

Abstract. Predicting complications associated with complex disease is a challenging task given imbalanced and highly correlated disease complications along with unmeasured or latent factors. To analyse the complications associated with complex disease, this article attempts to deal with complex imbalanced clinical data, whilst determining the influence of latent variables within causal networks generated from the observation. This work proposes appropriate Intelligent Data Analysis methods for building Dynamic Bayesian networks with latent variables, applied to small-sized clinical data (a case of Type 2 Diabetes complications). First, it adopts a Time Series Bootstrapping approach to re-sample the rare complication class with a replacement with respect to the dynamics of disease progression. Then, a combination of the Induction Causation algorithm and Link Strength metric (which is called IC*LS approach) is applied on the bootstrapped data for incrementally identifying latent variables. The most highlighted contribution of this paper gained insight into the disease progression by interpreting the latent states (with respect to the associated distributions of complications). An exploration of inference methods along with confidence interval assessed the influences of these latent variables. The obtained results demonstrated an improvement in the prediction performance.

Keywords: Latent Variable, Diabetes, Dynamic Bayesian Networks, Time Series Bootstrapping, Disease Prediction

1. Introduction

Diabetes UK reported that Type 2 Diabetes Mellitus (T2DM) is the most common form of diabetes, accounting for at least 90 per cent of all instances. The World Health Organisation (WHO) reported that in the next 10 years there will be about 550 million people suffering from this disease [1]. T2DM occurs because of impaired insulin secretion or opposition to insulin action or both, which is associated with severe long-term morbidities and large health maintenance costs to providers. In particular, the worsening level of these complications (e.g., retinopathy, neuropathy, nephropathy and hypertension is known as a significant cause of death [2]. The main motivation behind this work is to stratify patient groups by means of latent variables to discover how complications in diabetes interact. An Italian clinical data have been used to create the T2DM dataset in this study and build predictive model that combine knowledge associated with

[*] Corresponding author. E-mail: Leila.yousefi2@Brunel.ac.uk.

biomedical influences, including physiological, biological, epigenetic modification and behavioural inputs. The dataset consists of physical examinations such as cholesterol and blood pressure and laboratory data, including HbA1c measurements and lipid profile. Early prediction of T2DM complications while discovering the behaviour of associated aggressive risk factors can help to control the disease and improve a patients quality of life [3]. It also can enhance classification accuracy and boost user confidence in the classification models [4]. Nevertheless, prediction of these complications and mining complex clinical data are challenging tasks given the mixture of clinical test results (such as blood pressure, cholesterol level, etc), complication types (categorical, numerical, ordinal data types), unequal length of patient visits, highly correlated risk factors, presence of unmeasured factors, heterogeneity, biased data, etc. The non-stationary characteristic of clinical data collected as part of the monitoring of T2DM complications, creates a difficult context for effective forecasting [5]. This is because, in T2DM dataset with each visit, a patient has a unique profile of symptoms and complications, regardless of the phase of the disease.

Another challenge within complex clinical data, as Elidan and co-authors in [6] emphasised, is the importance of the presence of latent variables in clinical data. This variables are some unmeasured factors that clinicians fail to measure them and needed to be discovered at the early stage of diabetes. The latent variable discovery in causal structures has been introduced in [7]. Pearl in [8] utilise trees of hidden variables in order to render all observable variables independently. Similarly, Elidan and co-authors in [6] determined a hidden variable that interacts with observed variables within a Bayesian Network. In addition, they 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.

One amongst this research objectives is to use Intelligent Data Analysis (IDA) techniques to help explain the processes driving the complications of diabetes via the study of particular testimonials, time events as well as behavioural influences. Intelligent systems, whether biological or artificial, require the ability to make decisions under uncertainty using the available evidence. Several computational models in Artificial Intelligence (AI) and Machine Learning exhibit some of the required functionality to handle uncertainty. These models are judged by two main criteria: ease of creation and effectiveness in decision making. For example, Neural Networks (NNs) which represent complex input/output relations using combinations of simple nonlinear processing elements, are a familiar tool in AI and computational neuroscience. Having said that, these networks could not be suitable in clinical domain (such as the dataset used in this study), where each parameter was supported by little evidence, hence, the estimation of the parameters might not be not robust. As it is extremely important to learn a model from the clinical data with a small amount of training data with many parameters and few patients (samples). Alternatively, some other computational approaches has employed different methodology to reduce uncertainty in small-sized datasets. This study claims that Dynamic Bayesian Networks (DBNs) (which was introduced in [9]) can be incredibly beneficial in modelling T2DM dataset in which non-stationary risk factor dynamics are widespread and complicated. These networks can provide a topological description of the conditional independence relationships among variables. With these probabilistic graphical models, it is easy to interpret and provide information regarding the qualitative structure of the clinical domain. An appropriate non-stationary DBNs model can contribute to diabetes literature by discovering latent variables, potentially capture unmeasured effects from clinical data. This research is to incorporate these into existing predictive analytics frameworks to improve decision-making in patient practice. A key aim of this study is to bootstrap non-stationary data to improve our latent variable learning model due to the over representation of patients with specific comorbidities.

The rest of this article is organised as follows: The first half of the paper is dedicated to explaining how to balance time series clinical data and learn DBNs with and without latent variables. It begins with a descriptive data analysis while providing data definition and pre-processing methods. A set of models are learned from the data to evaluate the impact of adding latent variables and re-balancing the data via bootstrapping. It also involves techniques for analysing the strength of relationships between clinical and latent variables to better understand the meaning of the latent variables within the complex disease model and explores their effect. The second half of the paper is dedicated to analysing the results, in terms of classification (predicting two comorbidities associated with T2DM), validation and the potential for adoption in clinical practice. Finally, the findings are assessed by using a number of quantitative and qualitative validation strategies.

2. Related Works

There are various methodologies for T2DM prediction, e.g, risk-prediction equations and Markov models [10]. However, risk-prediction equations suffers from uncertainty as well as only performing one-step-ahead predictions. The Markov models are also limited to a small number of discrete risk factors. Among these, much of the existing literature on investigating the prognosis of T2DM complications, e.g., [11], focuses on logistic regression and Naïve Bayes methods. Moreover, most of the literature in T2DM prediction has often been restricted to modelling a limited number of visits. For example, Dagliati et al. in [12] presented a Hierarchical Bayesian Logistic Regression model to anticipate patients changes when the individual model parameters are estimated. For predicting comorbidities in [12], external and internal heterogeneity in T2DM patients were explored in cross-sectional data with just three horizons of time. Whilst, time series modelling and longitudinal study was not employed in the individual measurements. Similarly in [13], T2DM data was analysed to understand the influence of H2A1c and other risk factors in the development of the microvascular complications in 2-year time periods but did not model the data as a time series. In another work in [14], the authors failed to consider time series analysis, where a Bayes Network to predict diabetes was proposed on the Pima Indian Diabetes dataset. Marini in [15] simulated the health state and complications by using Bayesian inferred models, while applied to type 1 diabetes non-time series data.

Although extensive research has been carried out on the prediction of diabetic progression [11, 12, 15–17], no single study exists which has attempted to interpret the impact of correct number of latent variables in the presence of diabetic disorders. In a similar study [18] to the present paper, the authors provided a factor structure learning method that efficiently utilised hidden variables. 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 at the first place. Moreover, this method failed to consider prior belief in the factor structure and therefore, could not rely on the final structure. Unlike deep learning methods, DBNs can be invaluable where we do not want too much reliance on the training data which risks overfitting with poor generalisation capabilities. Many diseases involve structural changes based upon key stages in the progression, but many models do not appear to take this into account. Previous work on learning DBNs have inferred both network structures and parameters from (often incomplete) clinical data sets [9]. There has been some work in extending DBNs to model underlying processes that are non-stationary [19]. In [19], clinical features were modelled using a second order time series model but it was assumed that the temporal dependencies were time-invariant. Markov Chain Monte Carlo (MCMC) sampling algorithm and non-stationary DBN models in [20] were formalised for learning the structure of the model from time series biological data. Another work [21] 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. Similarly, Talih in [22] estimated the variance in the data structure parameter with a MCMC approach whilst the search space was limited to a fixed number of segments and indirect edges only, which is not suitable for T2DM data. Overall, such studies remained narrow 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.

T2DM dataset is highly imbalanced based on the disease common complications. To address the imbalance issue, so far many methods from weighting, generating new samples to one class classifiers have been proposed. Different learning techniques deal with imbalanced data, such as oversampling, undersampling, boosting, bagging, bootstrapping, and repeated random subsampling [23]. This work introduces a Bootstrapping approach which has been specifically designed for the longitudinal data to identify and re-balanced targeted complications. 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. TS Bootstrapping approach seems appropriate for T2DM dataset as the prediction in non-stationary models of data was difficult. Moreover, predicting rare cases in clinical data with an unbalanced distribution of a target complication is 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 [24]. A recent study, in [25], presented slightly similar re-balancing technique to this study but to analyse a different type of dataset (fisheries data) as well as different structure learning methodology.

This research expands the previous works to identify latent variables conducted following [26] on disease progression modelling with latent variables while proposes a bootstrap to both balance the data and calculate confidence bounds. Although, in [26], similar approaches to the current paper were employed in the context of the T2DM data, the imbalance issue was dealt using a DBNs model with only fixed single latent node. Moreover, the latent variable was leaned based on the standard approaches along with a basic bootstrapping technique, which was able to re-balance only one complication at time. Later, in [27, 28], an intuitive stepwise method, based upon the constraint based algorithm, was developed to learn the effects of multiple latent variables on the prediction performance. In addition to this, the obtained results were demonstrated on the diabetes dataset where re-sampled using the Pair-Sampling approach. The discovery of the optimum number of the latent variables was also challenging and often accuracy dropped as more latent variables were added due to overfitting. To address these issue, this article intends to incrementally identify the most influential latent variables. Therefore, it propose an enhanced variation on the stepwise method on bootstrapped data, which is called IC*LS approach.

3. Descriptive Data Analysis

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

3.1. Data Description and Data Collection

The data for this study is belonged to "MOdels and Simulation techniques for discovering diabetes Influence faCtors" (MOSAIC) project. Most of the information presented in the data section retrieved from MOSAIC website in [29], which were previously reported in [30, 31]. The MOSAIC project funds the data under the 7th Framework Program of the European Commission, Theme ICT2011.5.2 Virtual Physiological Human (600914) from 2009 to 2013. It consists of pre-diagnosed T2DM 1000 patients aged 25 to 65 years (inclusive) that were recruited from clinical followups at the "IRCCS Instituti Clinic Scientifici" (ICS) Maugeri of Pavia, Italy. This was extracted from external resources such as: the Hospital Fondazione Salvatore Maugeri (FSM), that mostly captured epidemiological records relevant to normal healthcare sector, and by the Local Healthcare Agency (Agenzia Sanitaria Locale, ASL), that accumulated measurements for institutional and technical transparency.

Since the definition of data varies among researchers, it is important to clarify how the final dataset was imputed throughout this research. T2DM data was chosen as a case study for this work as it suits the characteristic of complex clinical data (small-sized dataset with uneven number of visits per patient) after performing the centre profiling; a detailed analysis of the literature reported in [32]. For choosing the predictor variables to be used in the predictive model, this study mainly focused on the analysis of the diabetes literature mentioned earlier and found the variables which were usually related and accessible in the data with a significant risk of T2DM complications. Particularly, certain complications and risk factors (predictors) were selected based on existing literature on diabetes [33–37] and using recommendations from the clinicians at ICS. The selected T2DM complications were Retinopathy (RET), Hypertension (HYP), Nephropathy (NEP), Neuropathy (NEU) and Liver Disease (LIV). The predictors were identified and selected from the dataset, including: 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). For example, Fowler and co-authors in [38] analysed type 2 Diabetic American patients. They 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 to this, they considered LIV as a severe phenotype of diabetes and associated with T2DM complications (especially NEU) [39]. Litwak and co-authors analysed Russian diabetic patients in [40] referring to the influence of macro-vascular and micro-vascular disease on one anther. They showed that the most important features in T2DM dataset included blood pressure, HDL, lipid, BMI, and HbA1c influence diabetes complications. They also revealed that HDL has a negative effect on HYP, NEP, NEU, and RET, whereas HbA1c negatively associated with HYP. Another study conducted by Ramachandran [41] referred to the high prevalence of NEU and RET in Type 2 diabetes in India. Similar research in [35] 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. It also showed that there was a strong association among NEP, NEU, and RET.

Tables 1-2 represent the selected T2DM complications (comorbidities), risk factors and their clinical control values. It is necessary here to clarify exactly what is meant by Control Value and Discretised Value. 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). This work 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. Table 1 shows the binariased complications with two clinical level of High and Low. Alternatively, in Table 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 definitions [16, 17]. 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. Clinical risk factors are consists of three clinical level of risk, namely low (0), medium (1) and high (2). Node ID column, as seen in Table 2, 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. 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.

Table 1
The Description of T2DM Target Complications, Clinical Nodes, 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}

Table 2
The Description of the T2DM Clinical Features, Risk Factors, Control Values, and 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}

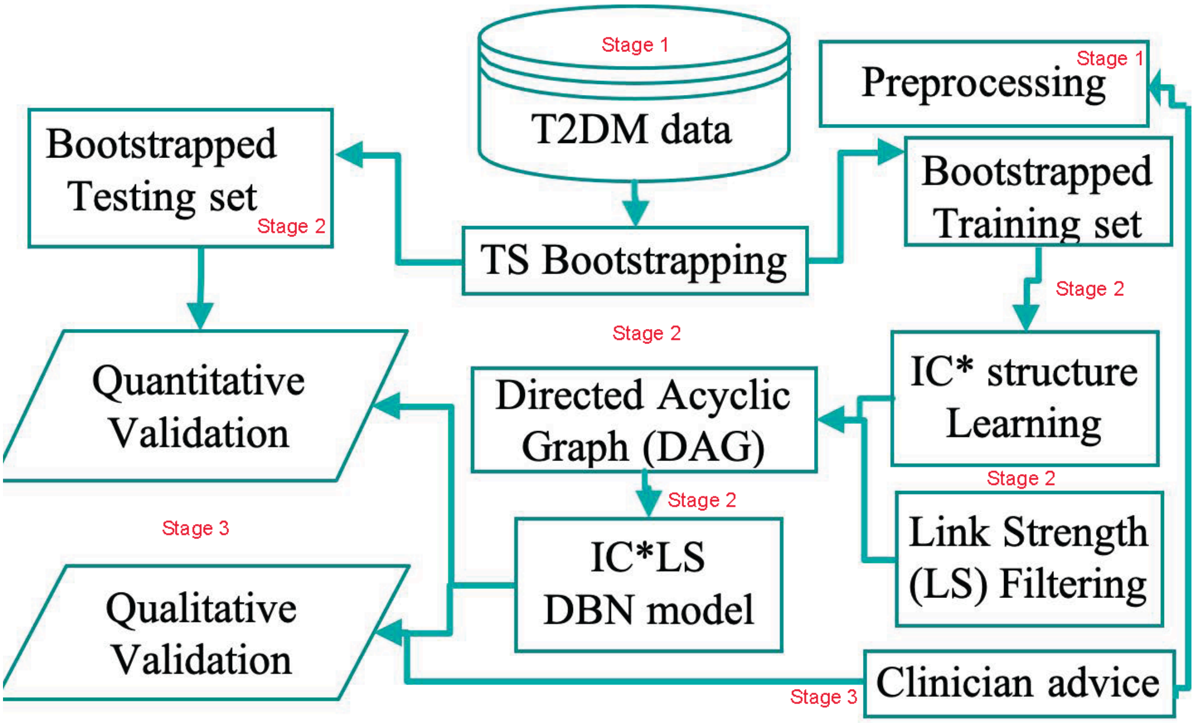

Fig. 1. IC*LS methods Diagram: The overall methodology of the proposed predictive model in three stages.

3.2. Pre-processing and Relational Models

In this research, different pre-processing techniques are used before modelling the clinical data such as: data representation and cleaning or removing all uninteresting and uninformative information about patients (e.g., dates of visits). The uninformative and bias records (e.g., a patient with only one visit) were truncated to filter out unnecessary data and reduce the outliers. In addition to this, Centre Profiling was employed to assess 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).

In these pre-processing stages, this study employed different tools such as Microsoft Access, Genie, Weka and SQL Server to manipulate, process and store the collected data at Pavia in the form of T2DM dataset. These software tools were chosen to interact with the database and extract the useful information gathered from all preliminary tables by using the Data Manipulation Language (DML). First, Relational Models in Database Management System (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 aided to build one single table from integration of primary tables (five tables were intuitively collected and cleaned 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.

3.3. Missing Values and Data Imputation

As explained previously, in this small-sized clinical dataset, 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 order to address the missing data and imputing the data, 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. The entire set of statistics was then changed by deleting value records randomly. To assess efficiency, only cases lacking missing data were taken into account. 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, thereby generating fictitious missing values to evaluate the ability of imputation [1].

4. Methods

The methods required to predict and validate the complications associated with T2DM are divided into three stages, which the overall methodology is represented in Figure 1 as well as a pseudo code illustrated in Algorithm 1 (Enhanced Stepwise Algorithm). In Figure 1, the first stage (Stage 1) describes the data and mostly focus on the descriptive data analysis and pre-processing approaches used in this research. Similarly, lines $1-13$ in Algorithm 1, shows the initialisation and prepossessing stages. The second stage, as was shown in 1-Stage 2 and line 14-18 of Algorithm 1, focuses on the proposed methodology and contributes to a re-balancing approach in a time series analysis with unequal lengths of patient visits (which was called TS bootstrapping). It then splits the T2DM data into train and test sets and then the bootstrap approach is applied to these sets.This stage explains how structure learning within causal networks are utilised to stratify patients based on the probabilities of latent variables during their visits. To learn the structure, the IC* algorithm, Fisher's Z test and the correlation matrix were employed to generate a Directed Acyclic Graph (DAG), as was illustrated in lines $19-20$ in Algorithm 1. Furthermore, line 21 trained a latent DBN structure to model the joint distribution of the domain representing probabilistic relationships between comorbidities and risk factors. Later, links within the DAG were filtered based upon their strength, as was represented in line 22. Therefore, a combination of IC* and Link Strength produced a predictive model (IC*LS DBN) by incrementally adding a latent variable to the observation. This information was used to visualise the identified latent variables that accounts more for understanding the latent variable.

The last stage, in 1-Stage 3, and lines $23-30$ in Algorithm 1, aims to increase the reliability of prediction measures and interpretation of the outcomes to only identifying the correct number of the latent variables by using the enhanced stepwise approach. Lines 27 in Algorithm 1 tested the latent DBN model and based on the accuracy of trained DBNs model, to make decision whether to repeat the learning/adding another latent variable in another step of the enhanced stepwise approach or not. Stage 3 in Figure 1 also compares the proposed methodology to the previous methods by introducing quantitative validation strategies, such as visit-based, patient-based, sensitive analysis, confidence interval. Consequently, it interprets the quantitative results based on clinician point of view and medical articles. Finally, it summarises and discusses how

[1] A detailed description of data are reported in the supplementary materials.

a greater focus on the smaller number of latent variables will be maintained to produce a better prediction results.

4.1. Prediction and Classification

Predicting the comorbidities has long been a question of great interest in a wide range of medical fields. 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.

This study attempts to predict the future state of a patient per visit by utilising a set of observed test based on the temporal complication. 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 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.[2] 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). A common problem with classifying/predicting complications in longitudinal data is that there may be many more visits where the complication does not manifest itself compared to those where it does (due to careful management), which is discussed below.

4.2. Imbalanced Issue In Complex Clinical Data

To predict a target complication, T2DM patients are classified into two categories (cases): positive and negative cases. In particular, if the overall number of patients in the positive case is far less than the negative case, the complication class is labelled as an imbalanced class. In this research, the minority class represents patients visit during which a complication is present and the patient that are suffering from the target complication. In fact, frequency of samples belonging to one class is severely different from the other ones. Therefore, binary classifiers bias to a class which demonstrates the majority of samples.

To re-balance the imbalanced data, it seemed necessary to note the actual incidence at which the result happened and proportional probability or likelihood ratios are reported (e.g., stating that one complication generated a certain result twice more probable than another complication). Thus, an unbalanced ratio is calculated as the ratio of negative to positive cases {number of negative cases}: {number of positive cases} for a specific complication to ensure a balance. For instance, the unbalance ratio of the majority to the minority based on the population size of responding binary class proportions in the dataset for RET, NEU, NEP, and LIV is defined as {3:1{, {4:1}, {3:1}, and {4:1}, respectively. Whilst this ratio for HYP is {1:5}.

[2] It is possible to show patterns of complications for each single visit with respect to any combination of complication co-occurrences, chosen from C as demonstrated in [31].

4.3. Re-balancing Strategy (Time Series Bootstrapping)

This section, in order to deal with the imbalance issue, describes a time series Bootstrapping methodology, which is called "TS Bootstrapping" and employed a variant on the re-sampling approaches introduced in [24–26, 42]. 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 re-balancing technique generally have been found to produce more accurate and reliable statistics [43]. 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. 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 the latent structure to be inferred. It also assumes that patient status at time $t-1$ depends on the corresponding hidden variable at a previous time t (Markov properties). As a result, the bootstrapped data contains an equal number of positive and negative cases for the target complication at time t. In the next section, the time series bootstrapped data is analysed within a DBNs model.

4.3.1. Dynamic Bayesian Networks Model

Dynamic Bayesian Networks (DBNs) are probabilistic graphical models for handling uncertain, noisy clinical time series data. These probabilistic networks are a more explicit representation of a domain through modelling the joint probability distribution (the probability of all possible outcomes in a domain). In this study, DBNs were used to compute the probabilities of the presence of comorbidities over time, given a set of risk factors. DBNs were trained on the bootstrap T2DM data and tested on their power to predict a complication at the next time point, before the latent variables were explored.

4.3.2. Latent Structure

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 [44]. The causal discovery indicates dependencies that are generated by structures with unmeasured factors, i.e., latent variables. The latent structure is a projection in which every latent variable is either a root node or a link to observed variables. One advantage of a latent structure is that they can better encode the actual dependencies and independencies in the data. In this approach, firstly, the probability of a high clinical level of the nodes and the learned hidden variables are then inferred using the BN inference. 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. The prior works to understand latent variable in [26] for learning the structure of the model, the standard structure learning algorithms e.g., K2 [45] to create the non-temporal links (Intra). REVerse Engineering ALgorithm (REVEAL) [46] was also utilised to identify the temporal links (Inter).

In the Latent DBN structure (as seen in Figure 2), nodes represent variables at distinct time slots and there are links between nodes over time, so they can be used to forecast into the future. Data mining and analysis were performed using MATLAB, Bayes Net toolbox [47], and “Graphviz” for visualisation. The networks with temporal associations were inferred from T2DM historical patients time series data whilst represented in two DBNs (t and t-1) under the Markov properties assumption. In the discrete-space/discrete-time DBNs, two-time steps are considered to show the relationship between risk factors. For instance, Figure 2 shows the first complication at time t-1 that affects states of all other comorbidities and risk factors at t. The Expectation-Maximization (EM) algorithm [48] are used to estimate the BN parameters, within a standard Bayesian network inference. Then, the resultant CPTs are used to indicate the probability of being in one state given the states of all associated risk factors from the relationship graphs. The main weakness of these algorithms was the failure to address how to learn a structure with the correct number of latent variables.

4.3.3. Induction Causation (IC)*

In order to address the issue discussed above, we used a constraint-based method (which is know as IC* algorithm) to calculate several conditional independence tests, while learning a latent variable structure associated with a set of observed variables. The IC* algorithm is similar to the PC algorithm, except that it can detect the presence of latent variables, which was introduced in [49]. It returns a partially Directed Acyclic Graph (DAG) to characterise the entire Markov equivalence class. The IC* relies on statistical significance tests to decide whether an arc exists between two variables and on its orientation[3]. In the next part, in order to discover the correct number of hidden variables, extra checks are conducted on the learnt DAG on the stepwise IC* algorithm.

4.3.4. Link Strength (LS) Metric

The Link Strength (LS) [50] is a measure to calculate the overall strength of the dependent links within the DBNs. It focuses on the most powerful dependencies between T2DM risk factors. It also enables us to observe the specific impact of each discovered edge in a DBN. The percentage points of uncertainty reduction in a variable were 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. For instance, if there was a link in the IC* adjacent matrix with LSTA greater or equal to some threshold (here 20 percent), a link in the final structure is retained; otherwise, it is deleted. We chose this threshold to avoid providing overly connected networks and loops in the DAG, as well as to decrease the risk of edge overfitting[4].

*4.3.5. Stepwise IC*LS Approach*

This paper proposed an extended IC* stepwise approach, which was reported in [28]. This method 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 proposed IC*LS 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. The key stages of implementing the enhance stepwise methodology are shown in Figure 1-Stage 2. As a result, the

[3] A default error rate ($\alpha = 0.05$) is used to find the correlation of T2DM risk factors using IC* algorithm.

[4] A detailed description of the LS metric is reported in the Supplementary Materials.

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.

Algorithm 1: Enhanced Stepwise Algorithm

```
Result: LatentDBNs
initialisation: $DS$ ← Original T2DM dataset;
Disease ← {RET,NEU,NEP,LIV,HYP};
Data = Discretise(DS);
Size(Data) ← 3959;
Count(Patients) ← 356;
Latent Variables ← {Hidden1, Hidden2, Hidden3, Hidden4};
Threshold = 0.25; alpha = 0.05;
n = Count(Observed Variables) = 13;
N = n;
overallAccurracy ← 0;
Accurracy ← 0.001;
for $Disease\ in\ DS$ do
    Data ← Cell Arrays of Patients based on their Visits;
    while $Accurracy > Max(overallAccurracy)$ do
        Sample a consecutive pair of visits from each cell array;
        Return the training indices for sampling with replacement;
        trainSet ← TS Bootstrap(TrainingData, Disease);
        testSet ← TS Bootstrap(TestingData, Disease);
        CorrMat ← Correlation Matrix(trainSet);
        DAG ← Structure Learning ($IC*$(Fisher Z test, CorrMat, N, alpha)) ;
        (DBNtrained, Intra, Inter) ← Train dbn(trainSet, N, Disease);
        LSTA ← Structure Learning(True Average Link Strength(DBNtrained));
        if $LSTA > Threshold$ then
            Accurracy ← Test Model(LSTA, N, Disease, testSet);
            overallAccurracy ← [overallAccurracy, Accurracy];
        end
        Add the discovered latent variable in the previous step to the observed nodes;
        N = n + Count(Discovered Latent Variables);
    end
end
```

5. Results

This section assessed the effectiveness of the bootstrap re-balancing method and the latent variable discovery approach in T2DM dataset. The results were documented for the following comparative structures:

- UNB-K2-REVEAL: a latent variable and a fully learned structure from unbalanced data by using the K2 algorithm for Intra links and the REVEAL algorithm for Inter links.

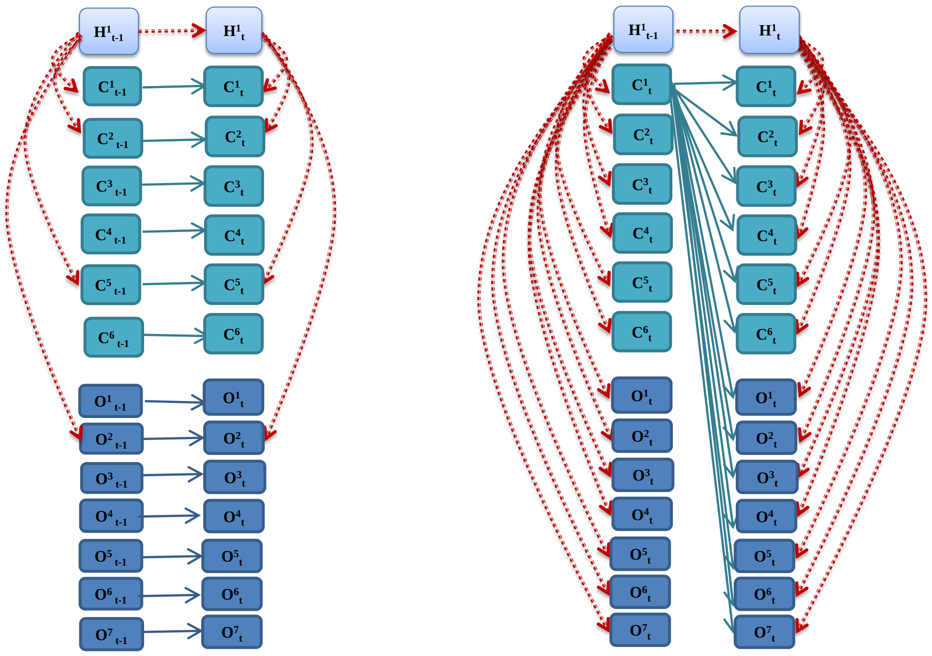

Fig. 2. Latent DBN Structure: The dynamic links, within time series structure in a DBNs model, observed by using the REVEAL algorithm (in the left-hand side) and Fully Auto-Regressive approach (in the right-hand side). The H, C, and O illustrate Hidden node, Complication, and Observed node, respectively.

- B-K2-REVEAL: a latent variable and a fully learned structure from the re-balanced (bootstrapped) K2 algorithm for Intra links and the REVEAL algorithm for Inter links with the balanced data using the TS Bootstrapping approach (shown in the left-hand side of Figure 2).
- No-latent: the network is fully learned from the re-balanced data by using PC algorithm without latent variable for Intra links. The dynamic structure for Inter links is Fully Auto-Regressive; each node is connected to the corresponding node in the next time slice.
- IC*: Several latent variables and a fully learned structure from the re-balanced data by using the IC* algorithm from the balanced data for Intra links and Fully Auto-Regressive structure for the Inter links, which was shown in [27].
- IC*LS: Several latent variables and a fully learned structure from the re-balanced data by using a combination of the IC* and LS filtering method for Intra links and Fully Auto-Regressive structure for the Inter links. Figure 3 represented four DAGs were learned in the stepwise IC*LS algorithm.

The proposed structure has been evaluated by performing the sensitivity analysis on the cohort based on two different perspectives: a “Visit-based” analysis and a “Patient-based” analysis and validation tests, as introduced bellow:

5.1. Visit-based Validation

The "Visit-based" analysis was defined with respect to each time point (as a single visit) belonged to a patient, which was scored individually as a "zero" or an "one". For example, once a

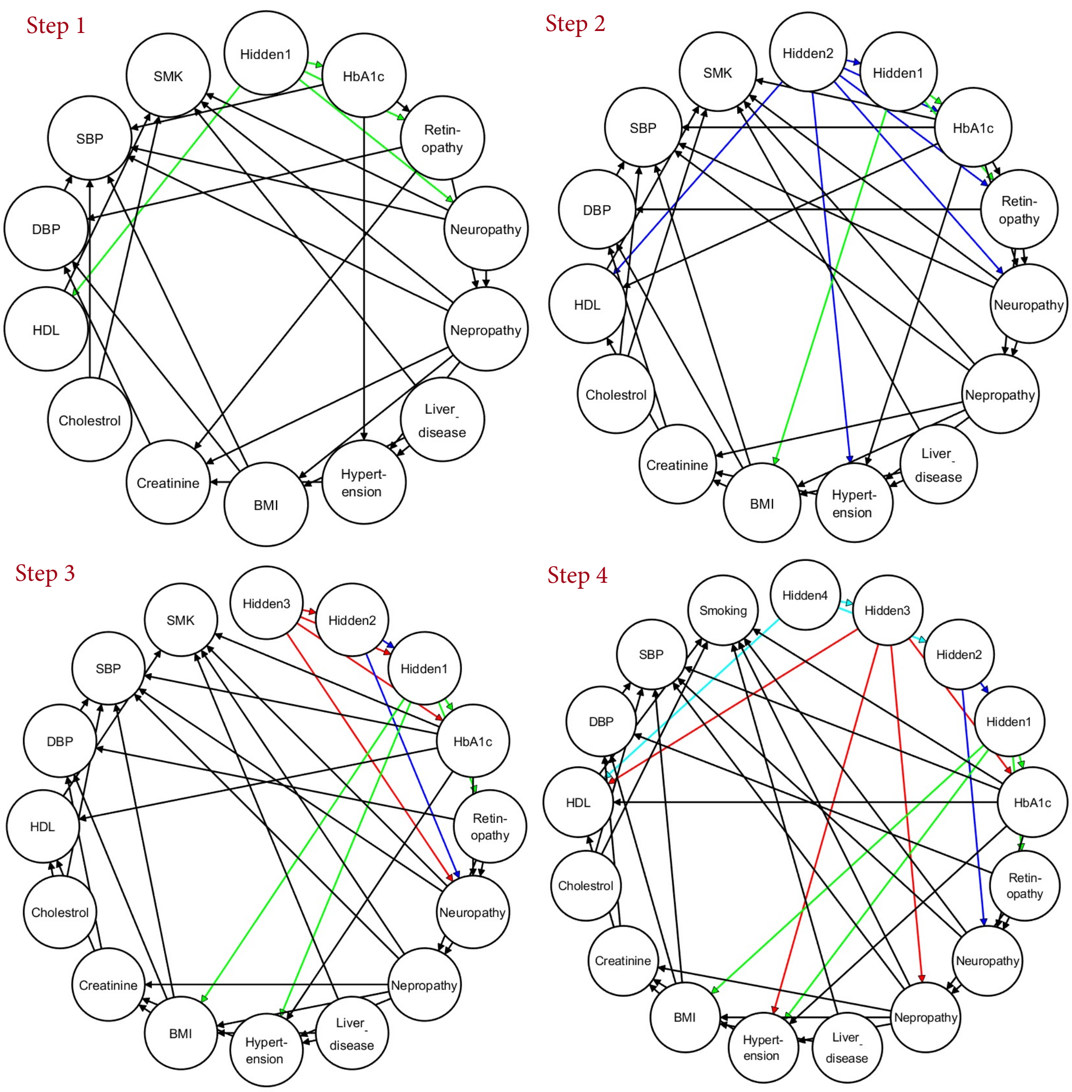

Fig. 3. IC*LS DAGs representing four steps in which a latent variable has been incrementally added to the prior latent structure. Hidden1, Hidden2, Hidden3, Hidden 4 are the first, second, third and fourth latent variables, which were learned in four steps of the enhanced approach.

target complication has occurred in a visit of a patient, an appearance of "one" on that specific visit was analysed regardless of the rest of visits for the corresponding patient. The findings obtained from the Visit-based sensitivity analysis showed that how many visits across all patients with having a specific complication were also correctly tested positive (comparison between previous and current values for the given visits). Visit-based specificity results revealed that how many visits which are equal to zero were also correctly tested negative. Table 3 illustrated that whether the complications were predicted correctly or not depending on the characteristics of time series data (based upon the Visit-based analysis). In this Table, the patient time series (corresponding to the first four visits of each T2DM patient) were assessed to obtain the classification accuracy in predicting retinopathy, liver disease and hypertension. The time series were analysed considering the Area Under Receiver Operator Characteristic Curves (AUCs). The overall results showed that the proposed TS Bootstrapping method (B-K2-REVEAL) provided more accurate prediction compared to the unbalanced model (UNB-K2-REVEAL) in Table 3. In addition, the IC* and IC*LS approaches were compared to a NO-latent method. It was sadistically evident that IC*LS based predictive model were more precise than the IC* based models, without an LS filter classification results were dropped considerably. This improvement was potentially achieved because the LS measure filtered out the less robust links, thus avoided overfitting. While the IC*LS approach was compared to the IC* approach, these two models were in a higher level of prediction accuracy over the models with no latent variable, hence, the use of the latent methods enhanced AUC values significantly.

Table 3

Visit-based performance assessment on the prediction results in percentage.

Performance Measure	UNB-K2-REVEAL	B-K2-REVEAL	NO-latent	IC*	IC*LS
AUC of Retinopathy	35	50	94	89	99
AUC of Liver Disease	38	51	71	92	99
AUC of Hypertension	60	51	65	83	99

5.2. Patient-based Validation

A "Patient-based" analysis was utilised to identify the appearance of a specific complication across a patient's time series. By detecting any "one" in any time point belong to a patient (over all patient's visits), it was assumed that that specific complication has occurred for the patient. Once a complication has been diagnosed in any visit belonging to a patient, where the patient was directed to join to the positive case, otherwise became a member of the negative case. Similarly, once a patient has been located in a positive case ($C_i = 1$), the patient stayed in that case throughout their time series, so, it was recorded for the rest of the visits (time series) as people were not recovered once have been diagnosed. As a result, those patients who have been already at a high risk of developing complications, it was assumed that they would not switch from a positive case to a negative case.

Table 4 represented the performance measures (sensitivity and specificity) obtained for the proportion of T2DM patients who correctly tested positive/negative in predicting two complications (retinopathy and liver disease in the Patient-based validation strategy) and then compared to the Visit-based strategy. Patient-based sensitivity results showed that how many patients that actually having a complication were identified correctly with that complication (comparison between predicted class value for a patient's complication and actual class value of the complication). Patient-based specificity validation test represented that how many patients without a specific complications were also correctly tested negative (not having the complication). Switching the methodology from B-K2-REVEAL to IC*LS in predicting liver disease showed a massive enhancement, first in sensitivity Patient-based assessment, from 71% to 90%; second, in Visit-based specificity experiments from 85% to 99% (as can be seen in Table 4-Liver disease (B-K2-REVEAL compared to IC*LS). Sensitivity was measured for retinopathy prediction by switching method from the standard methods (B-K2-REVEAL) to IC*LS increased sharply from 69% to 86% (Patient-based) and 75% to 92% (for Visit-based). Despite this, for retinopathy the total number of patients, which was predicted correctly without the disease (specificity or true negative rate) remained almost constant or improved slightly from 89% to 90%. According to these results, it seemed to be evident that better prediction performance had been generally achieved by using the IC*LS method[5].

Table 4
Comparison of Patient-based and Visit-based prediction accuracy percentage.

Complication Method	Retinopathy B-K2-REVEAL	Retinopathy IC*LS	Liver disease B-K2-REVEAL	Liver disease IC*LS
Sensitivity (Patient-based)	69	86	71	90
Sensitivity (Visit-based)	75	92	94	95
Specificity (Patient-based)	89	90	98	99
Specificity (Visit-based)	89	90	85	99

5.3. Confidence Interval Results

In this section the experimental findings and their significance were tested statistically by using the confidence interval. Looking at how the different structures were performed within a DBN for predicting the appearance of complications, which was illustrated in the fourth step of the enhance stepwise IC*LS seen in Figure 3, to report more precise results, confidence intervals to manage the uncertainty in the prediction results were 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. 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 4 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 4 revealed that the prediction accuracy of retinopathy in step one had been increased sharply by adding latent variables at step two to four and then dropped slightly at step five. Additionally, error bars on the top of the bar charts were illustrated. For example, the error bar in step "NO-Latent" was smaller than the first step, while the firth step was significantly at lower level comparing to the subsequent steps. The error bar in step two is

[5] A detailed explanation of the confusion matrix results is reported in the supplementary material.

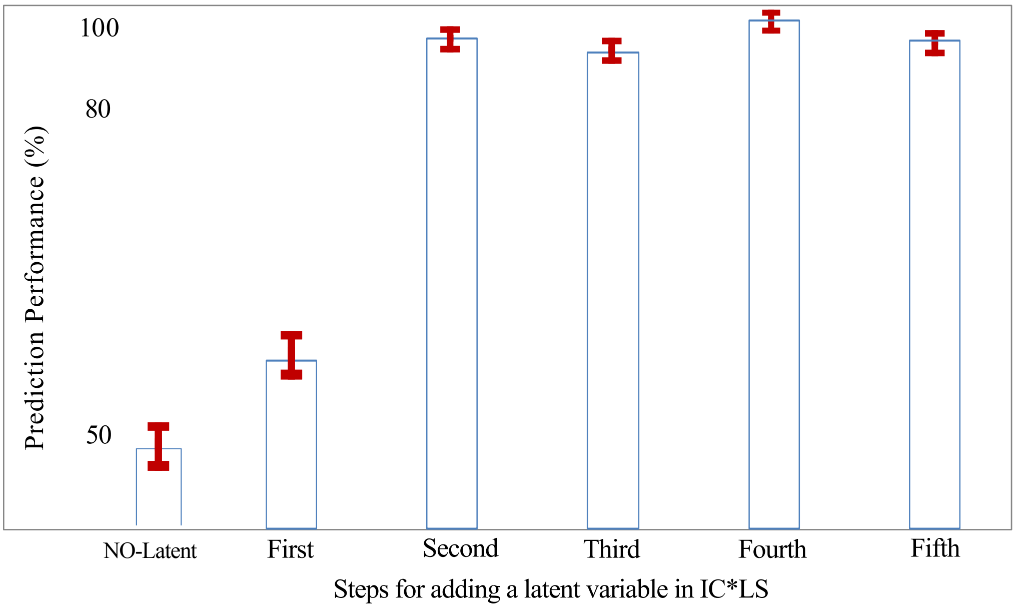

Fig. 4. Bootstrap Confidence Interval: A general improvement of overall performance percentages for predicting retinopathy. It compares the predictive models without a latent variable (the right hand side - "NO-Latent") and different steps of adding a latent variable using IC*LS (in Visit-based analysis).

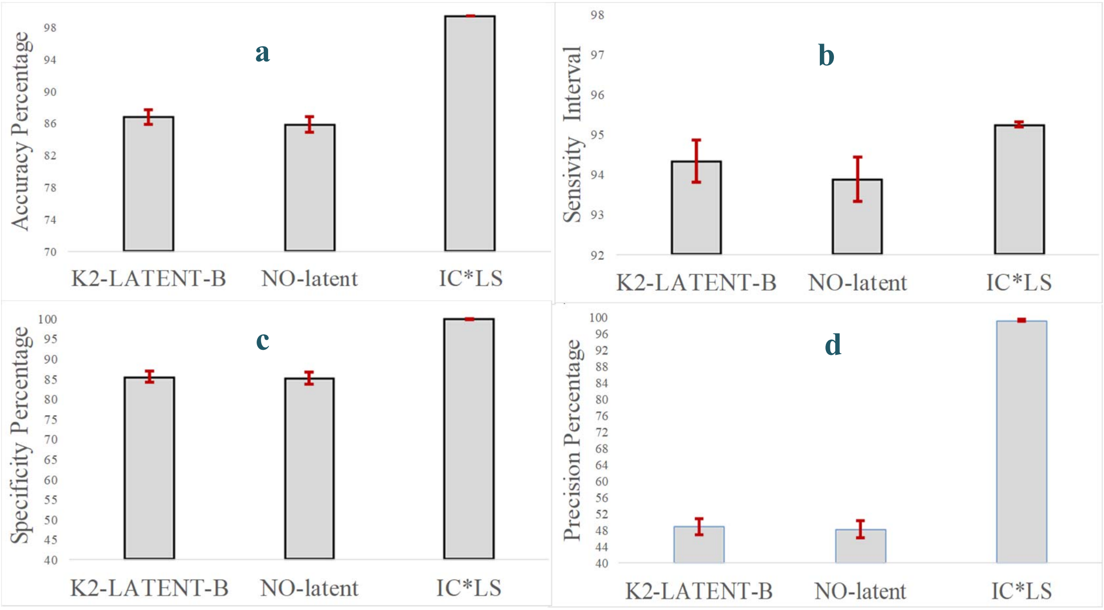

Fig. 5. Bootstrap Confidence Interval statistically checks the significance of accuracy, sensitivity, specificity, and precision to predict liver disease, which is compared to the standard approach as well as the "NO-Latent" models.

quite large due to a more considerable confidence interval of the successive steps, and in the forth step is the smallest. Overall, Figure 4 revealed that there could be a general trend to improvement in accuracy as more latent variables have been added. Surprisingly, this improvement levelled out after adding the fifth hidden variable.

Figure 5 illustrated the influence of the latent variable on the bootstrapped data in predicting liver disease. These results showed how the targeted use of latent variables improved prediction performance over standard approaches as well as aiding the understanding of relationships between these latent variables and disease complications/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 K2 and REVEAL algorithm as well as no latent variable approaches[6]

5.4. Latent Variable as Evidence

In this section, we look at the influence of the latent variable as evidence to predict the complications. As mentioned earlier, the aim of this research was to explore the impact of the targeted latent variable on prediction of the T2DM complications. Here, we validated the results to uncover influential factors in the diagnosis with regard to a set of diagnosis targets given the evidence. This target was possible to achieve relying on the nature of the Bayesian inference, where any of T2DM comorbidities or risk factors could be queried using a joint probability distribution. Therefore, any trigger or impact on the probabilities of complications was assessed/monitored once the clinical class as evidence has shifted or changed (in which the latent variable values were set to either the highest risk or the lowest risk level).

Figure 6 illustrated how prediction probabilities for retinopathy, liver disease and hypertension were affected by changing evidence (the fourth latent variable). The red arrows shows the changes in the evidence (the latent variable value is triggered from zero to one and one to zero). The bar chart in Figure 6-a showed how the probability of retinopathy being in its high value (the diagnosis point) changed by setting the evidence on the comorbidity-related risk factor. For instance, the marginal probability of one or more patients being diagnosed with retinopathy is 0.05 (no evidence set). According to the DBN model, setting evidence on the latent variable (from the latent variable = 0 to 1) increased the marginal probability of retinopathy from 0.05 to 0.29. This could be used to reassure that the latent variable has a significant impact on the target complication posterior probability. On the other hands, in Figure 6-b setting evidence on the latent variable (from the latent variable = 0 to 1) changes the posterior probability of liver disease slightly (from 0.99 to 0.97) and in Figure 6-c hypertension (from 0.95 to 0.88). Therefore, the discovered latent variable in the last (fourth) step of IC*LS could negatively associated to the development rate of liver disease and hypertension while it had a much stronger positive impact on retinopathy.

5.5. Latent Variable Validation Pattern

Figures 7-9 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 7-d

[6] The detailed findings were reported in the supplementary materials (Confidence Interval Results).

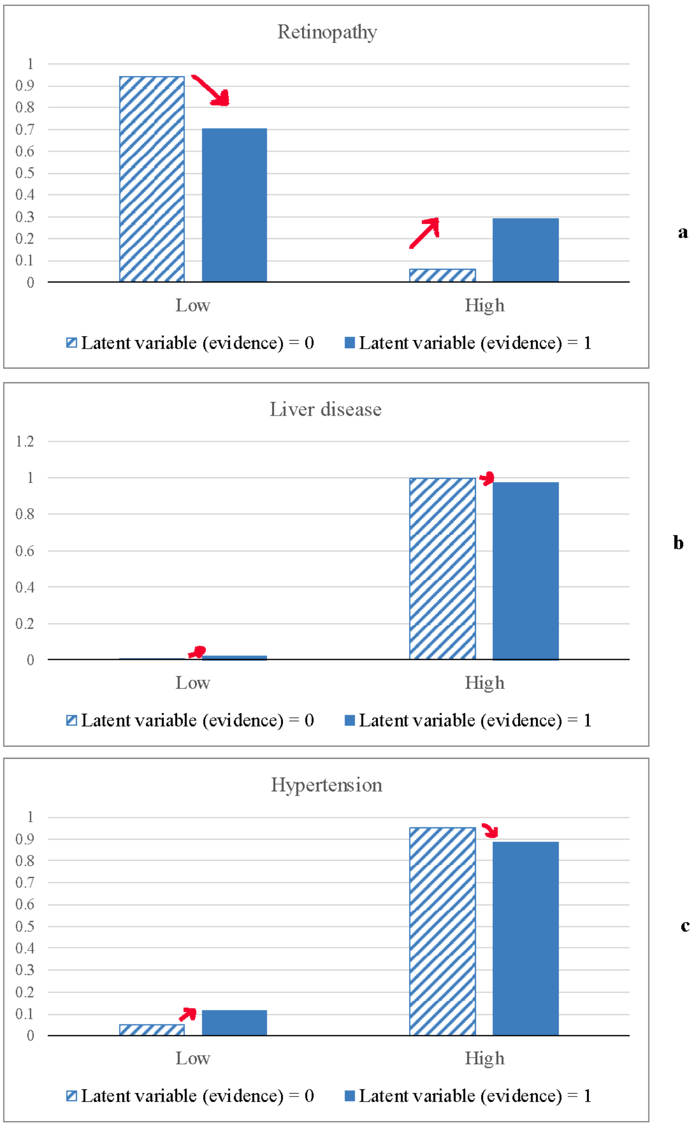

Fig. 6. The impact of the latent variable as evidence to predict the complications.

seemed to be to very similar to its observed probability shown in Figure 7-a, which indicated the complication occurrence slightly earlier than the prediction. The IC*LS latent approach, in Figure 7-c for liver disease, revealed a trigger around the clinician observation time, whereas the latent K2 process in Figure 7-b remained steady. A less significant predicted probability was also captured in Figure 8-d. This illustrated a fluctuation just before retinopathy has been monitored in Figure 8-a. Similarly, a trigger happened in two latent approaches in Figures 8-b-c.

These results 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

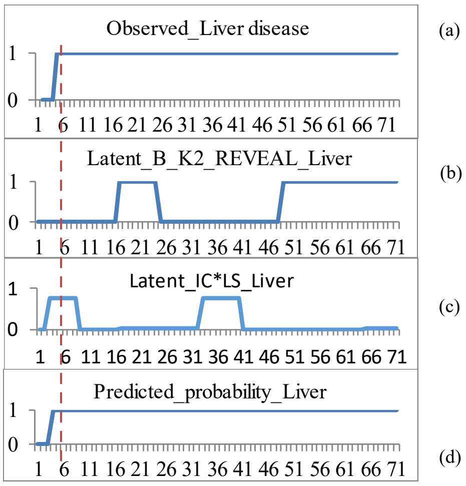

Fig. 7. Latent variable prediction pattern of liver disease over time (a patient follow-ups).

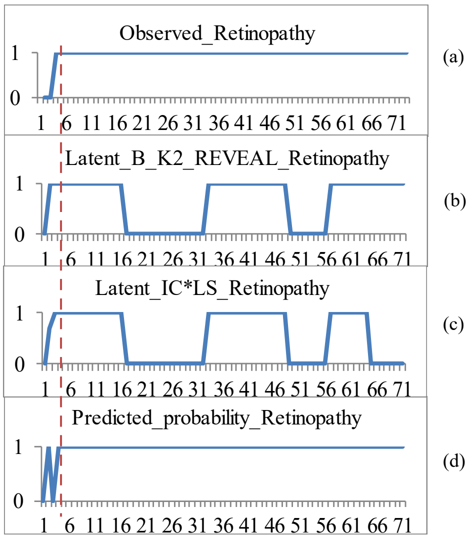

Fig. 8. Latent variable prediction pattern of retinopathy over time (a patient follow-ups).

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 Figure 9-d comparing to Figure 9-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 contrast, retinopathy as a chronic microvascular has been

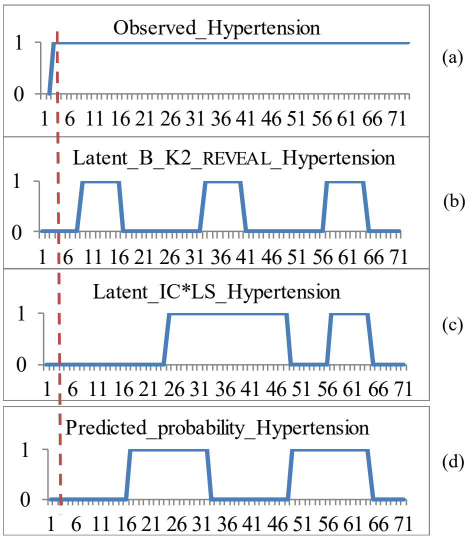

Fig. 9. Latent variable prediction pattern of hypertension over time (a patient follow-ups).

known very challenging to be caught at the earlier stage of the disease progression.

6. Discussion

This study has claimed that 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. 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 latent variables correlated with these components will differ as the trends for the observed ecosystem components change. The latent variable influenced the likelihood of developing a disease/complication depends on all the risk factors of which it has been related, and a shift in trends means that process relationships have shifted. According to these hypothesis, Figure 6 showed how the probabilities of retinopathy, liver disease and hypertension were influenced by adding the fourth latent variable to the DBNs. Surprisingly, in 6-b, a slight change was found in liver disease values, whilst the latent variable was in its highest value.

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. As a result, there was a significant negative correlation between the latent variable and hypertension, which was shown in 6-c. The AUC results obtained in Table 3-UNB-K2-REVEAL predicted hypertension accurately 60% of times comparing to 35% for retinopathy while data was imbalanced. This Table also 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. Although there was a direct link (correlation) between the latent variable and BMI in Figure 3- Step2,3,4, these results should be interpreted with caution, as this did not necessarily mean that the latent variable caused BMI.

As a result of including this latent variable, in Table 4 there was a steep rise in the prediction accuracy of hypertension from 69% to 86% (as the IC*LS was compared to NO-latent). Similarly, a positive correlation was found between the latent variable and retinopathy in Figure 6-a. It was apparent from Table 3 that retinopathy prediction was enhanced considerably from 94% (NO-latent) to 99% (IC*LS) by adding the forth 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 overall approach in this paper is abstracted in Figure 10. In the left-hand side of Figure 10, 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 (in the right-hand side). 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.

6.0.1. MOSAIC Tool

The data is belonged to the MOSAIC European Union project retrieved from MOSAIC website [29]. 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 [51].

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.

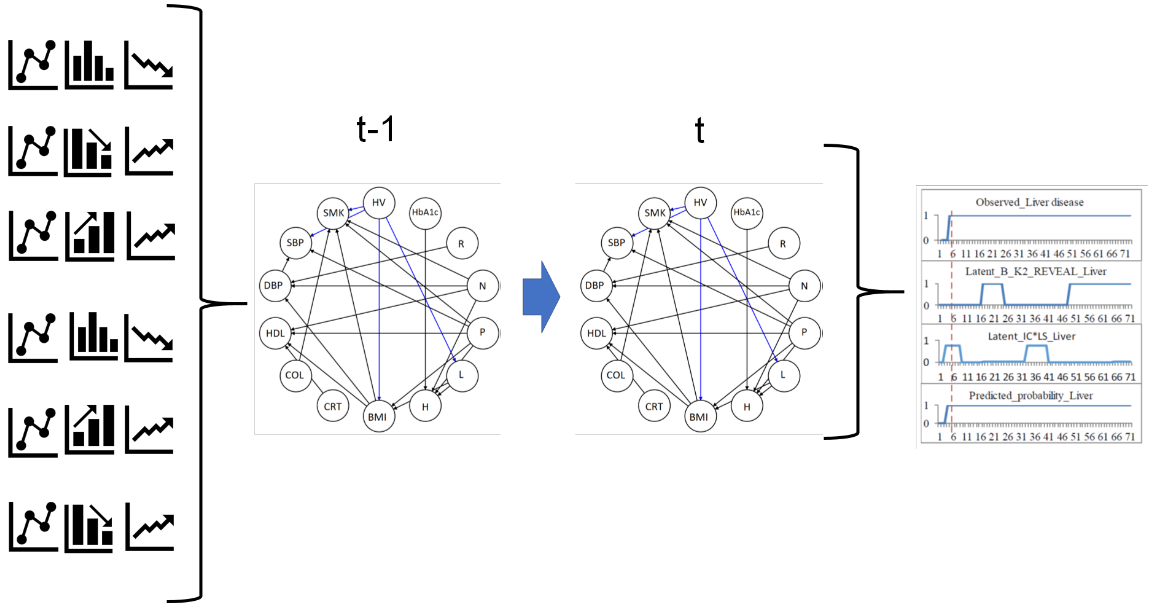

Fig. 10. A DBN Latent Model: From the left hand side, in the middle, and the right hand side demonstrate the patients history, the inferred latent variable probabilities, the prediction, respectively.

7. Conclusion

Diabetes specialists predict disease and comorbidities based on their knowledge of the disease and an individual patient's clinical history. This is a complex task because of the existence of unmeasured risk factors in the data, various responses to the disease, and heterogeneity in monitoring patients. Here, we modelled unmeasured factors by considering an approach to model progression using latent variables with a focus on trying to understand their behaviour and meaning. We exploited a DBN model because of the transparent way of modelling data as well as the flexibility in incorporating latent variables.

This paper contributed in several ways to our understanding of how the latent variable provides a basis for a better prediction of the T2DM complications. This paper was an extension of this paper author's previous works on the same dataset, where an intuitive stepwise method, based upon the IC* approach was developed to learn the effects of multiple hidden variables on the prediction performance. The contribution of this paper compared to the previous papers in which time series bootstrapping was used for re-balancing the data and providing a higher level of confidence in the prediction results. Analysis of the subset of patients with unequal number of visits is dealt with using a bootstrap technique that has been specifically designed for the longitudinal data to identify targeted complications among the trajectories which are key stages in the disease progression. Our results showed that our re-balancing approach by the use of TS bootstrapping method for an unequal number of time series visits demonstrated a better improvement in the prediction performance of the disease compered to the prior works.

Nevertheless, the discovery of the optimum number of the hidden variables was not easy and sometime accuracy dropped as more were added due to overfitting. To address this issue, this article adapted the IC*LS approach represented an enhanced variation on the stepwise IC* method for incrementally identifying hidden variables. We incorporated the IC* algorithm and a Mutual

Information based scoring metric to identify the strength of relationships between the latent variable and clinical risk factors. 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. Thus, the use of the IC*LS approach has provided a significant improvement on the accuracy of prediction while reducing uncertainty in the disease management. The most highlighted contribution of this paper gained insight by interpreting the latent states while the association among the disease complications are taken into consideration.

This work could be also applied to find the most influential latent variable as a temporal phenotype to identify the overall patterns of risk factors for each patient over time. This could lead to a better understanding of risk factors and patient-specific interventions. A natural progression of this work in the future involves extending the latent DBN models with more latent variables to capture a greater variety of factors to characterise critical changes. Our proposed approach will be useful for stratifying patients according to their probability of developing complications and clinician advice. For example, there is room for further progress in determining the optimal number of latent variables using Partial Least Squares (PLS). In addition, we will seek more advice from clinicians in interpreting hidden factors and their correlation toward other T2DM risk factors and complications as well as the disease prediction process. We also intend to look at more geographical and clinical factors, such as family history, pollution factors, and glucose levels.

References

[1] C.D. Mathers and D. Loncar, Projections of global mortality and burden of disease from 2002 to 2030, *PLoS medicine* **3**(11) (2006), e442.

[2] 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) (2005), 617–625.

[3] 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) (2001), 392–400.

[4] 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*, Morgan Kaufmann Publishers Inc., 1998, pp. 139–147.

[5] 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) (2008), 515–529.

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

[7] C. SPEARMAN, " General Intelligence," objectively determined and measured, *American Journal of Psychology* **15** (1904), 201–293.

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

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

[10] 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) (2006), 219–236.

[11] 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) (2016), 19–26.

[12] 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*, Vol. 2016, American Medical Informatics Association, 2016, p. 470.

[13] 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) (2018), 424–432.

[14] 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*, IEEE, 2012, pp. 471–472.

[15] 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** (2015), 369–376.

[16] 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*, IEEE, 2009, pp. 5629–5632.

[17] 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) (2011), 416–430.

[18] 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.

[19] A. Tucker, X. Liu and D. Garway-Heath, Spatial operators for evolving dynamic Bayesian networks from spatio-temporal data, in: *Genetic and Evolutionary ComputationGECCO 2003*, Springer, 2003, pp. 205–205.

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

[21] M. Grzegorczyk and D. Husmeier, Non-stationary continuous dynamic Bayesian networks, in: *Advances in Neural Information Processing Systems*, 2009, pp. 682–690.

[22] 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) (2005), 321–341.

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

[24] 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*, IEEE, 2016, pp. 282–291.

[25] 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** (2015), 142–158.

[26] 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*, IEEE, 2017, pp. 205–206.

[27] 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)*, IEEE, 2018, pp. 106–111.

[28] 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)*, IEEE, 2018, pp. 1040–1044.

[29] 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) (2013), R–162.

[30] 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)*, IEEE, 2019, pp. 198–203.

[31] 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).

[32] 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) (2015), 1119–1125.

[33] 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) (1998), 823–828.

[34] S. Colagiuri, Glycated haemoglobin (HbA1c) for the diagnosis of diabetes mellitus–practical implications., *Diabetes research and clinical practice* **93**(3) (2011), 312.

[35] 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) (1993), 817–817.

[36] 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) (1995), 407–415.

[37] 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) (2007), 734–743.

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

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

[40] 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) (2013), 57.

[41] 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) (1999), 1152–1156.

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

[43] L. SIMAR, An Invitation to the Bootstrap: Panacea for Statistical Inference?, *Institut de Statistique, Universite Catholique de Louvain, Louvain* (2008).

[44] X. Zhang, K.B. Korb, A.E. Nicholson and S. Mascaro, Latent Variable Discovery Using Dependency Patterns, *arXiv preprint arXiv:1607.06617* (2016).

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

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

[47] K. Murphy et al., The bayes net toolbox for matlab, *Computing science and statistics* **33**(2) (2001), 1024–1034.

[48] T.K. Moon, The expectation-maximization algorithm, *IEEE Signal processing magazine* **13**(6) (1996), 47–60.

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

[50] I. Ebert-Uphoff, Measuring connection strengths and link strengths in discrete Bayesian networks, Technical Report, Georgia Institute of Technology, 2007.

[51] 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)*, IEEE, 2014, pp. 244–247.