**Predicting on a latent variable in Dynamic Bayesian Networks for predicting the complications in patients with type 2 diabetes mellitus**

**Abstract** about 200 words

Type 2 diabetes mellitus (T2DM) is a rising public health concern worldwide. As may be expected, with a chronic disease that mainly affects adult individuals, T2DM is commonly complicated by its comorbidities (complications). In addition, T2DM patients are at increased danger of micro vascular comorbidities, such as, nephropathy, neuropathy, and retinopathy. Therefore, it is important to develop a prediction model based on data obtained from the follow up records of patients. With each visit, every patient faces a unique profile of symptoms and complications, regardless of the phase of the disease. However, Information on how to improve quality of life of patients through the utilization of predictive models is even restricted.

Models with the ability to interact with time series data are required to manage diabetic comorbidities. Consequently, challenges are being faced when mining time series in the prognosis of disease with rare positive values. Prior machine learning literature indicates that Dynamic Bayesian Networks (DBNs) are greatly enhanced and are capable to gather causal and temporal features of clinical records and likewise provide a predictive model that permits a detailed prediction of patient comorbidities. Furthermore, investigating latent variables is still a concern regarding learning DBNs. Although extensive research has been carried out on the prediction of diabetic data, no single study exists which has attempted to interpret the impact of the latent variables in the presence of unmeasured diabetic disorders.

Discovering Latent (Hidden) variables is an significant constituent in the medical diagnosis applications, and recreates a central part in improving learning DBN from clinical data. Understanding latent variables can contribute to a substantial advance in achieving a better example in application areas. On the other hand, much uncertainty still exists around the relationship between the comorbidities and latent variables. Hence, a systematic understanding of how latent variables contribute to the T2DM comorbidities is still missing.

The aim of this work is to show how to ameliorate the quality of life of T2DM patients by predicting the future phases of the diabetic complications, for different patients at their various medical visits. Specifically, we examine the function of the latent variable in the learning models from patient medical data. We demonstrate how statistical methods can help identify this hidden variable which is cut in the inference, because they are context-specific. We suggest that DBNs can improve our understanding of the true causal relationships hidden in the patient medical data. We aspire to determine the precise position of the latent variable and discover the dependencies between the latent variable and the observed variables. We employ a predictive model based on DBN to ease the decision-making process for doctors. A major strong suit of our approach, as opposed to other similar approaches in the literature, is the explicit representation of the kinship between the different risk factors and the footpaths. In summation, we went through a latent variable, which, along other risk factors influenced the results of our research and led to a better patient-specific interference.

Basically the key contribution is the combination IC\* to identify hidden variables with a DBN, as well as LinkStrength to assess the hidden variables. However, no research has found a reliable approach to locate a latent variable within DBNs.

**Introduction**

Diabetes is experienced as a silent murderer that is progressively made out as a serious, worldwide public health concern. Type 2 Diabetes Mellitus (T2DM) occurs because of impaired insulin secretion and/or opposition to insulin action, which is associated with severe long-term morbidity and large health maintenance costs to providers [ref]. As may be awaited with a chronic disease that mainly affects adult individuals, type 2 diabetes is commonly complicated by other medical conditions. Previous research has established that only 14% of patients with type 2 diabetes had no other comorbidities [1]. For instance, Hypertension is a major Cardiovascular disease (CVD) risk factors [1]. Up to 75% of adults with diabetes also have hypertension, and patients with hypertension alone often show evidence of insulin resistance. Moreover, in decision-making processes by physicians, it is necessary to capture the process of the disease and its common comorbidities. Therefore, the purpose of this study is to show how to improve the quality of life for such patients by anticipating the future stages of the diabetic complications for different patients at their various visits. Basically, our goal is to predict the onset of common T2DM comorbidities such as Microvascular Complications (Retinopathy, Neuropathy, Nephropathy), Hypertension and Liver disease. Specifically, we study the role of the latent variables in the learning models from the clinical data. We show how statistical methods can help us to discover these hidden variables. We suggest hidden variables are wrongly ignored in the inference, because they are context-specific. In summation, we show that contextual consideration can help us learn more about true causal relationships in the data.

At every visit, every diabetic patient has a unique profile of symptoms and complications that changes over time, regardless of the phase of the disease. Furthermore, these rapid changes at the time of different visits are having a severe effect on forecasting the solution. The non-stationary characteristics of the clinical data creates a difficult context for effectively forecasting time variations of data [2]. Moreover, considering how the state of the patient during each visit changes over time can be an important challenge for physicians preparing for future visits. Clinical data needs to be concerned as a time series data to provide a description of the progression of a disease over time [3]. Therefore, dealing with the time series patient records is known to be a major issue in the prognosis of comorbidities. Mining time series in the prognosis of disease with rare positive results is one of the challenging problems known to be the class imbalanced.

The issue of imbalanced data learning has received considerable critical attention in data mining [4]. There are several aspects that might influence the performance achieved by Bayesian network learning. It has been reported that one of these aspects is related to a class imbalance in which examples in the training data belonging to one class (here negative cases) heavily outnumber the examples in the other class [ref]. In this condition, which is ordinarily found in the clinical data describing an uncommon but important event, the learning system may have difficulties to learn the concept related to the minority class (the number of positive cases). The minority class represents that a patient suffers from the specific complications. In this situation, which is found in clinical data describing an infrequent but important event, the learning system may have difficulties in learning the concept related to the minority class (number of positive cases). To eliminate this class imbalance in the predictors, we use resampling and bootstrapping rare records. Furthermore, we sought a significant improvement of accuracy in prognosis of different complications. The bootstrapped helped us to quantify and deal with bias about the different patient comorbidities. In this way, our approaches to balance the data are different from previous researches.

we exploited Dynamic Bayesian Network (DBNs) which have more important advantages compared to other modelling tools and provided benefits of quality of numerical learnt from big data analytics.

In this paper, we briefly provide a model of the prognosis for the major comorbidities of patients diagnosed with diabetes. Furthermore, we aim to analyse the care received by patients with T2DM and specific comorbidities in the T2DM clinical data. The primary goal of this paper is to discover interesting relationship between latent variable and the clinical feature variables rather than just study learning of DBNs models. However, we propose the implementation of offline learning in Autoregressive Hidden Markov Models (HMM) in two-time series.

The outcome in the prediction tasks is based on the classification, a qualitative method, and investigating the presence of the latent variable in prognosis of the comorbidities. Only a few simulation models based on long term disease modeling have been carried out [5]. We have evaluated our model by measuring the Area Under the Curve (AUC), which is one of the best choices for evaluating the disease model [5]. Assessment of performance is based on the sensitivity and specificity. Meanwhile for the model, we tested the accuracy of prediction as a percentage of the correct prognosis of the specific comorbidities.

In the history of T2DM, development of common disorders has been characterised as a health care concern and they are the leading cause of death in the world. Therefore, the prediction of common disorders has been the focus of this paper. Risk factors found to be influencing T2DM have been explored in several studies [ref]. It has previously been observed that patients with type 2 diabetes mellitus (T2DM) are at increased risk of microvascular comorbidities including nephropathy, neuropathy, and retinopathy [11]. Predicting the comorbidities has long been a question of great interest in a wide range of medical fields. Learning the latent variable is a highlighted research on learning DBNs from clinical data. Latent variable discovery can lead to a better understanding of application domains. It can improve classification accuracy and boost user confidence in the classification models. Furthermore, understanding latent variables can lead to a significant improvement in reaching a better model in application domains [12]. Moreover, initial observations suggest that there may be a link between comorbidities and latent variables. In a study which set out to determine a hidden variable that interacts with observed variables and tried to locate the hidden in the right place in the Bayesian learning [13], the authors emphasized the importance of the presence of hidden variables. They highlighted the need to identify hidden variables. They noted that networks without hidden variables are clearly less useful because of the increased number of edges. Moreover, the problem would increase with limited amounts of data that cause the induced network to omit several dependencies in the model.

DBNs represent uncertainty in Artificial intelligence, in terms of effective inference, and efficient learning. Additionally, DBNs greatly benefit from being able to gather causal and temporal characteristics of clinical records to provide a prognostic model that provide a detailed prediction [6]. Previous works on learning DBNs have presented both network structures and parameters from the whole data set and learning parameters for a fixed network from incomplete data, in the presence of missing data or latent variables [4].

Previous researches have previously investigated the diagnosis of diabetes type II complications on the Pavia dataset [7] [8] [9]. A few recent studies presented a Bayesian method for extending the basic model to handle missing data and a latent variable [4]. Other researchers, however, used logistic regression and Naïve Bayes with diverse modelling strategies, different handling of unbalanced data, and different combinations of predictors [7] [8] [9]. Similarly, there have been a few investigations into diagnosing diabetes type I that have exploited Bayesian Networks [5]. Marini [5] simulated the health state and complications of diabetes type I patients by using partially and entirely Bayesian learned models. However, we use a different approach to the representation of the relationship between different comorbidities and the impact of the comorbidities on the final outcomes. This approach will be useful for stratifying patients according to their probability of developing complications in the near future. However, the major limitation of their work derives from time discretization in temporal time slices of one year. Most of the previously published studies in diabetes prediction have tended to focus on all patients as one integrated database rather than separating patients on different databases and using time series data per visit [Dagliati]. Dagliati’s paper in presenting a Hierarchical Bayesian Logistic Regression model to anticipate patients’s changes when the individual model parameters are estimated has some drawback. First, their parameters estimate used MCMC approaches, which is not suitable for large data sets. Moreover, time series model was not emplyed in the individual measurements. Considering these evidences, it seems that researchers have not analysed the impact of latent variables in the prediction of T2DM comorbidities in much detail. On the other hand, much uncertainty still exists about the relationship between unmeasured risk factors and latent variables.

Most of the Bayesian methods have not been designed to deal with such imbalanced data and just focused on improving accuracy or eliminating error rates. Liang dealt with the drawback of the over-sampling and under-sampling by exploiting the hybrid sampling technique to enhance bagging for imbalanced data [10].

Most studies in the field of structure learning have only focused on the fully learned structure or partially learned structure from previous literature [ref]. Moreover, they did not deal with discovering the latent variable. In Marini’s paper [ref], variables are connected within two-time series and within the same time slice. The assumption that the temporal dependencies are time invariant [ref] implies that the DBN is thoroughly described by exploring the edges and probability distribution between a time slice and its antecedent together with the probability distribution of the nodes at time t=0.

Until recently, there has been no reliable structure that shows the conditional dependencies between risk factors and diabetes comorbidities. In Marini [5] paper the health state and complications of diabetes type I patients were simulated by using partially and entirely Bayesian learned structured. For learning the network structures, they used a Tabu search, based on the hill climbing algorithm for BN.

Moreover, the unbalanced issue was fixed by sampling the data with stratification, according to the corresponding risk factors, and by excluding the test set before the learning step. With data mining methods, such as the Tabu search algorithm, they learned discretization and the network structure.

**Data processing and imputation:**

Data for this study were retrospectively collected from the rich longitudinal dataset. Samples were obtained with consent, from pre-diagnosed diabetic patients.

We categorised the risk factor using the model and prioritised features based on existing literature [ref]. For example, for the estimation of the Retinopathy, three types of Retinopathy disorder were mixed. In summation, for evaluating the specific analysis, macro vascular comorbidities (coronary artery disease, peripheral arterial disease, and stroke) were mined from the other diabetic disorders. The experiments were extended out over the various micro vascular comorbidities (diabetic nephropathy, neuropathy, and retinopathy), liver disease, and high blood pressure.

Anticipation of the common disorders such as Hypertension, Macro vascular disorder and Liver disease is the focus of this report. To study diabetes, a random sample of patients with different comorbidities was recruited from Pavia clinical data in Italy. Diabetes health status records were accumulated from 356 patients. Additionally, we performed physical examinations and gathered data (including BMI, SBP, DBP and laboratory data, including HbA1c) as our observed data from the FSM hospital in Italy.

We take an illustration as a T2DM patient in a single visit in the dataset. The total number of pre-diagnosed T2DM patients in our rich dataset is 3959 instances. We used data from the same clinical information in our former report [ref]. We stratified the patient visits based on their risk factors. We ignored patient visits missing data related to the risk factors. We split the dataset into a train set and a test set with the same number of patients in the sets.

Foremost, we discretized risk factors to be applied in the discrete space DBNs.

Having time series in diabetes data, we characterized longitudinal data with variables which are changing over time. Various studies on longitudinal datasets suggest an association between comorbidities and symptoms of the disease. We have retrieved evidence from longitudinal studies. And so we exploited them to figure how the fluctuations in the latent variables are moved by various comorbidities. Moreover, we anticipated the future condition of each patient during each future visit by using a set of observed trials. Each disease has an independent probability of causing each symptom. Thus, this method is especially appropriate for studying the microvascular complications (retinopathy, neuropathy, and kidney disease). In addition, our model represents some risk factors, which consist of Body Mass Index (BMI), Total Cholesterol, HDL, Triglycerides, Creatinine, Systolic Blood Pressure (SBP), Smoking habit. These risk factors have been proven as associated variables over time with diabetic complications [11][14]. [13]

We used a prediction model to ease the decision-making process for physicians. Injection solutions with our customised time series bootstrapping for the model were coded to reduce experimenter bias.

**Method**

In this paper, we used discrete space/discrete time DBNs and data mining techniques on a rich longitudinal database to analysis the latent variable. Discovering the latent (hidden) variables in DBNs is an important component in the medical diagnostic applications, and plays a key role in learning DBNs from data. In this section, we present a novel methodology to analysis the hidden variable whithin the DBN. We aim to find the likely relationship with the latent variable and T2DM risk factors and discover the dependencies between the latent variable and the observed variables.

Given the symptoms, the network could be used to compute the probabilities of the presence of various comorbidities. DBNs have advantages over other prediction models based on clinical data, such as the Markov models, that suffer from low performance by having high numbers of risk factors. In addition, risk prediction equations were not capable of performing predictions of feature visits. DBNs could represent the probabilistic relationships between comorbidities and symptoms. DBNs were used to show the Markov models in patient states under uncertainties. Therefore, we provided structured modelling techniques for understanding the latent variable. Moreover, we designed a probabilistic model, which represented the integration of the comorbidities and various findings from diabetic data. We infered DBNs with dynamic connection between the variables using the REVEAL algorithm [ref]. A definite hidden variable is employed for qualitatively different form of appearance of the entire set of risk factors. Therefore, we applied the method to specific diabetic risk factors for a better understanding of the relationship between different comorbidities and findings.

Each risk factor with a dynamic nature has qualitative states, e.g., high, medium, and low. Nevertheless, the comorbidities have binary states which are enormously imbalanced. In fact, the problem of class imbalance seems to be related to learning with too few minority class of positive cases in the presence of other complicating factors, such as class overlapping [4]. Thus, we exploited bootstrapping approach to regenerate our observed Time series (TS) visits per patient. We used this method for inference by re-sampling and concatenating different pairs of visits. Due to T2DM literature, once a patient has suffered from a given comorbidity was identified, we estimated the probabilities that the the patient would keep developing the comorbidity (we called this method as a pre-balancing approach). Following this update of the comorbidities data in the following visits, unbalanced data were reduced notably. Prior to training the data and undertaking the investigation, the TS bootstrapped approach was applied to the training set to remove the bias from the comorbidities. After balancing the data with the TS bootstrapping method, we train the data. The Time series data were analysed by DBNs learning models and balanced with a combination of pre-balancing and TS bootstrapping approaches, which was called pair re-sampling approach.

To implement our model, we made a few assumptions. Firstly, we used probabilistic directed model Hidden Markov Model (HMM), which is a simple type of DBNs with discrete latent variables [15]. Secondly, we tried to split out different complications at each visit for each patient into binary variables. Thirdly, the process of model specification and parameter estimation was carried out at the completion of the last visit of each patient. One of the most interesting observations is the variation of the specific complications over time for the patients in their health condition. Nevertheless, we assumed that temporal dependencies are time invariant. Then, for simplicity, we applied data reduction methods such as discretisation on T2DM risk factors and features. Then we used the variables as input arguments of the learning structure with the K2 and REVEAL algorithm [16] for creating the Intra links and the Inter links, respectively. Furthermore, we assumed that patient status at time t + 1 depends on the corresponding hidden variable at a previous time (t). Then, we added the hidden variable in an HMM to monitor the correlation of risk factors over time. On arrival at the clinic, patients were taken into different clinical tests. Moreover, the EM algorithm was set to run no more than 30 iterations. Risk factors and outcome variables were represented in two-time series DBN. Consequently, the outcome variables of our study were the prediction of nephropathy, hypertension, liver disease, retinopathy, and neuropathy comorbidities. Finally, prediction results for the common comorbidities were compared with the results obtained from the preliminary analysis of T2DM. data mining and analysis were performed using MATLAB and Bayes Net Toolbox [16]. For virtualization and drawing the graph, we relied on Graphviz toolbox [ref].

**Locating the Latent Variable strategy**

In this section, we illustrate the proposed technique to locate the latent variable within the DBN and find the best structure for the risk factors to predict the next stage of disease that each patient will experience. In our method, we exploited the expertise of practitioners in the diagnosis [ref]. It is essential to define how the latent variable and comorbidities are expected to affect the risk factors. Understanding the latent variables is a continuing concern within learning DBNs [13]. Thus, the key contribution in our work was exploiting the combination the IC\* algorithm [ref] to identify hidden variables with a DBN, as well as LinkStrength [ref] (LS) to assess the hidden variables. To this goal, providing prediction methods and sensitivity analysis to understand the unmeasured risk factors (act as the latent variables) of clinical data would be necessary.

Therefore, we employed local and global kinds of sensitivity analysis [ref], which consists of Mutual Information (MI), and Link Strength, respectively. There are two types of LS which result when measuring uncertainties, including True Average Link Strength (LSTA), and Blind Average Link Strength (LSBA). The MI between node X and node Y is a number of uncertainties in Y that we decrease by knowing the state of X (and vice versa). MI is distinct between any sets of nodes. However, if there are multiple edges, it might be that one which carries the most impact. Moreover, we utilised the percentage points of uncertainty reduction in Y by knowing the state of X, if the states of all other parent variables are known (averaged over the parent states using their actual joint probability (for LSTA) or assuming all parents are independent of each other and uniformly distributed (for LSBA)). An LS measure tolerates us to observe the specific impact of each edge. After conformational analysis of uncertainties with LS, it was necessary to evaluate the learned structure. Therefore, we replaced the Intra links in the LS with a high probability corresponding links obtained from the IC\* algorithm.

The IC\* algorithm is similar to the PC algorithm (Spirtes, Glymour, and Scheines 1993), which calculates several conditional independence tests, and combines these restraints into a Directed Acyclic Graph (DAG) to characterise the entire [Markov equivalence class](http://www.cs.ubc.ca/~murphyk/Software/BNT/usage_02nov13.html" \l "markov_equiv), except that they can detect the presence of latent variables. The IC\* algorithm learns a latent structure associated with a set of observed variables. The latent structure revealed is the projection in which every latent variable is either a root node or linked to exactly two observed variables.

The latent variable, in the projection, is represented using a bidirectional graph. Table 1 presents the adjacent matrix obtained from the IC\* algorithm with the preliminary analysis and the intercorrelation among the 13 features of diabetic data. The correlation of the dataset was used to ascertain the structure with the IC\* algorithm. The outcome of the IC\* correlation graph is summarized in figure 1. There are two different possibilities for finding the location of the latent variables using the IC\* algorithm. To distinguish between these two possibilities, the IC\* graph was illustrated in Figure 1, with two different colours, e.g., green was used if there was either a latent variable or a directed edge between the corresponding nodes while the location of the latent variable was identified by the red links. Then we comparedthe probabilities which we obtained from the Link Strength algorithm to assure that the best structure was provided by using the IC\* algorithm.

Table 1: IC\* adjacent matrix.



A screenshot of a cell phone

Description generated with high confidence

Figure 1: ICLS graph obtained from IC\* adjacent matrix with

the higher percentages of uncertainty reduction from LS approach.

Selecting the LSTA and LSBA percentages in the specific edges (that were chosen from those edges, which had been discovered previously in the IC\* adjacency matrix), we can find the possibility of the existing edges and location of the latent variable. For example, from table 1 you are assured that there is a latent variable between two variables (Retinopathy and DBP). Meanwhile, from figure 1, you can see the probability of the appearance of a latent variable between the corresponding variables is about 11% and 96.50% for LSTA and LSBA, respectively. The LSTA and LSBA percentages of appearance the latent variable are highlighted in red.

Comparison methods between the following approaches are desirable to indicate the validity of our approach:

1. Using One latent variable and the structure was fully learned from the K2 algorithm and the REVEAL algorithm on the imbalanced data. (Auto-regressive HMM that the latent variable was pointed to all the observed data.)
2. Using One latent variable and the structure was fully learned from the K2 algorithm and the REVEAL algorithm on the balanced data. (Auto-regressive HMM that the latent variable was pointed to all the observed data.)
3. Without using any latent variable, but the structure was obtained from ICLS approach.
4. Using One latent variable and the structure obtained from IC\* algorithm.
5. Using One latent variable and the structure obtained from ICLS approach.

We represented the result of this comaprison in the next section.

**Experimental results**

In this section, our convincing experimental results shows three discrete reasons of the contributions provided in this paper. First, using the latent variable in the prediction and the ICLS method to improve the classification trend. Then, we show that using our pair-balancing approach for removing the bias from the data causes a huge improvement in the classification accuracy. Finally, the correlation between the latent variable and the risk factors increases over time.

To this end, we compare the effects of learning DBNs in T2DM patient comorbidities in the latent variable presence or absence. Thus, to differentiate between these two possibilities, a simple statistical analysis based on the ROC (Recursive Operator Curve) was applied to interpret the improvement in the prediction results by adding a latent variable to the observed risk factors. In summation, we used an Auto-Regressive model to show the latent variable pointing to the entire set of risk factors within the same time slice. The latent variable in slice t+1 depended on the latent variable in slice t. The total risk factor set followed the structure obtained from the K2 algorithm and, for dependencies between two slices, the REVEAL algorithm. This was our basic structure and then, by interpreting this, which obtained from the ICLS approach, a huge improvement in the prediction model was developed.

For proving the importance of presenting a latent variable within a prediction model such as the DBNs, we compare the ROC curves as well as the confusion matrices of the prediction model. Additionally, to see that our approach (ICLS) in predicting the comorbidities is more precise than simply using the IC\* algorithm itself, we revealed the classification accuracy as well as sensitivity by using ROC and confusion matrix and AUC (Area Under the Curve).

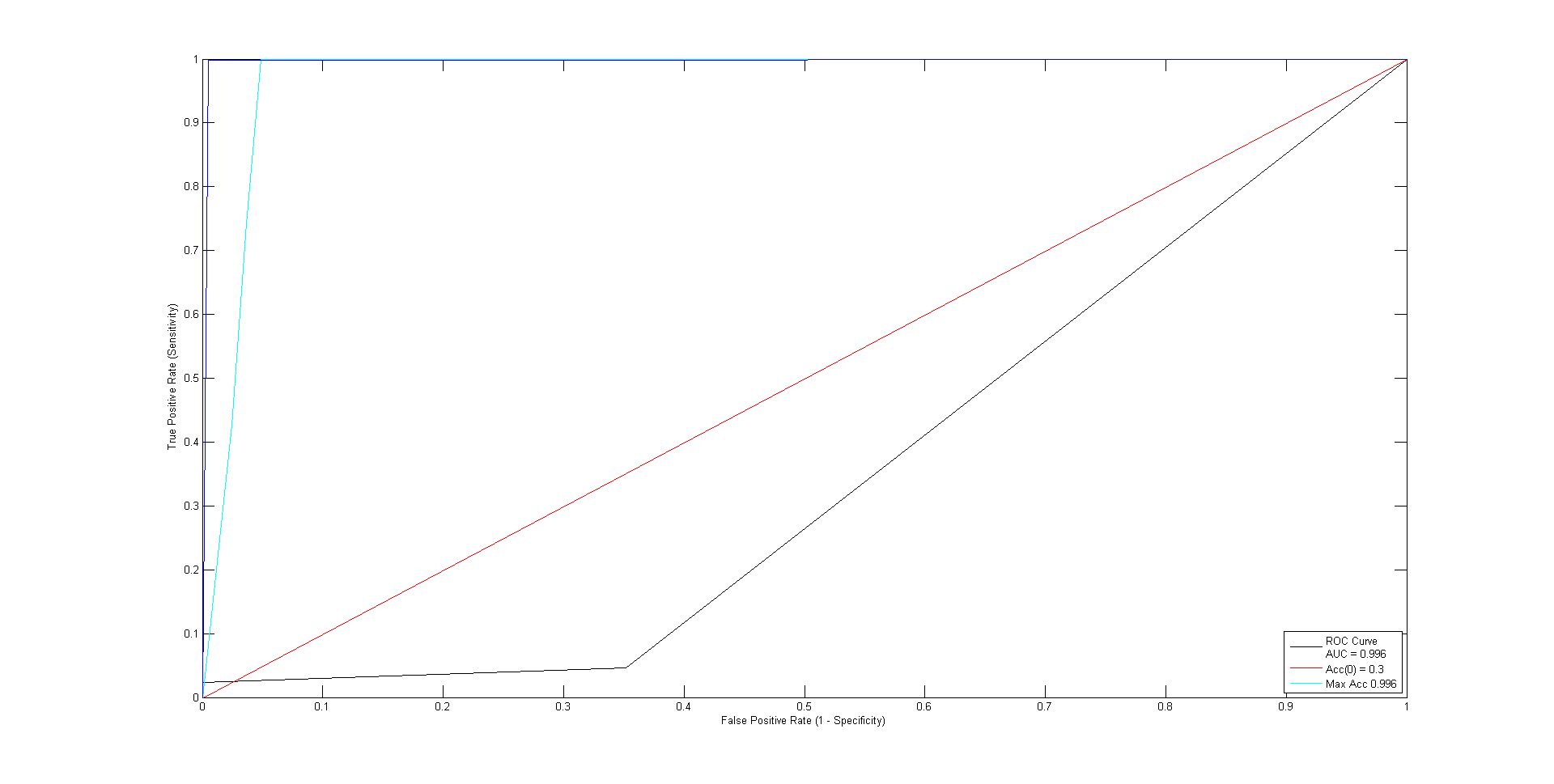
AUC is a measure of overall accuracy, that a classifier would rank a randomly chosen true instance (e.g. patient) higher than a random false one (e.g. control subject). In Figure 1, you can see a significant increase in the classification accuracy and AUC in the new method by improving the confidence in the existing links between variables by using the ICLS approach. Due to this, a reliable structure for representing the dependencies between the latent variable and the diabetic features is provided.

Table 1 Comparison characteristics of the different approaches imbalanced versus balanced versus one latent variable versus IC\* versus no latent variable versus ICLS.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Experiment 1 | Experiment 2 | Experiment 3 | Experiment 4 | Experiment 5 |
| Imbalanced data | Yes | No | Yes | Yes | Yes |
| Structure Learning Algorithmfor Intra and Inter links/latent variable structure | K2 and Reveal/ Auto-Regressive | K2 and Reveal/ Auto-Regressive | ICLS/ no latent variable | IC\*/IC\* | ICLS/ ICLS |
| Number of Latent variables | 1 | 1 | 0 | 1 | 1 |
| AUC |  |  |  |  |  |

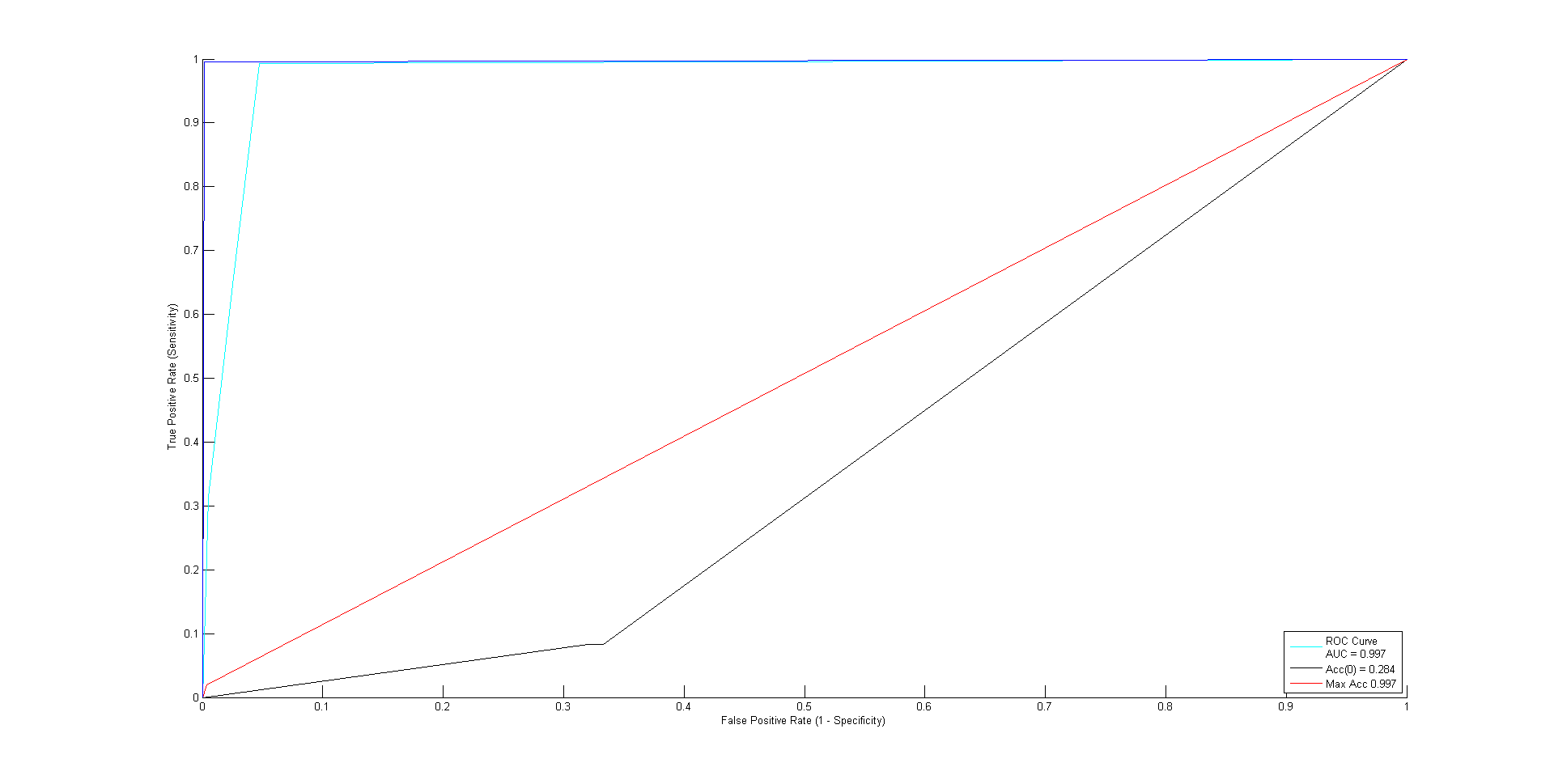
The classification accuracy, as well as sensitivity from the above list, confirmed that there is an increasing trend from top to bottom of the list. Our comparative experiments show that, in general, our approach in pre-balancing and then TS bootstrapping (pair re-sampling), which we have exploited in this step, provides more accurate results than unbalanced model, considering the area under the receiver operating characteristic (ROC) curve as well as the AUC (Area Under Curve) [15] [5].

Figure 1-3 reveals a clear trend of increased classification accuracy for the prediction of comorbidities from a latent variable using ICLS for locating the latent variable between the risk factors (coloured by blue). Moreover, there has been a notable rise in sensitivity (see cyan curve) for the results obtained from the balanced model using the K2 and the REVEAL algorithms (experiment 2 versus expriment 1 in table 2). In contrast without a hidden variable, sensitivity (see red curve) of results have dropped sharply over time (experiment 2 versus expriment 3 in table 2). As we can see in figures 2-3, the top half of the curve shows the sharp improvement of sensitivity (using bootstrapping), whereas the bottom half of the curve (unbalanced data) shows a steady decline in sensitivity (black curve).



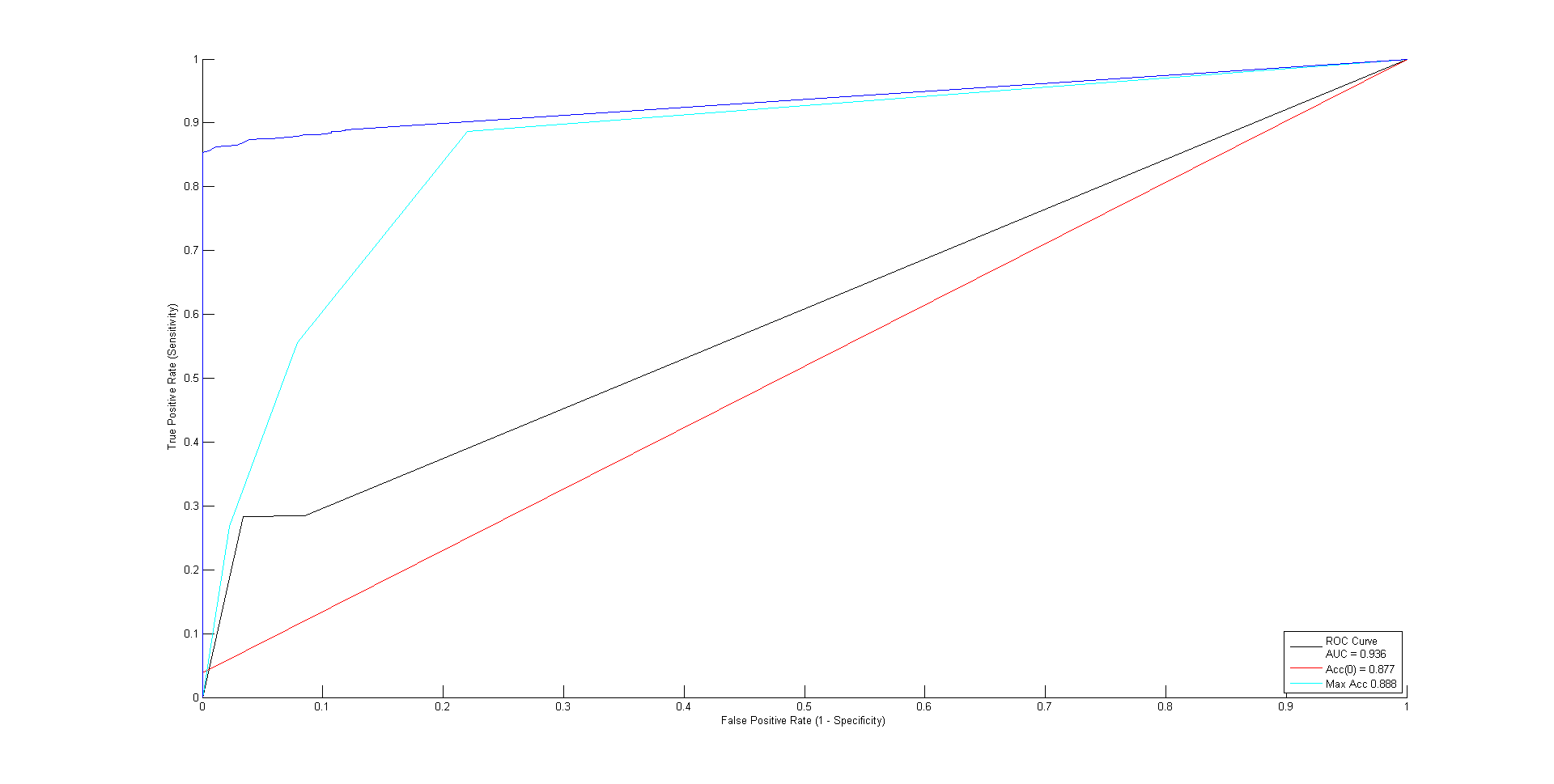
**Figure 2: Retinopathy ROC curves in blue, red, cyan, black shows the ROC**

**curves with ICLS, without Latent, with K2 and unbalanced, respectively.**



**Figure 3: Liver disease ROC curves in blue, red, cyan, black show the ROC**

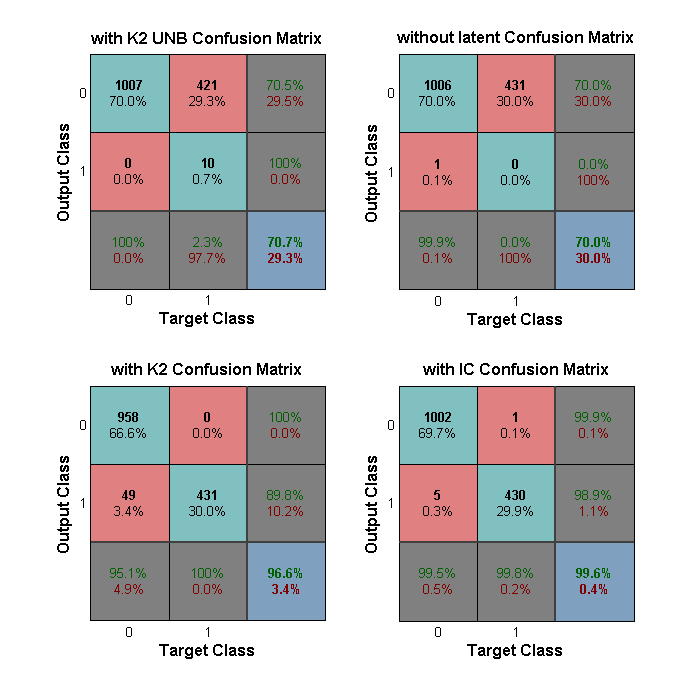
**curves with ICLS, without Latent, with K2 and unbalanced, respectively.**



**Figure 4: Hypertension ROC curves in blue, red, cyan, black show the ROC**

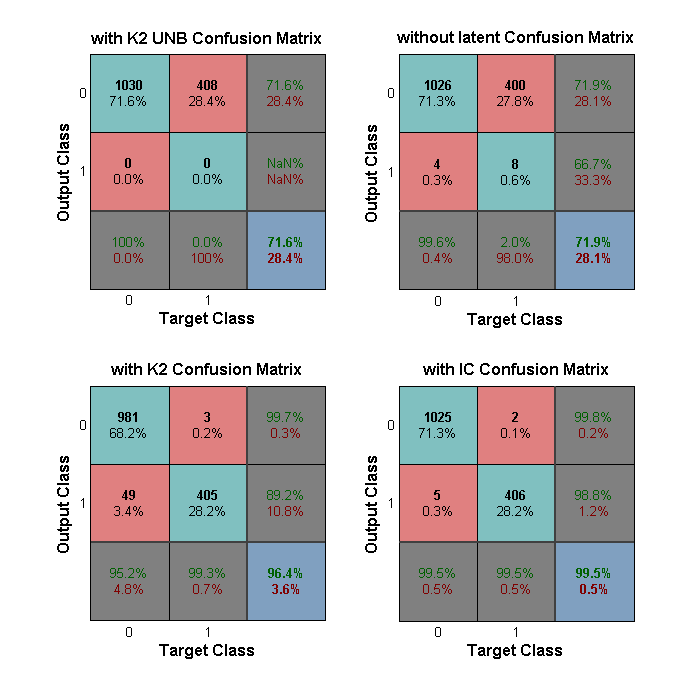
**curves with ICLS, without Latent, with K2 and unbalanced, respectively.**

As can see in confusion matrices in figures 7-9, our approach achieved perfect sensitivity and specificity when we use the ICLS method. On the other hand, the model accuracy changes notabely while the confidence of locating the latent variable within the balanced data grows.



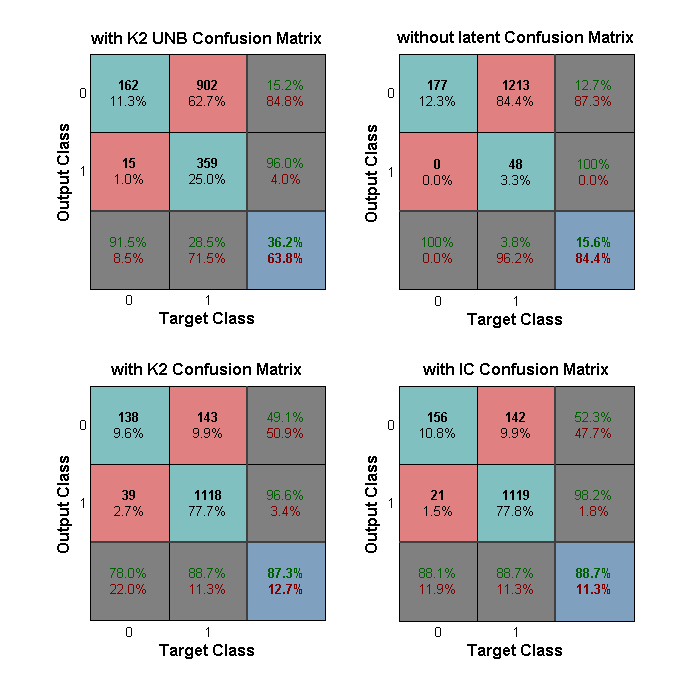
**Figure 8 Retinopathy Confusion Matrices comparing experiment 1 (top left),**

**experiment 2 (bottom left), experiment 3 (top right), experiment 5 (bottom right).**



**Figure 6 Liver disease Confusion Matrices comparing experiment 1 (top left),**

**experiment 2 (bottom left), experiment 3 (top right), experiment 5 (bottom right).**



**Figure 10 Hypertension Confusion Matrices comparing experiment 1 (top left),**

**experiment 2 (bottom left), experiment 3 (top right), experiment 5 (bottom right).**

The statistical analysis was used to represent the late, early, and same time prediction of comorbidities comparing the experiment 2 and the experiment 5. The results of multi class confusion matrices for early, same time and late prediction of the first four visits of T2DM patients are summarized in Table 1-3. In this experiment, we looked at the conditional likelihood and with respect to it, the resulting models have a very high prediction accuracy. This illustrates the fact that earlier prediction of the comorbidities necessarily implies better classification results.

Table 2: Retinopathy Multi Class Confusion Matrix (green, amber, red show the early, same time and late prediction of the first four visits of T2DM patients. The percentage (bottom left) of the Confusion Matrix represents the classification accuracy.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Retinopathy confusion matrix (without using ICLS) | | | | | | |  | Retinopathy confusion matrix (with using ICLS) | | | | | | |
| Predicted | | 0000 | 0011 | 0111 | 1111 | 0001 |  | Predicted | | 0000 | 0011 | 0111 | 1111 | 0001 |
| Real |  | 1 | 2 | 3 | 4 | 5 |  | Real |  | 1 | 2 | 3 | 4 | 5 |
| 0000 | 1 | 68 | 0 | 0 | 0 | 0 |  | 0000 | 1 | 73 | 1 | 0 | 0 | 0 |
| 0011 | 2 | 2 | 1 | 1 | 0 | 0 |  | 0011 | 2 | 1 | 1 | 0 | 0 | 0 |
| 0111 | 3 | 1 | 0 | 2 | 0 | 0 |  | 0111 | 3 | 0 | 0 | 3 | 0 | 0 |
| 1111 | 4 | 2 | 0 | 2 | 21 | 0 |  | 1111 | 4 | 0 | 0 | 2 | 21 | 0 |
| 0001 | 5 | 3 | 0 | 0 | 0 | 0 |  | 0001 | 5 | 1 | 0 | 0 | 0 | 0 |
| 90% |  |  |  |  |  |  |  | 95% |  |  |  |  |  |  |
| 8% |  |  |  |  |  |  |  | 1% |  |  |  |  |  |  |
| 2% |  |  |  |  |  |  |  | 2% |  |  |  |  |  |  |

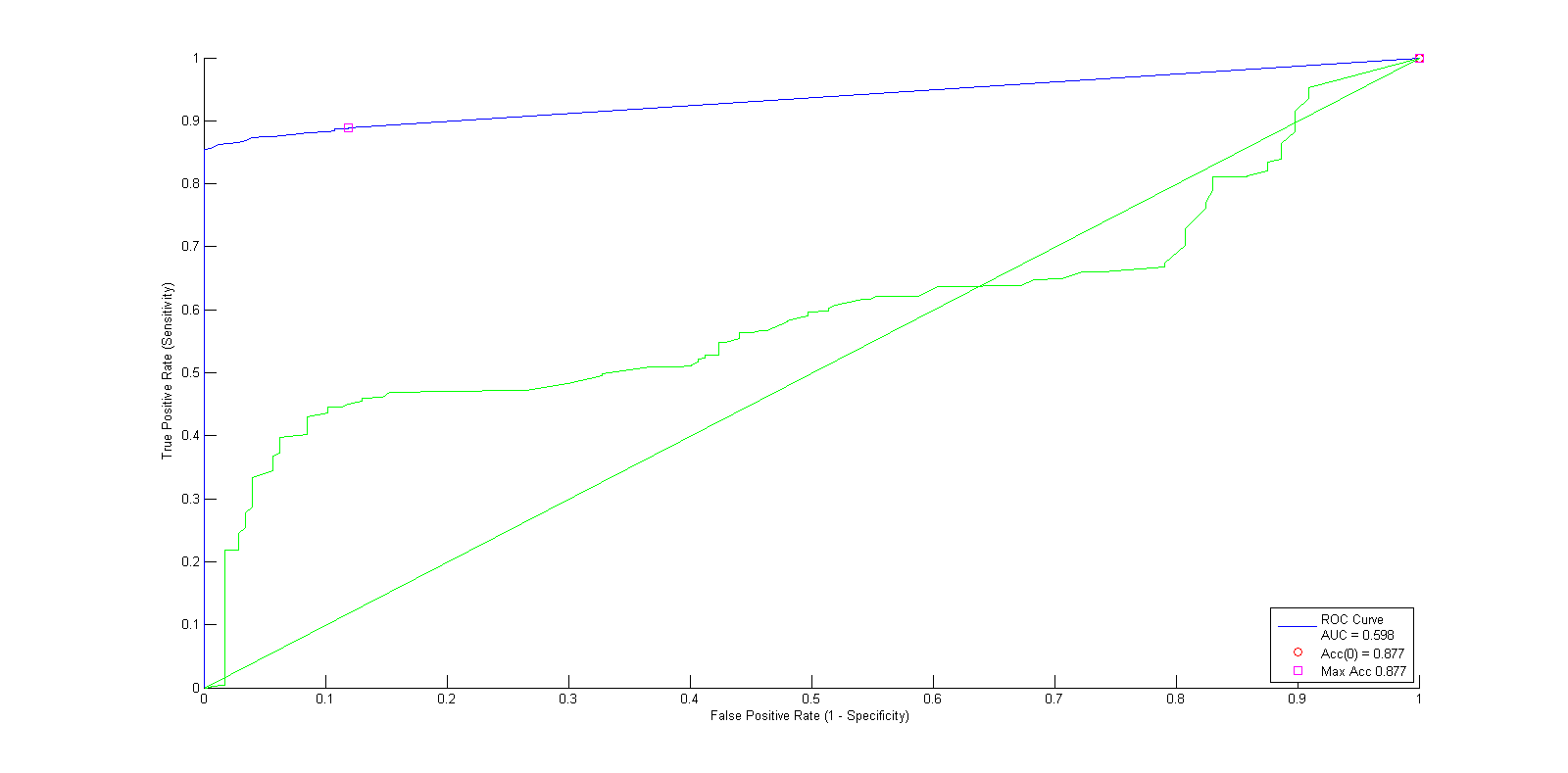
Table 3: Liver disease Multi Class Confusion Matrix (green, amber, red show the early, same time and late prediction of the first four visits of T2DM patients. The percentage (bottom left) of the Confusion Matrix represents the classification accuracy.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Liver confusion matrix (without using ICLS) | | | | | | |  | Liver confusion matrix (using ICLS) | | | | | | |
| Predicted | | 0000 | 0011 | 0111 | 1111 | 0001 |  | Predicted | | 0000 | 0011 | 0111 | 1111 | 0001 |
| Real | | 1 | 2 | 3 | 4 | 5 |  | Real | | 1 | 2 | 3 | 4 | 5 |
| 0000 | 1 | 75 | 0 | 0 | 0 | 0 |  | 0000 | 1 | 81 | 0 | 0 | 0 | 0 |
| 0011 | 2 | 0 | 0 | 0 | 0 | 0 |  | 0011 | 2 | 0 | 0 | 0 | 0 | 0 |
| 0111 | 3 | 6 | 0 | 2 | 1 | 0 |  | 0111 | 3 | 0 | 0 | 2 | 1 | 0 |
| 1111 | 4 | 1 | 0 | 0 | 16 | 0 |  | 1111 | 4 | 1 | 0 | 0 | 16 | 0 |
| 0001 | 5 | 1 | 1 | 0 | 0 | 0 |  | 0001 | 5 | 1 | 1 | 0 | 0 | 0 |
| 92% |  |  |  |  |  |  |  | 97% |  |  |  |  |  |  |
| 8% |  |  |  |  |  |  |  | 2% |  |  |  |  |  |  |
| 0% |  |  |  |  |  |  |  | 0% |  |  |  |  |  |  |

Table 3: Hypertension Multi Class Confusion Matrix (green, amber, red show the early, same time and late prediction of the first four visits of T2DM patients. The percentage (bottom left) of the Confusion Matrix represents the classification accuracy.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hypertension confusion matrix (without using ICLS) | | | | | | |  | Hypertension confusion matrix (using ICLS) | | | | | | |
| Predicted | | 0000 | 0011 | 0111 | 1111 | 0001 |  | Predicted | | 0000 | 0011 | 0111 | 1111 | 0001 |
| Real | | 1 | 2 | 3 | 4 | 5 |  | Real | | 1 | 2 | 3 | 4 | 5 |
| 0000 | 1 | 11 | 0 | 0 | 1 | 0 |  | 0000 | 1 | 12 | 0 | 1 | 2 | 0 |
| 0011 | 2 | 1 | 0 | 1 | 1 | 0 |  | 0011 | 2 | 1 | 0 | 0 | 3 | 0 |
| 0111 | 3 | 0 | 0 | 0 | 0 | 0 |  | 0111 | 3 | 0 | 0 | 2 | 14 | 0 |
| 1111 | 4 | 2 | 0 | 2 | 82 | 0 |  | 1111 | 4 | 2 | 0 | 0 | 65 | 0 |
| 0001 | 5 | 1 | 0 | 0 | 1 | 0 |  | 0001 | 5 | 0 | 0 | 0 | 1 | 0 |
| 93% |  |  |  |  |  |  |  | 94% |  |  |  |  |  |  |
| 5% |  |  |  |  |  |  |  | 3% |  |  |  |  |  |  |
| 2% |  |  |  |  |  |  |  | 0% |  |  |  |  |  |  |

Comparison of Intra links obtained only from the IC\* algorithm alone and links in the adjacent matrix with higher probabilities in the LS algorithm (ICLS) found significant improvement. (Figures 3-4).



**Figure 2: Comparison of Hypertension ROC curve using ICLS (blue curve) and using IC\* (green curve).**

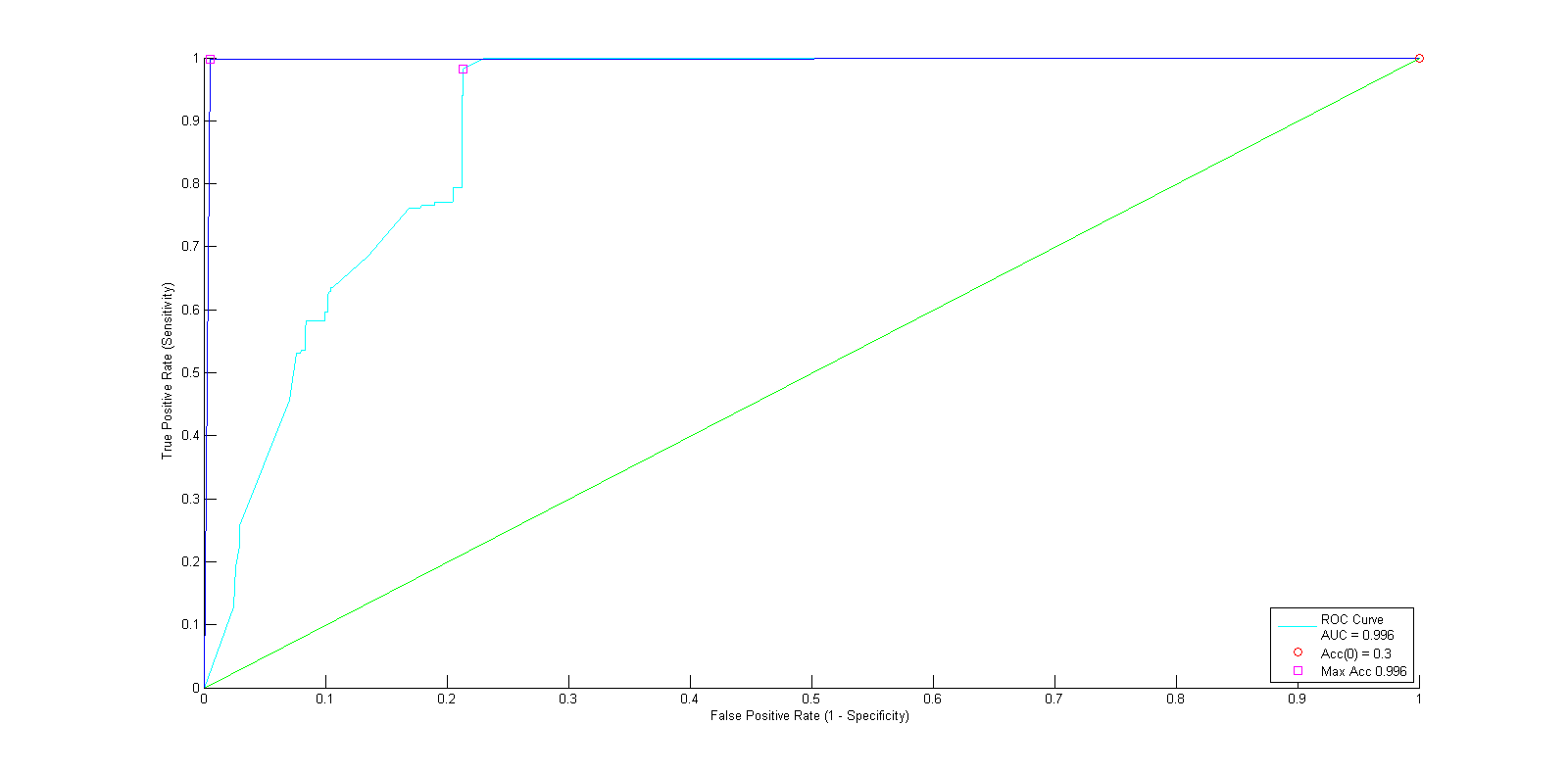
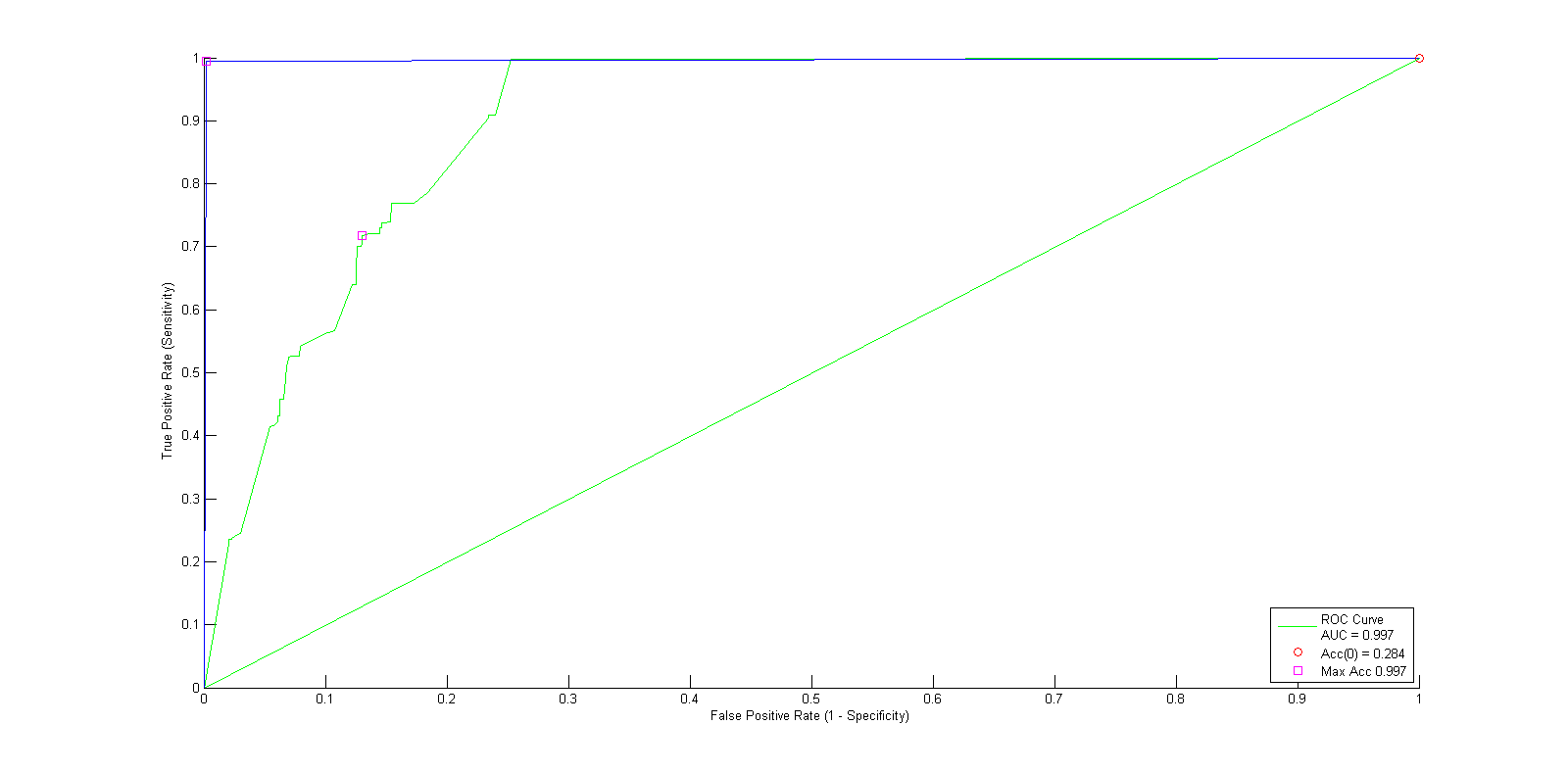


Figure 3: Comparison of Retinopathy ROC curve using ICLS (blue curve) and using IC\* (green curve).



**Figure 4: Comparison of Liver disease ROC curve using ICLS (blue curve) and using IC\* (green curve).**

The sensitivity analysis was performed on the cohort of pre-diagnosed T2DM over 15 years. Therefore, the states percentages of corrected prediction, were computed and estimated almost 90% for the common comorbidities. Moreover, mapping the latent variable to the associated risk factors is at the heart of our understanding of which disease risk factors and comorbidities will be influenced by the latent variable. On the other hand, the outcome of latent variables may represent different types of predictors, such as life expectancy, quality of life, or the spread of specific risk factors or comorbidities. Further analysis shows how a latent variable is connected strongly to the comorbidities variables in figures 10-20. Taken together, these results suggest that there is an association between latent variables and common complications in the prognosis of the diabetic patients.

**Conclusion**

Clinicians may need to resolve the future care based on their knowledge and previous experience. Therefore, there is a time consuming and unreliable decision which may cost a great deal. The reason caused from either the importance of unmeasured risk factors in the dataset or different response to the disease by different patients. Due to unmeasured risk factors and missing information in diabetes dataset, there was not any sufficient possess for a perfect prediction. Thus, we utilized the latent variable in DBNs to support the rest of observed variables and fill the vacant space of unmeasured variables in the dataset. Here we aimed to offer an explicit explanation to decision makers of the potential result of the latent variable.

Skewed class distribution causes a misreading of common assessment as it leads to a biased classification. One of the features that might affect the performance achieved by Bayesian network learning is related to a class imbalance. In this situation, the training data fit into one class (negative cases) which severely outnumber the examples in the positive class. This study analysed clinical data which will help to address the above research gaps.

Our goal is to decrease the cost of illness and improve the quality of life of T2DM patients, which suffer from the disease over a long period of time. One purpose of this study was to assess the extent to which the latent variable predicted T2DM comorbidities. The single most striking observation to emerge from the data comparison was the improvement of accuracy and sensitivity by adding latent variables. The clear benefit of anticipating latent variable in the prediction of comorbidities could be identified in this analysis.

In this paper, we predicted whether the specific comorbidities would have developed by subsequent patient visits or not. Moreover, we compared the effect of applying bootstrapped algorithms and creating the balanced dataset. We observed considerable differences between learning with latent variables and without employing them. It was obvious that by using our model, the accuracy increased significantly. In summation, our methods for learning and bootstrapping developed good predictive models and an excellent balance between genuine and false positive rates for predicting the occurrence of comorbidities in the different health stage of patients for the T2DM data.

We implemented the DBN structure using the REVEAL algorithm [ref], and the K2 algorithm [ref], to learn the dynamic links, and the static links, respectively. Then we proceeded by finding the location of the latent variable and its relationship to the other observed nodes from the IC\* algorithm. In summation, we induced a correlation matrix for 13 risk factors, derived from 3959 samples. We exploited the IC\* matrix to learn a causal model of the diabetic data. Later we computed the probabilities of links by using the Link Strength (LS) algorithm to guarantee that we had chosen the correct set of related links and locations of the latent variable. Consequently, we have updated the final structure by obtaining the latent variable from the IC\* algorithm adjacent matrix and assessing the uncertainties by using the LS method. Thus, we modelled the conditional dependencies using IC\* integrated with the Link Strength (we called it ICLS) approach to depict the exact links and dependencies between T2DM variables.

We determined a probabilistic simulation of the development of T2DM and its complications over time. Firstly, we imputed variables to each diabetic risk factor. Secondly, we determined a qualitative definition for them. Thirdly, we specify the state space for variables. Furthermore, we recognised basic predictors and their relationships with eachother. In addition, we estimated the risk factors using the model and prioritise features based on existing literature. Then, we compared the sensitivity and specificity of the learned model with the balanced model in terms of the structure and marginal distributions over the latent variable and the risk factors. Consequencely, explained how the latent variable affected the prediction result. As can be seen in the experimental result section, our approach is forecasting the presence and development of the most common comorbidities that each patient will experience during her visits.

Overall, the statistical result is quite revealing in several ways. Frist, the most striking result to emerge from the data shows a significant correlation between the diabetic comorbidities and risk factors. Then, it enhances our knowledge in understanding the dependencies between various risk factors and comorbidities. More significantly, it gives away the position of the latent variable in the learned structure. Interestingly, on that point are also differences in the LSTA and LSBA percentages, which symbolizes the importance of identifying the parents and their dependencies with the kids.

Our primary goal is to focus the importance of an understanding of latent variable in the clinical data with unmeasured attributes. In conditions of uncertainty reduction, when some aspects of the model are not exactly known, we utilized DBNs with help from other methods (IC\* and LS). We apply this appraisal method to fulfil the importance of exploiting latent variable in a clinical model. In fact, the prediction accuracy in our approach was significantly improved. However, the current research was not specifically designed to evaluate continuous factors as well as time-variant model. A further study should concentrate on the investigation of continuous T2DM features while modeling DBNs rather than discrete model. Anatural progression of this work is to involve exploring the extension of these models with more latent variables to capture a greater variety of factors [Factorial] that characterizes key changes in the clinical and complications data. Ultimately, we intend to use these latent status to facilitate us to identify different cohorts of patients who have different dynamics and therefore stratify them so that more can be inferred around the different expressions of the disease and its advancement. It is suggested that the association of these risk factors and different complications is investigated in future studies.

**Appendix I**

The output of the IC\* algorithm is a matrix P, defined as follows (Pearl (2000)):

* The possibility of existing a latent variable between variable i (rows) and j (columns) is highlighted in green color (P (i, j) = P(j, i) = 2, if there is a latent variable L such that i<-L->j).
* The possibility of existing a latent variable or a link between variable i (rows) and j (columns) is highlighted in yellow color (P (i, j) = -1, if there is either a latent variable L such that i <-L->j or there is a directed edge from i->j).
* The possibility of existing a link between variable i (rows) and j (columns) is highlighted in orange colour (P (i, j) = -2, if there is a marked directed i-\*>j edge).

We characterized comorbidities into seven groups:

1. Diabetic nephropathy or Kidney disease:  
         - Essential hypertension  
         - Nephritis and nephropathy  
         - Essential hypertension begin
2. Liver disease  
         - Chronic liver disease without mention of alcohol

- Another non-alcoholic chronic liver disease

1. Neuropathy  
         - Polyneuropathy in diabetes
2. Retinopathy  
         - Diabetic Retinopathy  
         - Diabetic Retinopathy Proliferative  
         - Diabetic Retinopathy Simple  
         - Retinopathy Simple Unspecified
3. Hypertension

There are five steps that must be taken to generate the result obtained from the method:

1. Introducing diabetic risk factors.
2. Defining how the latent variable and comorbidities are expected to affect the risk factors.
3. Implementing the HMM as our probabilistic model and parameters.
4. Dealing with the unbalanced data; Imbalanced data are a major clinical problem for predicting a model and the main cause of low predictive accuracy of rare cases.
5. Assessing and evaluating the model, by using different evaluation models and enumerating uncertainties and the error rate.

Our simulation was implemented using MATLAB with the Bayes Net Toolbox [16].

We computed results under different assumptions to prove the influence of a variable under sensitivity analysis can be beneficial for many reasons, including:

* Identifying important connections between observations, diabetes risk factors, and predictions, which leads to the development of a better model.
* Enhancing the understanding of the relationships between risk factors and different comorbidities in a Bayesian model.
* Simplifying the model by removing some risk factors in diabetes that have less impact on the prediction of comorbidities.
* Exploiting uncertainty reduction methods such as the IC\* algorithm, and the LS approach, to increase the robustness of our structure.

We built our model based on three primary aims:

1. To tackle the imbalanced data.

2. To investigate learning DBNs in the presence of latent variables.

3. To ascertain Latent variables and yield better classification accuracy by adding a latent variable.

The model consists of following steps:

1. Classifying the relevant risk factors and predictors. The test results and comorbidities have been identified as predictors.
2. Modelling the Bayesian probabilities and dependencies of the observed data and latent variable, using qualitative methodology.
3. Exploring the latent variable and exploiting knowledge from the experience with the learning structure and conditional dependencies. This information was assembled into a Hidden Markov Model (HMM) to investigate the likely impact of the latent variable.
4. Employing an initiative assessment methodology by building the latent variable structure learned from the IC\* algorithm (Pearl and Verma, 1991). We monitored the latent variable behaviour to discover the dynamic nature of comorbidities and other risk factors and their dependencies.

Then, the analysis was checked when initially performed and then checked again at the end of the process. [ref]

The result of the IC\* correlation analysis is summarised in Table 2. Moreover, the result of True Average Link Strength (LSTA), and Blind Average Link Strength (LSBA) on the same data in the table 0 can be seen. Data from this Table 1 can be compared with the data in Table 2 which shows the confidence in understanding a suitable structure for the latent and observed variables.

Figure 11: Latent variable prediction pattern for the patient visits (early prediction by using ICLS latent variable).

Figure 12: Latent variable prediction pattern for the patient visits (same time prediction by using ICLS latent variable).

Figure 13: Latent variable prediction pattern for the patient visits (early prediction by using ICLS latent variable).