# Opening the Black Box: Discovering and Explaining Hidden Variables in Type 2 Diabetic Patient Modelling

Leila Yousefi[*], Stephen Swift[*], Mahir Arzoky[*], Lucia Saachi[†], Luca Chiovato[‡], Allan Tucker[*]

[*]Dept. of Computer Science, Brunel University London,London, United Kingdom
[†]Dept. of Computer Science and system,University of Pavia,Pavia, Italy
,
[‡]Unit of Endocrinology,Pavia, Italy
Email: Leila.Yousefi,Stephen.Swift,Mahir.Arzoky,Allan.Tucker@brunel.ac.uk,Lucia.Sacchi,Luca.Chiovato@unipv.it

*Abstract*—Clinicians predict disease and related complications based on prior knowledge and each individual patient's clinical history. The prediction process is complex due to the existence of unmeasured risk factors, the unexpected development of complications and varying responses of patients to disease over time. Exploiting these unmeasured risk factors (hidden variables) can improve the modeling of disease progression and thus enables clinicians to focus on early diagnosis and treatment of unexpected conditions. However, the overuse of hidden variables can lead to complex models that can overfit and are not well understood (being 'black box' in nature). Identifying and understanding groups of patients with similar disease profiles (based on discovered hidden variables) makes it possible to better understand disease progression in different patients while improving prediction. We explore the use of a stepwise method for incrementally identifying hidden variables based on the Induction Causation (IC*) algorithm. We exploit Dynamic Time Warping and hierarchical clustering to cluster patients based upon these hidden variables to uncover their meaning with respect to the complications of Type 2 Diabetes Mellitus patients. Our results reveal that inferring a small number of targeted hidden variables and using them to cluster patients not only leads to an improvement in the prediction accuracy but also assists the explanation of different discovered sub-groups.

*Index Terms*—Time-series clustering, hidden variable discovery, Temporal Phenotype, Diabetic Patient Modeling, Dynamic Bayesian Networks

## I. INTRODUCTION

Type 2 Diabetes is traditionally known as Type 2 Diabetes Mellitus (T2DM), and has been known for thousands of years. According to the World Health Organization, the number of people with T2DM has increased four-fold since 1980 [1]. In addition, T2DM patients are at increased danger of vascular comorbidities, such as nephropathy, neuropathy, liver disease, and retinopathy. Early diagnosis and prevention techniques are needed to reduce the associated mortality and morbidity caused by these complications [2]. However, studies that enables early predictions of diabetes using predictive models are limited [3]. 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, which help patient stratification, predict complications and understand the disease process. Hidden variable modelling [4] has a long tradition in causal discovery. Moreover, it can improve classification accuracy and boost user confidence in the classification models [5]. Among these, Bayesian models suit disease progression analysis very well because of their ability in modeling relationships and their ease of interpretation [6]. In particular, in biomedical science and clinical decision support, Bayesian Networks (BNs) have become a popular representation for dealing with uncertainty domain knowledge [7]. The learning of BNs is a critical research area, which depends on exploring the space of models for those which can best capture a pattern of data [8]. For investigating the prognosis of diabetes type 2 complications, some researchers used logistic regression and Nave Bayes with diverse modelling strategies, including methods for unbalanced data [9]–[11]. Ceccon and coauthors [12] exploited a variation of the naive Bayes classifier, which explored a hidden variable for segmenting patients into disease subtypes while classifying Glacuoma patients based upon visual field data. They used BNs inferred from cross-sectional data with one hidden variable connected to all variables. An extension of BNs for modeling temporal processes is often used in medicine, the Dynamic Bayesian Network (DBNs) [13]. DBNs represent clinical knowledge explicitly in terms of conditional independencies over time. Here, we consider time series data involving follow-up visits and an extended version of DBNs that can encode non-stationarity DBNs [6]. Time series clustering is often problematic [14], 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. For instance, [15] proposed an algorithm to cluster patients based on clinical data whilst utilizing the clustering information for identifying distinct patterns. There are some methods that uncover temporal patterns and relationships among clinical variables include causal information [16], numeric time series analysis [17].

### A. Data

The T2DM dataset was exploited in this work contains 13 clinical risk factors for a binary classification problem of 314 patients that have suffered from type 2 diabetes, aged 25 to 65 years inclusive, were recruited between 2009 and 2013

| Node | risk factors | | Clinical feature | Complication | Hidden variable |
|---|---|---|---|---|---|
| 1 | HbA1c (%) | $6.6 \pm 1.2$ | YES | NO | NO |
| 2 | Retinopathy | 0,1 | NO | YES | NO |
| 3 | Neuropathy | 0,1 | NO | YES | NO |
| 4 | Nepropathy | 0,1 | NO | YES | NO |
| 5 | Liver Disease | 0,1 | NO | YES | NO |
| 6 | Hypertension | 0,1 | NO | YES | NO |
| 7 | BMI (kg/m2) | $26.4 \pm 2.4$ | YES | NO | NO |
| 8 | Creatinine (mg/dL) | $0.9 \pm 0.2$ | YES | NO | NO |
| 9 | Cholestrol (mg/dL) | $0.9 \pm 0.2$ | YES | NO | NO |
| 10 | HDL cholesterol (mmol/l) | $1.1 \pm 0.3$ | YES | NO | NO |
| 11 | Diastolic blood pressure(DBP) (mmHg) | $91 \pm 12$ | YES | NO | NO |
| 12 | Systolic blood pressure(SBP) (mmHg) | $148 \pm 19$ | YES | NO | NO |
| 13 | Smoking Habit | 0,1,2 | YES | NO | NO |
| H1 | First Hidden variable | [0,1] | NO | NO | YES |
| H2 | Second Hidden variable | [0,1] | NO | NO | YES |
| H3 | Third Hidden variable | [0,1] | NO | NO | YES |
| H4 | Fourth Hidden variable | [0,1] | NO | NO | YES |
| H5 | Fifth Hidden variable | [0,1] | NO | NO | YES |

from clinical follow-ups of diabetes patients at the IRCCS Istituti Clinici Scientifici (ICS) Maugeri of Pavia, Italy [18]. Data mining and analysis were performed using MATLAB and Bayes Net toolbox [19] and for visualization we used Graphviz.

## II. METHODOLOGY

A systematic understanding of how latent variables contribute to T2DM complications is still lacking. To the best of our knowledge, no previous work on prediction process of Diabetes patients focused deeply on explaining the discovering hidden variables, in a DBN framework, in terms of how they capture the temporal phenotype for groups of patients. Thus, we aim to show here how a systematic understanding of hidden variables contribute to the modeling of the disease. Previously, [20], we have developed methods to discover the relationship between learned hidden variables and observed variables, using a DBNs model [13] to predict T2DM complications leading to a better understanding of the disease risk factors, while balancing data based on our time-series bootstrapped approach. In [18], we have made a start by developing an intuitive stepwise method to learn these latent effects based upon the IC* algorithm, using Pair-sampling for balancing data. For enhancing our previous methodology in [18], instead of using Pair-sampling, we balanced data simply by oversampling the minority patient cases (Positives) up to the majority patient cases (Negatives) (e.g., retinopathy has the imbalanced ratio of 3.2, comparing Positives to Negatives). Then, we exploited an enhanced version of a Stepwise approach in discovering more hidden factors. we focused on models that captured unmeasured effects in a more targeted manner so that knowledge about disease processes can be extracted, while accuracy is improved at the same time. To this end, we employed the Stepwise IC* (Induction Causation) approach, which incrementally adds hidden variables based on the IC* algorithm in classifying T2DM data in a structure learning process, an enhanced variation of our previous work in [18]. The IC* is a constraint-based method which applies conditional independence analyses to infer causal structures and learns a partially-oriented directed acyclic graph (DAG) with hidden variables [21]. The probability of a high state of any learned

hidden variables is then inferred using a standard Bayesian network inference. The inferred probabilities of the hidden variable are treated as observations which means that we can then treat the hidden variable as an observed variable in the subsequent step. In the next step, IC* is applied again to see if the new observed variable uncovers any new hidden variables. This is repeated until either no other hidden variables are discovered or chain connections between hidden variables are destroyed and classification performance is reduced. To find more hidden variables, we need an informative DAG, [22] representing the Markov equivalence class of our Bayesian model. In this paper, we add extra steps to the IC* algorithm (in [18]) with oversampling randomly positive patient cases (minority class) and then apply IC* one more time to give another chance for DAG to learn more hidden variables. Furthermore, we utilised a better stopping point.

In this study, we have discretized risk factors (as variables) and their parent in Bayesian learning models. Therefore, their associated factor can be interpreted as a simple look-up table or Conditional Probability Table (CPT), which indicates the probability of being in one state given the states of all correlated risk factors from relationship graphs. We used these CPTs to complete the DBN using parameter estimation. Our goal is to define similarity among patients by distance over either hidden factors or correlated risk factors. Having discovered the hidden variables, we built a DBN predictive model. We perform parameter estimation using the Expectation Maximization (EM) algorithm. For learning the temporal links of our DBNs, we used REVerse Engineering ALgorithm (REVEAL) algorithm [19], following work in [20]. Furthermore, we assumed that patient status at time $t$ depends on the corresponding hidden variable at a previous time $t$-1. All time series visits per patient has a different profile of 13 different risk factors, which they are correlated to each other as well as hidden variable. For identifying patient groups (clusters) in the clinical time series data set, a vector of uneven length of follow-up visits for each hidden variable should be considered as a comparing pair of patients. However, manifesting a right distance function for comparing the pair of patients would be a challenge. Capturing these local and dynamic correlation across the similar pattern among risk factors while calculation an average for each patient time series risk factor would be another challenge. On the other hand, discovery of such clusters of patients is essential in revealing substantial correlations in T2DM risk factors in response to the disease over time. Thus, the concept of similarity in one cluster is based on distances between two patients across their unequal follow-up visits. We exploited Dynamic Time Warping (DTW) [23] as a distance metric for the discovered hidden variables. We then used discovered hidden variables probabilities to group patients into clusters. DTW distance is a suitable measure to evaluate the similarities and dissimilarities of time series with respect to their shape. In our work, univariate 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.

We need to keep 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. In order to characterize the profile of each discovered cluster, we apply medoid analysis [24] to the DTW distance matrices to extract the patient with the smallest inter-patient distance from $n$ clusters of sub-groups. Thus, any sub-group can be represented by a hidden variable pattern, which can be thought of as a temporal phenotype for the cluster of patients.

## III. EXPERIMENTAL RESULTS

Each patient has a unique profile at each visit (time series). In addition, the status of two T2DM patients' risk factors during their time series is subject to change, as their level may rise and fall over time. Based on our clustering model, two patients are similar if they exhibit a similarity in their discovered hidden variables. Interestingly, our results demonstrates at the end of process, patients in a different sub-group share a similar profile of observed risk factors, while these factors were not taken into account in the cluster decision making process. This will ease prediction process for clinicians.

In Figure 1, a DAG for each step of the Stepwise IC* approach is learned. 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 1).

Looking at how these differing structures perform within a DBN for predicting the appearance of complications, Figure 2 illustrates that there is a general trend to improvement in accuracy as more hidden variables are added but this improvement levels out after the addition of the fifth hidden variable. A precise estimate of the uncertainty related with parameter estimation is essential to avoid misleading inference. This uncertainty is typically outlined by a confidence interval, which is professed to incorporate the true parameter value with a predefined likelihood. We confine ourselves to confidence intervals to manage the uncertainty in the prediction results derived from a randomly selected subset of T2DM patients.

The effect of adding a hidden variable at each step of the enhanced Stepwise is assessed on the randomly oversampled T2DM patients data in predicting a common complication of T2DM (retinopathy). Clustered column charts in Figure 2, compare the classification accuracy percentages average among 250 times randomly oversampling, for five steps of the enhanced Stepwise method. Additionally, error bars on the top of bar charts is illustrated. These results in Figure 2, reveal that classification accuracy of retinopathy in step 1 with discovering the first hidden variable has been increased sharply by adding a new hidden variable in step 2-4 and then dropped slightly at step 5, while error bar in step 1 is bigger than others. The error bar in step 2 is quite big due to a bigger confidence
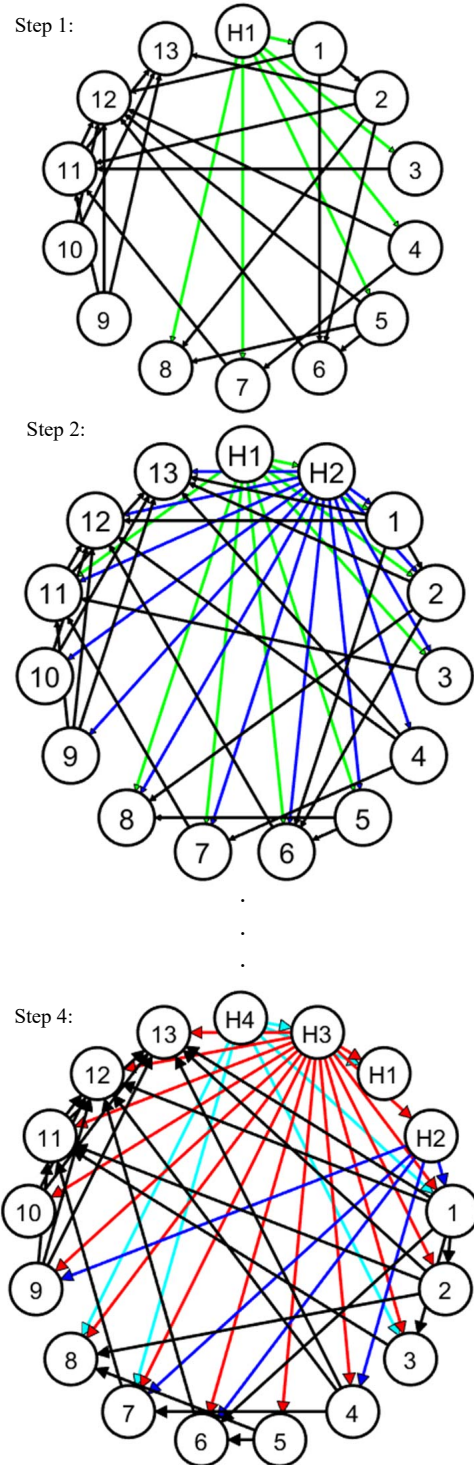


Fig. 1. DAG of static relationships among T2DM risk factors by applying Step 1, 2 and 4 of the Stepwise IC* approach. Conditional dependency for the hidden variable observed in the first, second and fourth steps are plotted in green, blue and cyan colors, respectively.

interval of others. Classification accuracy is a starting point for evaluating prediction results, which is the proportion of correct

TABLE II
COMPARISON OF OUR NEW AND ENHANCED STEPWISE IC* APPROACH
WITH ITS PREVIOUS VERSION IN [18].

| | Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|---|
| No Hidden variable | 0.40 | 0.50 | 0.40 | 0.40 |
| Stepwise in [18](Step1) | 0.60 | 0.40 | 0.80 | 0.70 |
| Enhanced Stepwise(Step1) | 0.82 | 0.30 | 1.00 | 0.99 |
| Stepwise in [18](Step2) | 0.80 | 1.00 | 0.60 | 0.70 |
| Enhanced Stepwise(Step2) | 0.97 | 0.82 | 0.98 | 0.88 |
| Stepwise in [18](Step3) | 0.80 | 1.00 | 0.60 | 0.70 |
| Enhanced Stepwise(Step3) | 0.97 | 0.83 | 0.98 | 0.84 |
| Enhanced Stepwise(Step4) | 0.98 | 0.83 | 0.99 | 0.94 |
| Enhanced Stepwise(Step5) | 0.97 | 0.84 | 0.99 | 0.87 |

predictions made divided by the total number of predictions made. Overall, we are sure more than 95 percent that accuracy in retinopathy prediction with leaning more than one hidden variable in our step-wise IC* approach is bigger than learning model with only one hidden variable.
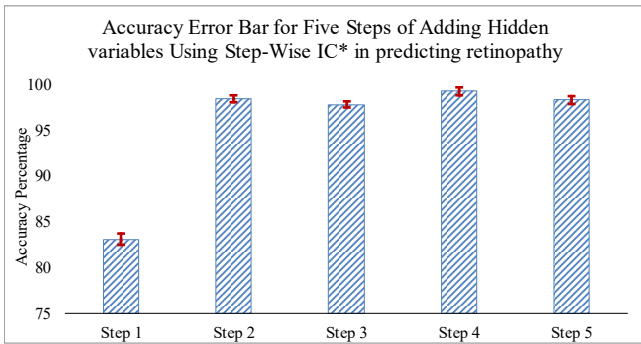


Fig. 2. An error bar is obtained for calculating confidence interval for average classification accuracy (for 250 times) of predicting retinopathy at 5 steps of the enhanced Stepwise IC* approach.

In Table II, we compared our enhanced Stepwise approach to the previous work in [18]. It is clear that the enhanced Step-wise method, have achieved a better performance measurements in predicting retinopathy. We now explore further the meaning of the hidden variables beyond their structure within the network. After the hidden variables are clustered for all patients using hierarchical clustering (distance metrics of DTW), we identify different Hidden Cluster 'Profiles'; dendogram of hierarchical clustering with temporal phenotypes for step 1 and step 2 of our approach is shown in Figure 3. For each sub-group of patient (each cluster), we provide the medoid pattern for the first and second learned hidden variables in Figure 3 (that is the hidden variable profile of the patient at the 'center' of the cluster [25]). The medoid of a cluster is a patient of that cluster whose average dissimilarity with the other patients of the cluster is the smallest. These can be thought of as 'deep phenotypes' that capture some hidden dynamics between the clinical factors. It can be seen in Figure 3 that these profiles capture quite different behaviors: one is fluctuating between the higher state and lower state of the hidden variable (Cluster 3 in step 1), involves a switch-
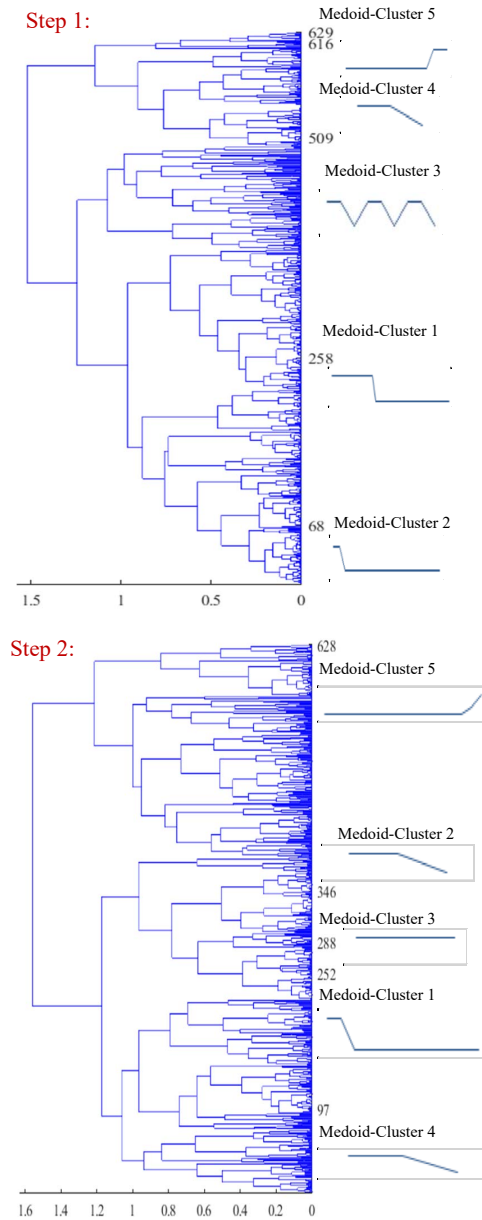


Fig. 3. Temporal phenotypes (The First Hidden Clusters "Profiles") in hierarchical clustering. Deprograms of Hierarchical clustering (complete) for the first and second hidden variable with the DTW distance metric. The x-axis represents is a measure of closeness of either individual data points or clusters, while y-axis is representing patient IDs as data points.

like behavior, one involves a general decreasing trend (cluster4 in all two steps), and another is flat-lining (Cluster 3 in step 2). If we look at the associated mean values for the clinical variables associated with each cluster, it is clear that the data has generated clearly separated cohorts of patients. Figure 4-C4 reveals interesting 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 in yellow (Cluster 1), whilst patients in Cluster 3 (with the flat-lining hidden cluster
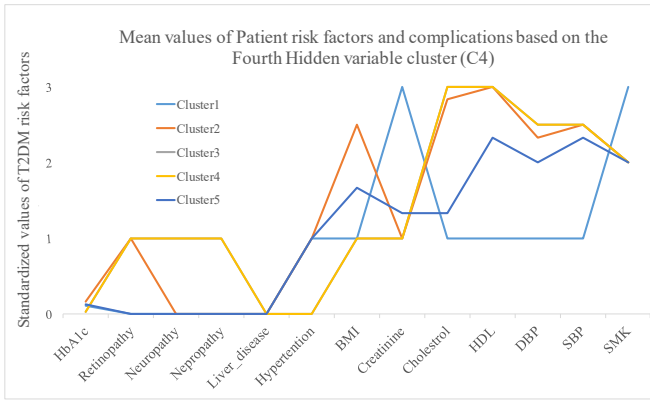
Fig. 4. Cluster Profile on mean values of patient risk factors and complications. Patients clustered using the fourth hidden variable obtained from the fourth step of the enhanced Stepwise IC* algorithm (C4).

profile) have generally low BMI and high HDL. Cluster 4, with the decreasing hidden cluster profile has much higher BMI values amongst the patients and very low HDL and DBP.

## IV. CONCLUSION AND FUTURE WORK

In this paper, we addressed three goals. Firstly, we identified targeted locations of hidden variables within clinical data from patients suffering DT2M using an enhanced version of stepwise IC* approach with more robust stopping points, which is able to learn more hidden factors. We then revealed how these hidden variables can improve prediction with applying confidence interval. Moreover, we clustered patients based upon the hidden variables and used the Medoid hidden variable profile of each cluster to characterize the "deep temporal phenotype" of that set of patients. We demonstrated how the discovery and understanding of hidden variables results in an improvement in the stratification of patients and aids in understanding interactions between risk factors and unmeasured variables. Although some attempts have been taken to find a meaningful temporal pattern, we believe this is the first use of DTW and Hierarchical clustering to stratify patients based on just one hidden variable, not other risk factors. We will extend this work by exploring how the discovered hidden variables interact amongst themselves and clinical variables by using inference techniques on different complications. We also would like to explore how the approach performs when integrating other related data such as environmental data, as well as on other longitudinal disease progression data. We will seek more advice from clinicians in interpreting hidden factors and their correlation toward other T2DM risk factors as well as disease prediction process.

## REFERENCES

[1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS medicine*, vol. 3, no. 11, p. e442, 2006.

[2] 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*, vol. 2012. American Medical Informatics Association, 2012, p. 606.

[3] S. E. Inzucchi and R. S. Sherwin, "The prevention of type 2 diabetes mellitus," *Endocrinology and Metabolism Clinics*, vol. 34, no. 1, pp. 199–219, 2005.

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

[5] 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.

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

[7] R. Daly, Q. Shen, and S. Aitken, "Learning bayesian networks: approaches and issues," *The knowledge engineering review*, vol. 26, no. 2, pp. 99–157, 2011.

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

[9] 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*, vol. 10, no. 1, pp. 19–26, 2016.

[10] S. Marini, E. Trifoglio, N. Barbarini, F. Sambo, A. 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*, vol. 57, pp. 369–376, 2015.

[11] L. Sacchi, A. Dagliati, D. Segagni, P. Leporati, L. Chiovato, and R. Bellazzi, "Improving risk-stratification of diabetes complications using temporal data mining," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 2131–2134.

[12] 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*, vol. 18, no. 3, pp. 1008–1014, 2014.

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

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

[15] 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*, vol. 10, no. 2, pp. 254–263, 2006.

[16] S. Mani and G. F. Cooper, "Causal discovery using a bayesian local causal discovery algorithm." in *Medinfo*, 2004, pp. 731–735.

[17] 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*, vol. 4, no. 3, pp. 181–192, 2008.

[18] 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.

[19] K. Murphy *et al.*, "The bayes net toolbox for matlab," *Computing science and statistics*, vol. 33, no. 2, pp. 1024–1034, 2001.

[20] L. Yousefi, L. Saachi, R. Bellazzi, and L. C. A. Tucker, "Predicting comorbidities using resampling and dynamic bayesian networks with latent variables."

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

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

[23] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[24] 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*, vol. 101, no. 11, pp. 3797–3802, 2004.

[25] A. Struyf, M. Hubert, P. Rousseeuw *et al.*, "Clustering in an object-oriented environment," *Journal of Statistical Software*, vol. 1, no. 4, pp. 1–30, 1997.