University of Reading

Analysis for warning factors of type 2 diabetes mellitus complications with Markov blanket based on a Bayesian network model

Article

Accepted Version

Liu, S., Zhang, R., Shang, X. and Li, W. ORCID: https://orcid.org/0000-0003-2878-3185 (2020) Analysis for warning factors of type 2 diabetes mellitus complications with Markov blanket based on a Bayesian network model. Computer Methods and Programs in Biomedicine, 188. 105302. ISSN 1872-7565 doi: 10.1016/j.cmpb.2019.105302 Available at https://centaur.reading.ac.uk/88400/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1016/j.cmpb.2019.105302

Publisher: Elsevier

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

University of Reading

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Analysis for warning factors of type 2 diabetes mellitus complications with Markov blanket based on a Bayesian network model

Siying Liu[a], Runtong Zhang[a], Xiaopu Shang[a], Weizi Li[b]

[a]School of Economics and Management, Beijing Jiaotong University, Beijing 100044, PR China

[b]Informatics Research Center, University of Reading, Berkshire RG6 6AH, United Kingdom

Abstract

Background and objective: Type 2 diabetes mellitus (T2DM) complications seriously affect the quality of life and could not be cured completely. Actions should be taken for prevention and self-management. Analysis of warning factors is beneficial for patients, on which some previous studies focused. They generally used the professional medical test factors or complete factors to predict and prevent, but it was inconvenient and impractical for patients to self-manage. With this in mind, this study built a Bayesian network (BN) model, from the perspective of diabetic patients' self-management and prevention, to predict six complications of T2DM using the selected warning factors which patients could have access from medical examination. Furthermore, the model was analyzed to explore the relationships between physiological variables and T2DM complications, as well as the complications themselves. The model aims to help patients with T2DM self-manage and prevent themselves from complications.

Methods: The dataset was collected from a well-known data center called the National Health Clinical Center between 1st January 2009 and 31st December 2009. After preprocess and impute the data, a BN model merging expert knowledge was built

with Bootstrap and Tabu search algorithm. Markov Blanket (MB) was used to select the warning factors and predict T2DM complications. Moreover, a Bayesian network without prior information (BN-wopi) model learned using 10-fold cross-validation both in structure and in parameters was added to compare with other classifiers learned using 10-fold cross-validation fairly. The warning factors were selected according the structure learned in each fold and were used to predict. Finally, the performance of two BN models using warning features were compared with Naïve Bayes model, Random Forest model, and C5.0 Decision Tree model, which used all features to predict. Besides, the validation parameters of the proposed model were also compared with those in existing studies using some other variables in clinical data or biomedical data to predict T2DM complications.

Results: Experimental results indicated that the BN models using warning factors performed statistically better than their counterparts using all other variables in predicting T2DM complications. In addition, the proposed BN model were effective and significant in predicting diabetic nephropathy (DN) (AUC: 0.831), diabetic foot (DF) (AUC: 0.905), diabetic macrovascular complications (DMV) (AUC: 0.753) and diabetic ketoacidosis (DK) (AUC: 0.877) with the selected warning factors compared with other experiments.

Conclusions: The warning factors of DN, DF, DMV, and DK selected by MB in this research might be able to help predict certain T2DM complications effectively, and the proposed BN model might be used as a general tool for prevention, monitoring, and self-management.

Keywords： Bayesian network, Type 2 Diabetes Mellitus complications, prevention, self-management, warning factors

1. Introduction

Diabetes Mellitus (DM) is associated with various metabolic disorders, heavily enlarging the charge of non-communicable diseases [1], and chronic hyperglycemia due to insufficient insulin action is the main feature [2]. Type 2 diabetes mellitus (T2DM) accounts for 85% to 90% of all diabetic cases, which imposes the burden on not only individuals but the healthcare system [1,3]. Moreover, T2DM leads to complications such as kidney failure, blindness, cardiovascular diseases, nerve damage, ketoacidosis and foot problems, which seriously affect the quality of life, cause significant economic burdens to family and society, and even result in death [4-9]. The number of people with T2DM complications has been increasing, especially for young populations, which yearly increases the socio-economic societal burden of them [10]. According to a research carried out by Chapman et al. [11], the microvascular complications contribute extensively to the economic burden, and with the condition has been being more serious, the overall cost was exceeding £ 70 million over the two years. Therefore, the prevention of T2DM complications is significant for hospitals and patients with diabetes.

There are two methods of preventing complications that continuous medical care and long-term self-management. T2DM could not be cured completely but be controlled by medication, diet, and lifestyle changes [12-13]. In other words, the long-term prevention of complications is necessary for the patients with T2DM. However, nowadays, healthcare resources are limited and costly, especially in the impoverished area [14-15]. It is impractical to rely entirely on the medical care in the hospital. Moreover, doctors are more likely to advise patients to do diabetic self-care at home if the symptom is not severe. Consequently, actions should be taken for prevention and self-management.

Warning factors are strong contributors to certain outcome variables, which could be used widely to predict and support decision making in medical informatics [16-17]. For patients, warning factors could prevent them from the presence of a particular disease. In other words, analysis for warning factors plays a vital role in disease prevention. Machine Learning has been developing rapidly in recent years, and simulation models built with it are increasingly applied to derive predictions and analyze the warning factors of T2DM complications. Cho et al. [18] compared the prediction results of diabetic nephropathy between several classification methods with 39 features and showed the effect of each feature on the decision using the visualization tool. A fuzzy classification model was employed to predict heart and kidney complications with six attributes of clinical data [19]. Leung et al. [20] using 10 clinical attributes and 5 genetic attributes built different models with seven machine learning methods to predict warning patterns in diabetic kidney disease. There are also some studies suggesting the biomedical factors such as endostatin [21-22], microRNAs (miRNAs) [23] and red blood cell deformability index [24] to predict T2DM complications or predict the T2DM development in certain patients.

It is essential for diabetic patients to self-manage and to improve self-care skills with known warning features so that they can prevent themselves from complications to avoid severe conditions [25]. Therefore, the selection of the features from datasets is significant. It is inconvenient and unpractical for patients to obtain the features referring to professional medical test indicators and to input all features to the models that not dealing with the incomplete data sets when they take actions to self-manage and make prevention. However, there are few studies analyzing warning factors of T2DM complications from the perspective of diabetic patients' self-management and prevention, and most of them only focus on one complication.

Bayesian networks (BNs) representing the conditional probability between random variables graphically is an increasingly popular method that deals with uncertain and complex fields. They are more capable of dealing with incomplete data sets with better performance than many other models and revealing the causal relationships between variables [26-27]. Our paper mainly aims to build a Bayesian network (BN) model to

find out and analyze the warning factors of T2DM complications with Markov blanket (MB) and use them to predict complications of T2DM. The warning features consist of urine test data, glycated hemoglobin (HbA1c) and biochemical parameters, which patients could have access from medical examination. We analyze the model to explore the relationships between physiological variables and T2DM complications, as well as the complications themselves. The model could probably be beneficial for prevention, monitoring and self-management, and that might be more convenient and manageable for patients. Besides, the model is probably capable of applying in different scenarios based on hypothetical cases when new observations are instantiated according to the patients' situation.

This paper is organized as follows. Section 2 explains the collected data used in this study and its processing methods, displays the flow of building a BN model, and analyzes the final T2DM complications model and the warning features selection. Section 3 adds a Bayesian network without prior information model (BN-wopi), and describes the comparison results with the other three models used in the diagnosis and presents the effectiveness of the previous studies' prediction to demonstrate the performance of warning factors of the BN model in prediction further. Section 4 discusses the warning factors of T2DM complications using MB and several specific scenarios. We conclude the study and suggest our future topics and applications in the last section.

2. Materials and Methods

In this section, we will demonstrate some concepts of methods we used in the study and the learning flow of our BN model in detail. First, we will describe the data collected and the approaches of preprocessing and imputation. Then, we will introduce BNs and MB, and some related basic concepts of connection and separation in BNs. After that, the process of building a T2DM complication model will be presented, and the structural and parametric learning approaches of the model will be described in

detail. We will describe the final T2DM complications model and analyze the warning features of complications with MB in the last section.

Our models are developed in RStudio, with mice, vice, bnlearn and ROCR packages mainly. The BNs model is operated in Netica to analyze the warning features of T2DM complications.

2.1 Data Collection

Based on data from a well-known data center called the National Health Clinical Center, we learned the structure and the probability distributions of our BN model. All inpatient sample databases related complications of diabetes were provided by the General Hospital of the People's Liberation Army (PLAGH), and they were extracted for the period 1st January 2009 to 31st December 2009. The original data collected particularly from the Hospital Information System (HIS) were followed the principle of authenticity and professional characteristic that is recording the accurate and unprocessed value of each test and real information about the inpatients. Each case was taken down after diagnosis during hospitalization. It could be classified into three categories according to the property of data: basic information, physiological information, and complications information. Specifically, physiological information consists of urine test variables, HbA1C test variables and biochemical test variables.

In this data set, there are 43 features in all. To protect the privacy of patients, we removed the ID column. Moreover, we also removed lipase (LPS), ferrum (Fe) and unsaturated iron-binding capacity (TIBC) variables because of the large portion of missing values. Therefore, the dataset we used in this study to build the model and assess the warning feature contains 39 features including age and gender, 13 items related to the urine test, glycated hemoglobin (HbA1c) and 23 items related to the biochemical test. There are six complications variables, which are diabetic nephropathy (DN), diabetic retinopathy (DR), diabetic foot (DF), diabetic macrovascular complications (DMV), diabetic peripheral neuropathy (DPN), and diabetic ketoacidosis (DK) in this study.

Considering the readability, the convenience of training model using these data,

and the influence of outlier on the result, we integrated the information of each patient into one case and then removed the value of the same variable that was contradictory during one stay in the hospital, which leads to the independence between the cases. The contradictory value here means that the ones of discretized variables, and there are three situation (for one patient in one stay in the hospital): (i) if there are less 30% values of the variable in other states, whereas other values are in the certain state, the value of the variable is regarded as in this state. If there are more than 30% of values are abnormal values, we are concerned about them; (ii) If the abnormal values are in the same state, such as Low or High, the value of the variable is regarded as in this state; (iii) if the abnormal values are in a different state, the value of the variable is regarded as a missing value. As a result, the total number of data used in the model is 1485 and the records with complications are 755.

2.2 Data Preprocessing and Imputation

We decided to rely on discrete state BNs. Thus each continuous variable has to be discretized. According to the criterion of age provided by the World Health Organization and the age distribution in the dataset, we set three cut-offs of it. For variables present nominally, we indexed them with the discrete values directly. In addition, for variables with normal range values, we set two or three discrete values of normal and high or low normal and high. The details and the abbreviations of variables are listed in Table 1.

Table 1: List of variables and their relative states.

Variables	TYPE	Reference Value	States	%missing
Age	Numeric	/	[20-55, 56-69,≥70]	0.00%
Gender	Nominal	/	[male, female]	0.00%
Urine Leucocyte (U-LEU)	Nominal	0-36/ul	[normal, high]	0.20%
Specific Gravity (SG)	Nominal	neg	[normal, low, high]	0.27%
Urobilinogen (URO)	Nominal	neg	[neg, pos]	0.20%
Urine Bilirubin (U-BIL)	Nominal	neg	[neg, pos]	0.20%
Red Blood Cell (RBC)	Numeric	0-27/ul	[normal, high]	0.20%
Yeast-Like Cells (YLC)	Numeric	0-0/ul	[normal, high]	0.27%
Glucosuria (U-GLU)	Numeric	0-0mg/dl	[normal, high]	0.20%

Crystaluria (CRY)	Numeric	0	[normal, high]	0.34%
Urine PH value (PH)	Numeric	4.5-7.9	[normal, high]	0.20%
Urine Color (COL)	Nominal	yellow、light	[normal, dark, other]	8.35%
Griess Test (NIT)	Nominal	neg	[neg, pos]	0.20%
Urine Turbidity (TUR)	Nominal	clear	[normal, slight, turbid]	13.60%
Urine Ketone (U-KET)	Nominal	neg	[normal, high]	0.20%
Glycated Hemoglobin (HbA1C)	Numeric	4.1-6.5%	[normal, low, high]	36.50%
Alanine Aminotransferase (ALT)	Numeric	0-40U/L	[normal, high]	2.09%
Aspartate Aminotransferase (AST)	Numeric	0-40U/L	[normal, high]	2.15%
Total Protein (TP)	Numeric	55-80g/L	[normal, low, high]	8.82%
Serum Albumin (ALB)	Numeric	35-50g/L	[normal, low, high]	5.05%
Total Bilirubin (TB)	Numeric	0-21umol/L	[normal, low, high]	10.24%
Direct Bilirubin (DBIL)	Numeric	0-8.6umol/L	[normal, high]	10.91%
Alkaline Phosphatase (ALP)	Numeric	0-130U/L	[normal, high]	15.42%
Urea (UREA)	Numeric	1.8-7.5mmol/L	[normal, low, high]	2.36%
γ Glutamyl Transferase (GGT)	Numeric	0-50U/L	[normal, high, higher]	14.14%
Creatinine (Cr)	Numeric	30-110umol/L	[normal, low, high]	2.36%
Glucose (GLU)	Numeric	3.4-6.1mmol/L	[normal, low, high]	5.25%
Triglycerides (TRIG)	Numeric	0.4-1.7mmol/L	[normal, high, chyle]	22.96%
Uric Acid (UA)	Numeric	104-444mmol/L	[normal, low, high]	8.35%
Total Cholesterol (TC)	Numeric	3.1-5.7mmol/L	[normal, low, high]	22.56%
Creatine Kinase (CK)	Numeric	2-200U/L	[normal, high, hemolysis]	21.35%
Lactate Dehydrogenase (LDH)	Numeric	40-250U/L	[normal, low, high, hemolysis]	18.86%
Calcium (Ca)	Numeric	2.09-2.54mmol/L	[normal, low, high]	12.46%
Sodium (Na)	Numeric	130-150mmol/L	[normal, low, high]	4.65%
Potassium (K)	Numeric	3.5-5.5mmol/L	[normal, low, high]	3.97%

Chloride (CL)	Numeric	94-110mmol/L	[normal, low, high]	6.33%
Inorganic Phosphorus (IP)	Numeric	0.89-1.6mmol/L	[normal, low, high]	20.13%
Magnesium (Mg)	Numeric	0.6-1.4mmol/L	[normal, low, high]	21.75%
High-Density Lipoprotein (HDL)	Numeric	1-1.6mmol/L	[normal, low, high]	35.42%

There are three types of missing data prevalent in statistics literature [28], (a) Missing Completely at Random (MCAR), (b) Missing at Random (MAR) and (c) Missing Not at Random (MNAR). The probability of missing data classified as "MCAR" is independent of observed values and missing data. The probability of missing data classified as "MAR" does not depend on the missing values but the observed data of other features, while the probability of missing data classified as "MNAR" depends on both of them. In our dataset, we could regard all of the missing data as "MCAR" and "MAR" under the assumption that something like the staff in hospital deleted values on purpose or patients refused to do the tests would not happen. It means that we could use some imputation methods to process the missing data.

Multiple imputations are proved to be preferable rather than removing data entirely in some areas [29]. It is used in some literature for preprocessing missing values where the method called predictive mean matching usually presents the best performance when there are fewer than 50 percent cases including missing values [30-33]. Observing some missing values in the dataset, we used the VIM package embedded in RStudio to evaluate the distribution of missing values (refer to Figure 1). The red line takes up a part of the area, meaning that most of the missing values are about HbA1C and High-Density Lipoprotein (HDL), which, however, take up fewer than 50 percent cases reported in Table 1 specifically.

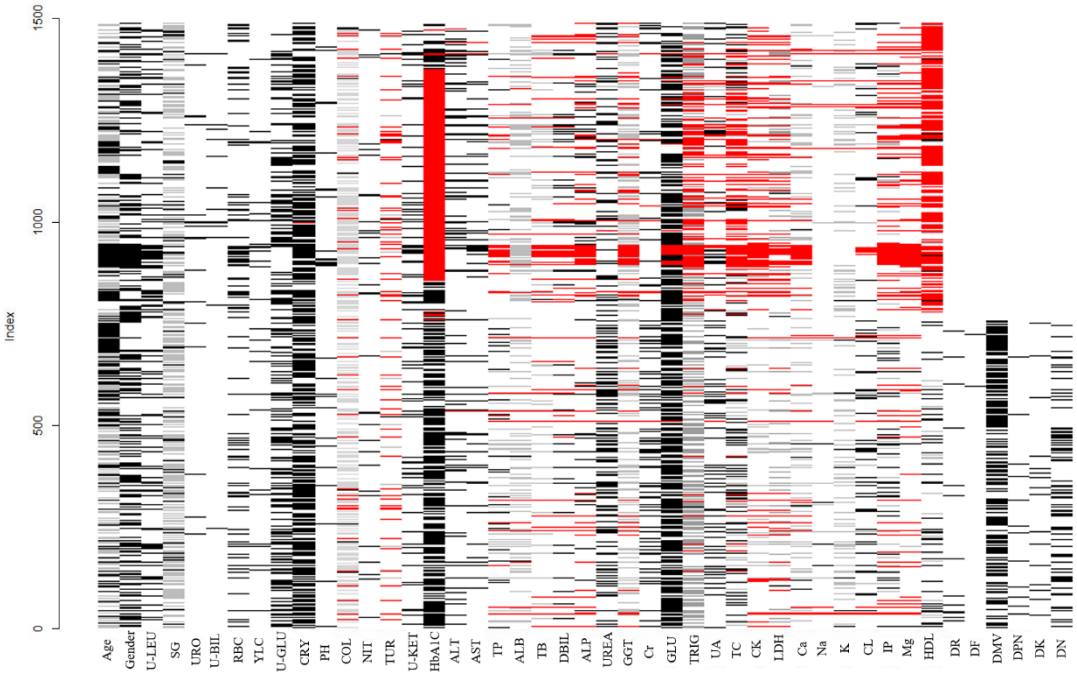

Figure 1: Distribution of missing values.

Consequently, we did multiple imputations with predictive mean matching (PMM) for missing data [34-35]. Using the MICE package embedded in RStudio, we set the number of iterations to 50 to reduce the impact of random factors. The density of difference between before and after imputation is shown in Figure 2. It is worth noting that a good fitting effect is reported, so the dataset after processing could be used for model training.

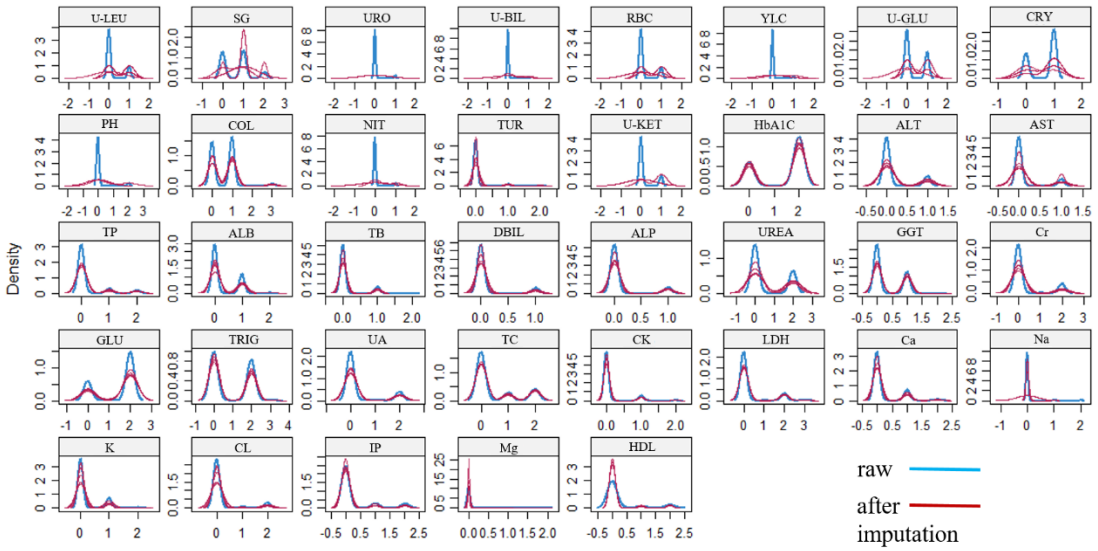

Figure 2: Density plot.

For the discrete and orderly value of all variables, data are analyzed using Kendall rank correlation coefficient. As clear showed in Figure 3, the correlation between multiple variables reveals nearly linear independence, whereas the linear correlation shows between others. The Bayesian network is a method that can deal with the complex correlation between the variables and indicate more information about data. Therefore, we build the T2DM complication model based on it.

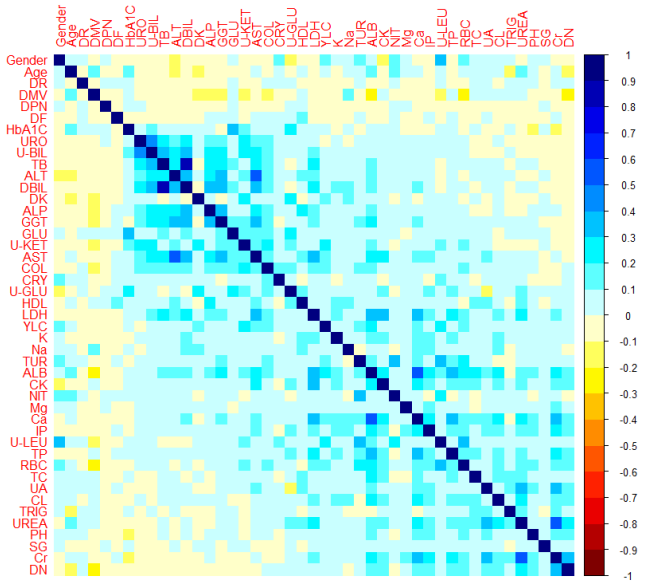

Figure 3: Maps of the correlation coefficients of variables.

The dataset was split into ten subsets of approximately equal size using 10-fold cross-validation [36], namely nine subsets used for training and the rest one subset for testing in turn for ten times. It aims to prevent the BNs model from overfitting through learning the model with the training sets and using the test to measure its performance. In our study, we used all data with the bootstrap approach (refer to Section 2.4.1) to learn the structure of BNs and used the training set with 10-fold cross-validation to do the parameters learning of BNs, calculating the mean values as the learning results. Ten times of probability prediction will be presented through some validation parameters as the final validation, with the purpose of a robust and effective model.

2.3 Bayesian Networks and Markov Blanket

Bayesian networks (BNs) is annotated directed graphs that represent a set of variables as nodes in a network, connected by edges representing the conditional

probabilistic relations between them. A pair (G, T), where G represents a directed acyclic graph (DAG) and T is the set of parameters quantifying the network, specifies a Bayesian network (BN) [37-38]. A Bayesian network B always defines a joint probability distribution that could be factorized as a result of several conditional distributions over a set of random variables [39-40]:

$$P_B(X_1, \cdots, X_n) = \prod_{i=1}^{n} P_B\left(X_i \middle| \pi_{X_i}\right) = \prod_{i=1}^{n} \theta_{\left(X_i \middle| \pi_{X_i}\right)} \qquad (1)$$

Note that $\prod X_i$ denotes the parent nodes of the random variable X_i and θ represents the conditional probability. It suggests that any node given the value of its parent nodes is conditionally independent of other all nodes that are not its descendants. This is known as Markov property [41]. The Markov blanket (MB) is the smallest subset of Bayesian network instantiating the property that all variables outside the MB could be deleted without influence on the target node and thus will have no impact on the accuracy of classification. It could be displayed as the following equation [42]:

$$P\left(T \middle| Y, MB(T)\right) = P\left(T \middle| MB(T)\right) \propto P(\pi_T) P(T|\pi_T) P(D_T|T) P\left(D_T \middle| \pi_{D_T}, T\right) P\left(\pi_{D_T}\right) (2)$$

$MB(T)$ is the Markov blanket of target node T. π_T represents the set of parent nodes of T whose child nodes are illustrated as D_T, and thus π_{D_T} describes the other parent nodes of D_T except the node T. Hence, MB is usually used in the feature selection [43-44]. There are three types of connections and d-separation in MB. One of the connections is called serial ($X \rightarrow Z \rightarrow Y$ or $X \leftarrow Z \leftarrow Y$), known as the intermediate cause where Z makes X and Y independent. The diverging connection is $X \leftarrow Z \rightarrow Y$, where X is independent with Y if Z is instantiated and vice versa. That is known as the common cause. If a trail is shown as $X \rightarrow Z \leftarrow Y$ and $Z \rightarrow R$, it is regarded as a converging connection. It makes X and Y independent only if not knowing neither of Z and R, which means common effect. The independence between X and Y reported above could be called that X and Y are d-separated. It could be concluded that if X and Y are d-separated by Z, X and Y are conditionally independent given by Z. Therefore, it is apparent that any node in the BN is d-separated of the nodes included in the non-Markov blanket given its Markov blanket.

2.4 Learning Bayesian Networks

There are two necessary steps to obtain a BN model: (i) structural learning to find the global optimum global structure proved as an NP problem, and (ii) parametric learning to estimate the conditional probability among nodes given a DAG. The entire process of building a T2DM complication model based on BN is presented in Figure 4.

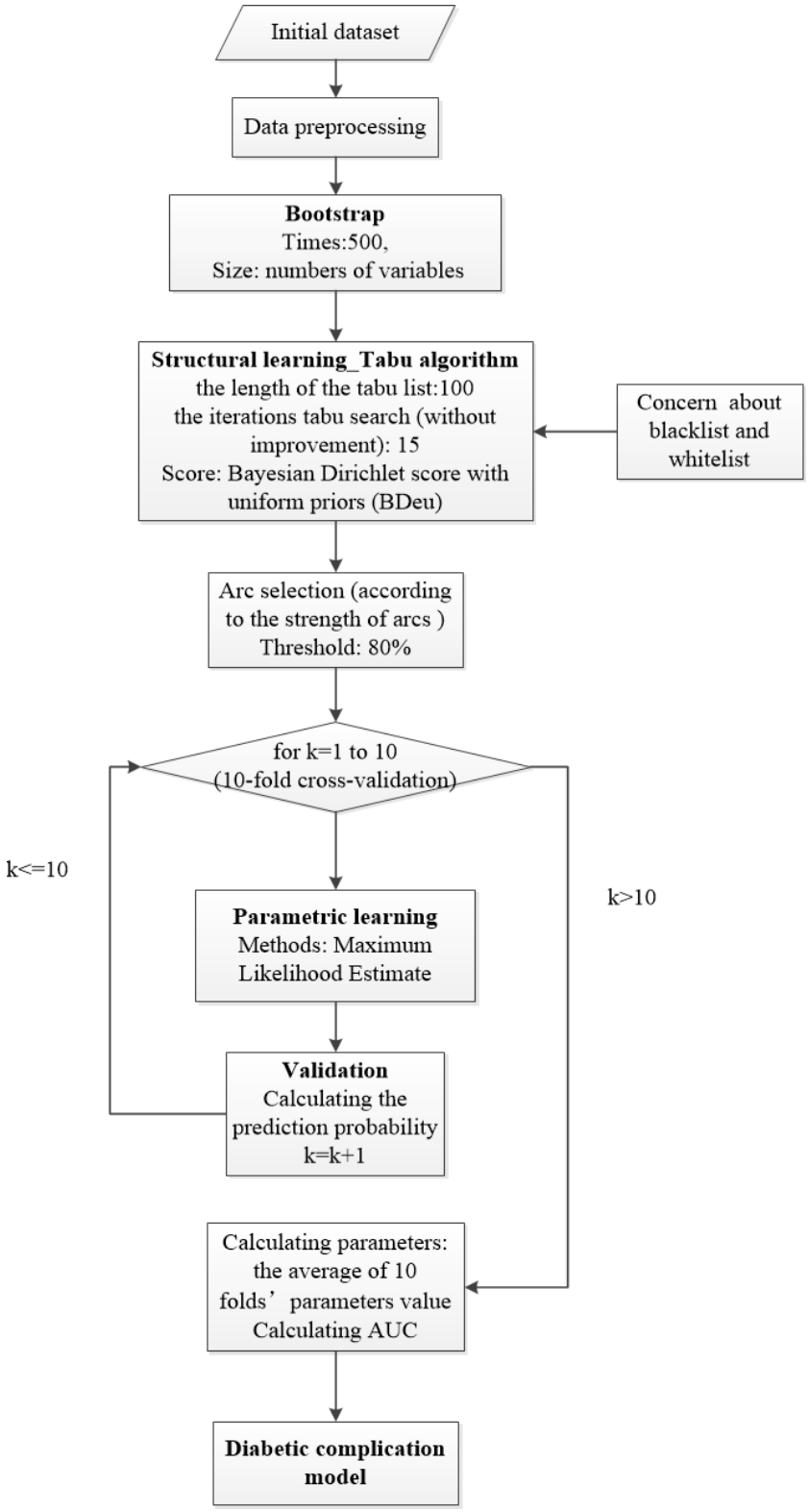

Figure 4: The implementation flow of the T2DM model.

2.4.1 Structural Learning

BNs learns the real probability distribution by updating the posterior distributions

according to the observed evidence, based on prior knowledge. Therefore, we focus on combining data-driven evidence with prior knowledge derived from previous research when learning the BN structure.

Sharma et al. [45] reported that total cholesterol (TC) and high-density lipoprotein (HDL) are related to cardiovascular diseases. According to Doliba et al. [46], Na+ levels may associate with the pathology of diabetic cardiomyopathy. Consequently, we assume that HDL, TC, and Na are related to DMV. Furthermore, Yang et al. [47] explored that $Na_V1.3$ and $Na_V1.7$, which are encoded by the Sodium (Na), contribute to the cause-effect relation between diabetes and painful neuropathy. Fadini et al. suggested that lower TRIG may protect kidney function from lip toxicity [48]. In other words, Na also seems to play an important role in the development of DPN and TRIG probably has an impact on DN.

Based on the prior knowledge mentioned above, we created a whitelist, which is summarized in Table 2. It presents the forced edges on outcome nodes in our structure. In addition, age and gender are the variables not determined by the model. More in general, the state of the two factors does not depend on the rest of the model. Therefore, we put the edges from two nodes to other nodes into a blacklist. The whitelist and the blacklist are integrated into the structure learning of BN classifier, which we will present in the following details.

Table 2: Forced edges on outcome nodes.

Outcome Nodes	Forced Parent Node
DN	TRIG
DF	/
DMV	TC, Na, HDL
DPN	Na
DK	/
DR	/

Two options could be chosen if several candidates' models can be accessed. One option is the most effective model, and the other is the average model that average over

the other models. The final structure of our model was obtained with the approach called bootstrap by repeating 500 times structure learning, namely 500 BNs were learned. Each network was explored with a Tabu Search algorithm (tabu) [49] according to the likelihood-equivalence Bayesian Dirichlet score with uniform priors (BDeu), merging the whitelist listed above and the blacklist where age and gender variables were never the parent nodes. We set the length of the Tabu list to 100 and the Tabu search iterations to 15 [50]. To ensure the robustness of the model, the network learned the averaged the arc strength of 500 models calculated by the conditional probability of two connected nodes. Consequently, arcs, whose strengths are above 0.8, remained in the final structure given a threshold valued 80% (refer to Figure 5).

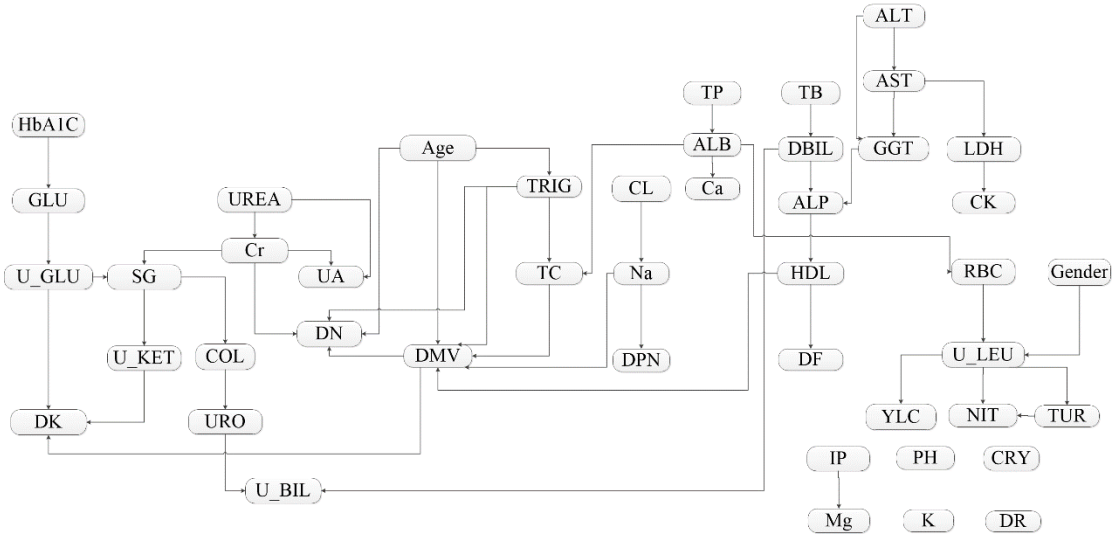

Figure 5: The network of T2DM complications model.

2.4.2 Parametric Learning

Parameters are the probabilities distribution of all variables. As mentioned above, 10-fold cross-validation was used to split the dataset into the training sets and test sets in parameter learning and make the estimation of the model accuracy. We can see the process of parametric learning in Figure 4. With ten subsets of approximately equal size divided in the beginning, the process repeated 10 times. There were nine subsets to train the model parameters with Maximum Likelihood Estimate (MLE) [51], and the rest subset was to validate the model performance with predicting probability using warning factors. After repeating for ten times, we calculate the average of conditional

probabilities (θ) of every node as the result of parameter learning and there were validation sets consisted of each test set and their predicting probability.

2.5 T2DM Complications Model and Warning Feature Analysis.

To obtain a legible and unambiguous graphical representation from the structure and parameters of the T2DM complications model learned with bnlearn package in R language, we operated the BN in Netica (refer to Figure 6). As shown in Figure 6, the number of edges is high. Note that BNs structure learned by data-driven integrated with expert knowledge might have the ability to demonstrate more relationships between T2DM complications and physiological variables.

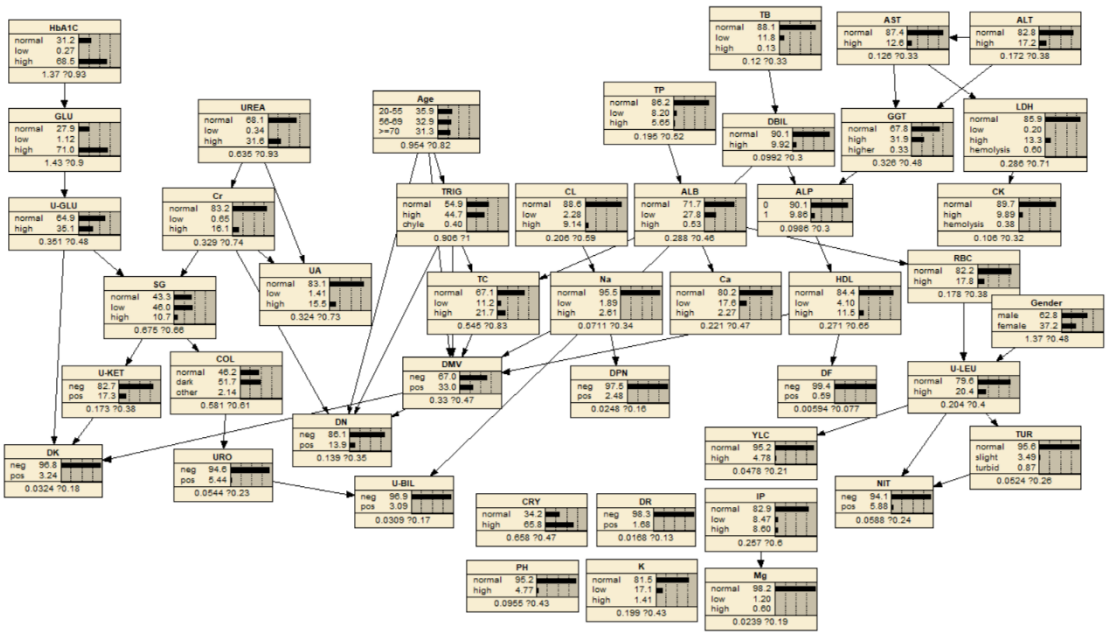

Figure 6: T2DM complications model operated in Netica.

As we can observe in Figure 6, the joint probability distribution of the BN model is factorized in 45 conditional probability tables (CPTs) where each table is for each node conditioned to the set of its parent nodes and presents the mean and standard deviation. It is worth noting that T2DM complications model seems to split the decision nodes and require two separate sub-networks to represent the original data distribution. Out of six T2DM complications, DR has no relationship with urine test items, HbA1c

and biochemical test items. In other words, urine test, HbA1c and biochemical test could not be able to predict DR, which would be explored in other features to examine the warning factors.

As for the other five complications, DK and DN variables are the child nodes of DMV variable, which is known as a common cause, and the connection between DK variable and DN variable is broken if DMV variable is initiated. Moreover, DMV variable is connected to DPN variable and DF variable through Na and HDL respectively.

The Market blanket of a variable consists of the set of its parent nodes, child nodes and the parent nodes of its child nodes as previously mentioned. The MB of DN variable is Age, TRIG, Cr and DMV variables, all of which are its parent nodes. That means the connection to DN variable via common cause trial and intermediate cause trial is broken if the MB is given. For DF and DPN variables, only the variables called HDL and Na are their MB respectively, which broke the trail to DMV variable. There are 10 variables including Age, TRIG, TC, Na, HDL that are parent nodes while DK and DN are child nodes. U-GLU, U-KET, Cr make the Markov blanket of DMV variables. As for DK variable, U-GLU, U-KET and DMV variables compose the MB. Table 3 summarizes the MB of five T2DM complications clearly.

Table 3: The Markov blanket of T2DM complications variables.

Complications Variables	Markov Blanket	Size of MB
DN	Age, TRIG, Cr, DMV	4
DF	HDL	1
DMV	Age, TRIG, TC, Na, HDL, DK, DN,U-GLU, U-KET, Cr	10
DPN	Na	1
DK	U-GLU, U-KET, DMV	3

3. Results

To ensure the effectiveness and robustness of our T2DM complication model, we have two kinds of performance comparison in this section. The first part refers to the

different baseline models including the Naïve Bayes model (NB) [52], Random Forest model (RF) [53], the C5.0 Decision Tree model (C5.0) [54]. We compare the performance from the perspective of several parameters with Area Under Curve (AUC) [55], 95% Confidence interval (95% CI), sensitivity and specificity. Following the formula described in (3) and (4), we could calculate sensitivity, which is also known as the true positive rate, and specificity which is known as true negative rate. Then we compare our validation parameters, especially AUC, with those in existing studies using some other variables in clinical data or biomedical data predicting T2DM complications to validate the performance of warning factors in the BN model in prediction in the second part.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{4}$$

3.1 Performance Comparison between Models

After the model learning, we validated the performance of prediction of the BN model using warning factors, and compared it with the NB, RF, and C5.0, which were used in the prediction of diseases with 10-fold cross-validation. In order to make a comparison with other classifiers fair, a Bayesian network without prior information (BN-wopi) model was studied, where both structural learning and parametric learning were performed using 10-fold cross-validation. The warning factors were selected according to the structure learned in each fold and were used to predict. The parameters of RF were set as follows: the number of variables randomly sampled as candidates at each split is set to three, and there are 100 trees allowed to grow. In relation to C5.0, the number of iteration was set to five. Note that warning factors were used to predict each outcome variable in the BN model and BN-wopi model, whereas all variables except the outcome variable were applied in classification tasks in NB, RF, and C5.0.

Because of dataset distribution and the number of positive cases and negative cases, it is not reasonable to take the threshold as 0.5 then calculate the confusion matrix or error ratio. However, AUC is a preferable method to validate the effect of models [56]

that takes the different predicted probabilities as the threshold and calculates the sensitivity and specificity respectively. Then there will be a Receiver Operating Characteristic Curve (ROC) and the area under it is the AUC.

We added the prediction probability and test data of each fold of different models to four lists in two columns respectively. Accomplished with each iteration, the prediction probability column in training datasets was taken as the threshold one by one, except the same value, to calculate the sensitivity, specificity and AUC, which were used together with 95% CI to illustrate the predictive effect of models in the last. Figure 7 describes the AUC, sensitivity and specificity indices of five models predicting the five complication variables each fold. In general, the two BN models outperformed other models in AUC and sensitivity.

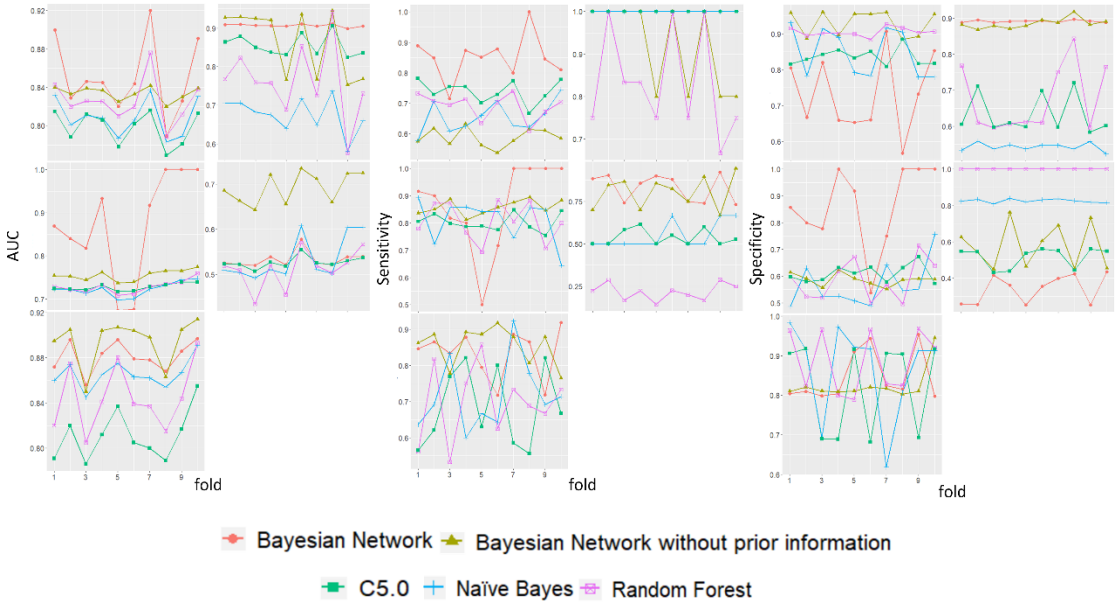

Figure 7: Fold by fold comparison in AUC, sensitivity and specificity. From top to bottom and left to right, the complications are DN, DF, DMV, DPN and DK orderly.

The analysis of variance (ANOVA) test indicated that at least two models for the all evaluated indices are statistically significantly different ($p < 0.05$) with 10 folds. Then, the p-values of the paired test regarding the AUC are summarized in Table 4. The statistical analysis indicated that two BN models performed statistically better than or similar to their counterparts. By comparing BN model and BN-wopi model, though there is a statistically significant difference in predicting some of the complications, they show basically equally powerful prediction.

In addition, as shown by the paired test results in Table 5, the sensitivity of two BN models are better than the other three models generally, whereas both BN models performed statistically similarly. In relation to specificity, other models statistically outperformed the BN models. However, note that their sensitivity reached a lower rate, which is undesirable for T2DM complications prediction using waring factors especially for self-management of patients. On the other hand, BN-wopi performed statistically better than or similar to BN model in the specificity of most of the complications except DPN complications.

Table 4 : The p-values of the paired test for the AUC in the prediction of five complications (DN and DMV are in the upper triangular part, DF, DPN and DK are in the lower triangular part). The symbols in parentheses denote: (=) not statistically different; (+) statistically different, where the row is superior to the column; and (−) statistically different, where the row is inferior to the column.

DF\DN	BN	BN-wopi	NB	RF	C5.0
BN	-	0.46(=)	0.00(+)	0.14(=)	0.00(+)
BN-wopi	0.27(=)	-	0.14(=)	0.67(=)	0.00(+)
NB	0.00(-)	0.00(-)	-	0.46(=)	0.67(=)
RF	0.00(-)	0.00(-)	0.01(=)	-	0.09(=)
C5.0	0.22(=)	0.76(=)	0.00(+)	0.00(+)	-

DPN\DMV	BN	BN-wopi	NB	RF	C5.0
BN	-	0.00(+)	0.00(+)	0.00(+)	0.00(+)
BNs-wopi	0.00(+)	-	1.00(=)	1.00(=)	1.00(=)
NB	1.00(=)	0.00(-)	-	1.00(=)	1.00(=)
RF	0.92(=)	0.00(-)	0.91(=)	-	1.00(=)
C5.0	1.00(=)	0.00(-)	1.00(=)	1.00(=)	-

DK	BN	BNs-wopi	NB	RF	C5.0
BN	-	-	-	-	-
BNs-wopi	0.19(=)	-	-	-	-
NB	0.19(=)	0.01(=)	-	-	-
RF	0.00(-)	0.00(-)	0.09(=)	-	-
C5.0	0.00(-)	0.00(-)	0.00(-)	0.00(-)	-

Table 6 lists detailed comparison result and Figure 8 shows the AUC of four models predicting the five complication variables. In relation to the different performance

between the two BN models, BN-wopi outperformed the BN model with prior information in the variables of DPN and DK. It is most likely because the expert knowledge was not reflected in the dataset consisting of a limit number of cases so that the fitting between data and structure of the model is not very well in the BN model. Moreover, the performance in sensitivity and specificity between the two BN models is opposite in the variables of DN and DF. The BN model performed better in sensitivity, whereas BN-wopi performed better in specificity. The values of sensitivity and specificity of a model depend on the point which is the closest to (1,1) in order to maximize both of them on the coordinate plane (refer to Figure 8). To some extent, they are inverse. Due to the different structures, it is possible for two BN models that performed differently in sensitivity and specificity, which has little impact on the result of the comparison. Out of the four models learned using 10-fold cross-validation totally, it is clear that the BN-wopi model performs the best. When the confidence level is identical, narrower confidence intervention leads to the higher significance the AUC is. For the variables of DN, DPN and DK, the 95%CI of BN-wopi model is the narrowest, and for DMV variable, BN-wopi model performs the second best. Therefore, the AUC could represent the effectiveness of models to some extent, which means BN models using warning factors could give the best classification performances out of the other three models.

Table 5: The p-values of the paired test for the sensitivity (upper triangular part) and specificity (lower triangular part) indices. The symbols in parentheses denote: (=) not statistically different; (+) statistically different, where the row is superior to the column; and (−) statistically different, where the row is inferior to the column.

DN	BN	BN-wopi	NB	RF	C5.0
BN	-	0.00(+)	0.00(+)	0.00(+)	0.00(+)
BN-wopi	0.00(+)	-	0.01(=)	0.00(-)	0.00(-)
NB	0.00(+)	0.00(-)	-	0.10(=)	0.00(-)
RF	0.00(+)	0.70(=)	0.12(=)	-	0.07(=)
C5.0	0.00(+)	0.00(-)	0.70(=)	0.06(=)	-

DF	BN	BN-wopi	NB	RF	C5.0
BN	-	0.10(=)	1.00(=)	0.00(+)	1.00(=)

BN-wopi	0.79(=)	-	0.10(=)	0.07(=)	0.10(=)
NB	0.00(-)	0.00(-)	-	0.00(+)	1.00(=)
RF	0.00(-)	0.00(-)	0.00(+)	-	0.00(-)
C5.0	0.00(-)	0.00(-)	0.00(+)	0.11(=)	-

DMV	BN	BN-wopi	NB	RF	C5.0
BN	-	1.00(=)	1.00(=)	1.00(=)	1.00(=)
BN-wopi	0.00(+)	-	1.00(=)	1.00(=)	1.00(=)
NB	0.00(-)	1.00(=)	-	1.00(=)	1.00(=)
RF	0.00(+)	1.00(=)	1.00(=)	-	1.00(=)
C5.0	0.00(+)	1.00(=)	1.00(=)	1.00(=)	-

DPN	BN	BN-wopi	NB	RF	C5.0
BN	-	0.85(=)	0.00(+)	0.00(+)	0.00(+)
BN-wopi	0.00(-)	-	0.00(+)	0.00(+)	0.00(+)
NB	0.00(+)	0.00(+)	-	0.00(+)	0.85(=)
RF	0.00(+)	0.00(+)	0.00(+)	-	0.00(-)
C5.0	0.00(+)	0.05(=)	0.00(-)	0.00(-)	-

DK	BN	BN-wopi	NB	RF	C5.0
BN	-	1.00(=)	0.00(+)	0.00(+)	0.00(+)
BN-wopi	1.00(=)	-	0.00(+)	0.00(+)	0.00(+)
NB	0.00(+)	0.00(+)	-	1.00(=)	1.00(=)
RF	0.00(+)	0.00(+)	1.00(=)	-	1.00(=)
C5.0	0.00(+)	0.00(+)	1.00(=)	1.00(=)	-

Table 6: Performance of different models. The value only in bold performs the best in the row out of the last four models. The value in bold and italic performs the best in the row out of all models and is from the BN model.

	Parameters\Models	BN	BN-wopi	NB	RF	C5.0
DN	AUC	***0.831***	**0.83**	0.826	**0.83**	0.806
	95%CI	***0.7947-0.8665***	**0.7931-0.866**	0.7765-0.8761	0.7747-0.8857	0.7481-0.8639
	Sensitivity	***0.86***	0.511	0.727	0.694	**0.756**
	Specificity	0.655	**0.959**	0.788	0.906	0.826
DF	AUC	***0.905***	0.788	0.704	0.761	**0.851**
	95%CI	***0.8841-0.9268***	0.5109-1	0.5531-0.8545	0.555-0.9664	**0.7532-0.9496**
	Sensitivity	***1***	0.833	**1**	0.833	**1**
	Specificity	***0.891***	**0.884**	0.541	0.604	0.602
	AUC	***0.753***	**0.749**	0.723	0.745	0.726

DMV	95%CI	0.6861-0.8193	0.7056-0.7916	0.6615-0.7841	0.6693-0.8198	**0.6992-0.7524**
	Sensitivity	0.827	0.855	**0.861**	0.765	0.804
	Specificity	0.563	0.584	0.503	**0.633**	0.599
DPN	AUC	0.545	**0.685**	0.505	0.516	0.527
	95%CI	0.4129-0.6767	**0.5885-0.7807**	0.04515-0.9649	0.3111-0.7219	0.4238-0.6305
	Sensitivity	0.75	**0.875**	0.5	0.2	0.5
	Specificity	0.475	0.457	0.827	**1**	0.554
DK	AUC	0.877	**0.898**	0.876	0.858	0.819
	95%CI	0.8182-0.9362	**0.8553-0.9402**	0.7875-0.9651	0.7665-0.949	0.7354-0.9018
	Sensitivity	0.867	**0.875**	0.688	0.619	0.606
	Specificity	0.76	0.817	0.917	**0.968**	0.912

Consequently, our T2DM complications model with the Bayesian network classifies the condition of patient T2DM complications with the best performance using fewer variables. Furthermore, their graphical representation is very informative.

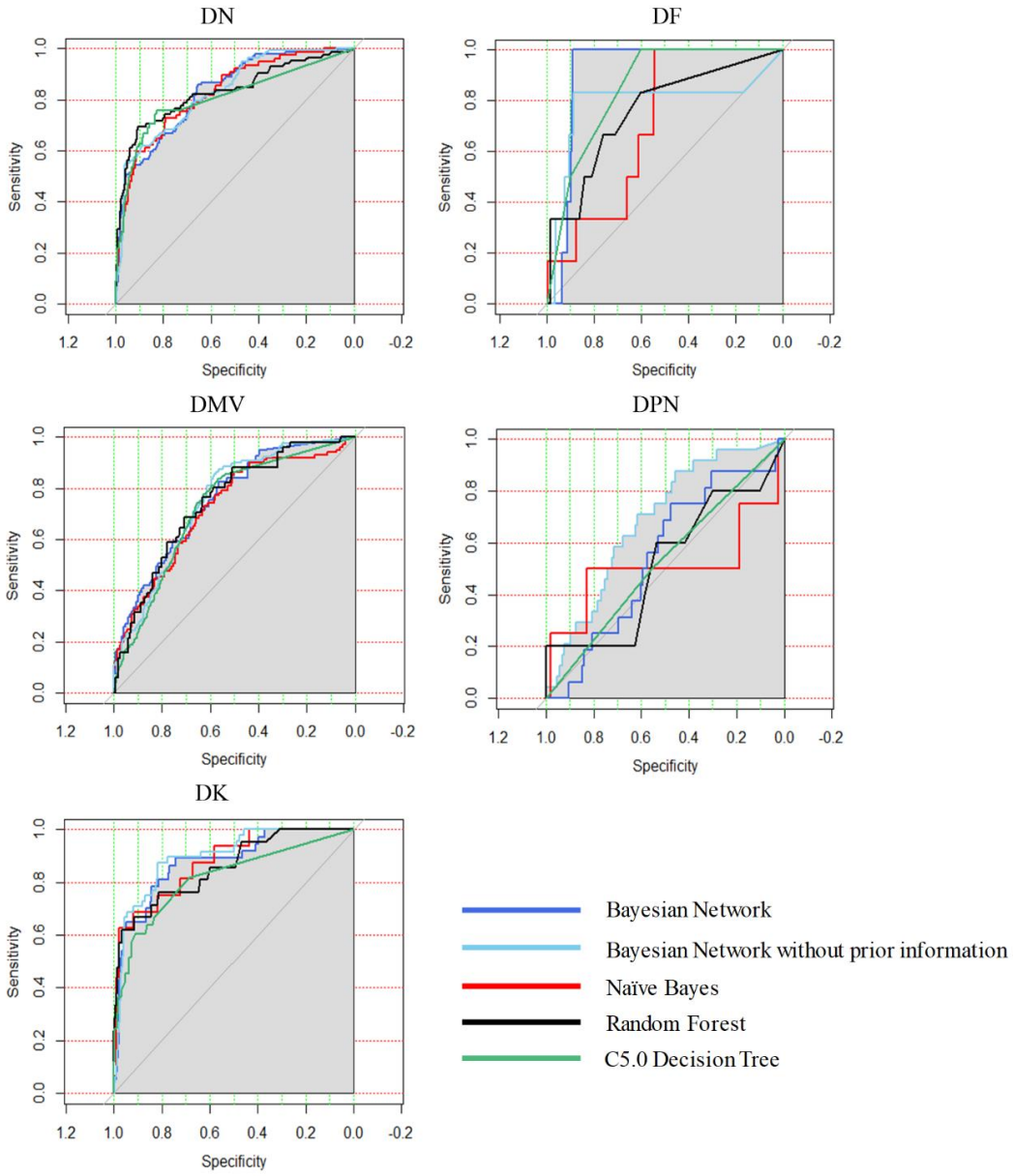

Figure 8: AUC of four models in five complication variables.

3.2 Performance Comparison with Other Experiments

In order to demonstrate the performance of the BN model further in prediction and the ability of complications warning using features selected by MB methods, we summarize the performance of other experiments conducted previously in predicting T2DM complications.

There are some related studies in Table 7 and most of them validate the performance with AUC of about 80%. For DK variables in five studies, more features demonstrate better effectiveness rather than one feature which needs to be examined

through the professional test. However, more feature means more resources are needed. In our model, there are four features that are available in medical examination to predict DK. Although the AUC in different models or studies does not have comparability of suggesting a better model, the AUC of 0.831 [95%CI: 0.7947-0.8665] is capable of proving that our model performs well in predicting DK. Moreover, the prediction of DF in our T2DM complications model are confirmed to be well with AUC of 0.905 [95%CI: 0.8841-0.9268], sensitivity of 1.0, and specificity of 0.891, while Irene et al. explored two methods with AUC of 0.776 [95% CI: 0.702–0.849], sensitivity of 0.83, specificity of 0.50, and AUC of 0.816 [95% CI: 0.757–0.874], sensitivity of 1.0, specificity of 0.32, respectively. Our model has a higher AUC of DMV prediction than AUC in existing related studies that is 0.75 on average. In other words, prediction in DMV also performs effectively with sensitivity of 0.827, higher than other related studies' listed in Table 7. However, the AUC of 0.545 in prediction in DPN means that the Na variable does not predict DPN very well, whereas Lin et al revealed that serum uric acid demonstrates stronger predictive power to DPN. Few studies predict DK with clinical data or biomedical factors. Nevertheless, AUC of 0.877 [95% CI: 0.8182-0.9362] could be seen as a well-formed model in general. Figure 9 shows the comparison of AUC of different related studies. Warmer color in the spectrum means a higher number of features.

Therefore, our model is proven to be effective in predicting DN, DF, DMV and DK based on the dataset. It might be beneficial in the prevention and self-management of diabetic patients in daily life.

Table 7: Performance and application of other relative studies

Relative studies	**T2DM complications**	**Features**	**Performance**	**Application**
Cho et al. [18]	Diabetic nephropathy	39 features	AUC: 0.969	Helping physicians to plan effective and proper treatment strategies
Song et al. [57]	Diabetic kidney disease	440 features	AUC: 0.82 [95%CI: 0.81–0.83]	Knowledge discovery
Leung et al. [20]	Diabetic kidney disease	10 clinical attributes, 5	ACC: about 90%	Research

		genetic attributes		
Chauhan et al. [21]	Diabetic kidney disease	Plasma endostatin	AUC: 0.77	Improving warning discrimination over traditional predictors
Lee et al. [24]	Diabetic nephropathy	(Fibrinogen ×ESR)/ EI	Sensitivity of 74.5%, Specificity of 63.1%, AUC: 0.762	Methods of assessment of DN
El-Ashmawy et al. [22]	Coronary artery calcification	Serum endostatin level	Sensitivity: 74.1%, Specificity: 71.4%, AUC: 0.776	Presence and progression of atherosclerosis in T2DM patients.
Sone et al. [58]	Coronary heart disease	HDL-cholesterol, the total cholesterol/ HDL-cholesterol	AUC: 0.726 AUC: 0.718	Clinical approaches to warning reduction among East Asians with diabetes.
Lin et al. [59]	Diabetic peripheral neuropathy	Serum uric acid	Sensitivity: 70.6%, Specificity: 65.2%, AUC: 0.65 [95%CI: 0.53–0.77]	Delaying the development of DPN.
Irene et al. [60]	Diabetic foot ulcers	sudomotor function test (SFT)	Sensitivity of 83.33%, Specificity of 50.47%, AUC: 0.776 [95% CI: 0.702–0.849] Sensitivity of 100%, Specificity of 31.53%, AUC: 0.816 [95% CI: 0.757–0.874]	SFT could be added in a care setting

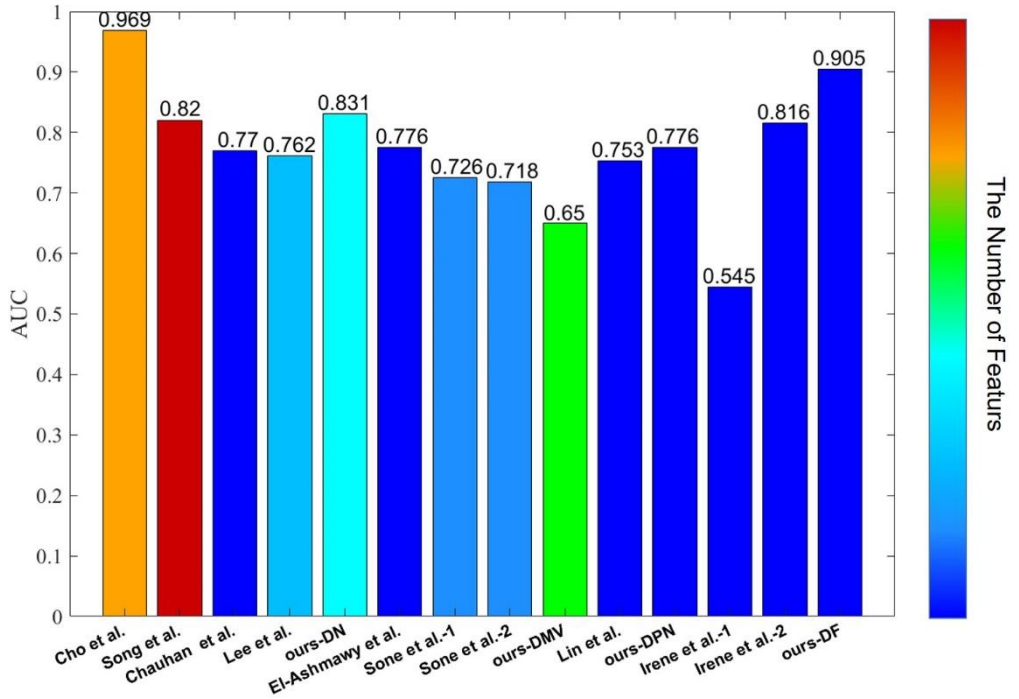

Figure 9: Comparison with AUC of different studies.

4. Discussion

Following the results explored above, the BN model in this study is beneficial and useful to aid diabetic patients in preventing from four kinds of complications and to make self-care with just a few features that could be obtained in general medical examination. Moreover, we could instantiate some variables in the MB of every complication variable to attain more information about them.

First, we focus on the DN variable. It is worth noting that the people aged between 20 and 55 are more likely to have the DN complications with the probability of 18.5%, whereas the probability is under 14% in the other age categories. In addition, if the Cr variable has an abnormal value, the probability for DN will increase from 13.9% to above 40% (refer to Figure 10). When the TRIG variable is instantiated to Chyle, Figure 10 shows the highest probability of DN variable of 41.8%. Besides the impact of feature variables, DMV variable is also associated with DN variable. When the DMV is in positive state, there is a probability of 4.4% at having DN at the same time.

Consequently, diabetic patients could pay more attention to Cr and TRIG, especially for young people aged between 20 and 55.

Because the positive cases of DF are below 3% of the sample size, the probability of DF in positive state is only 0.59%. HDL is the only variable that needs to be paid attention to, which in high condition will lead to a rise to 3.43% in the probability of DF (refer to Figure 11).

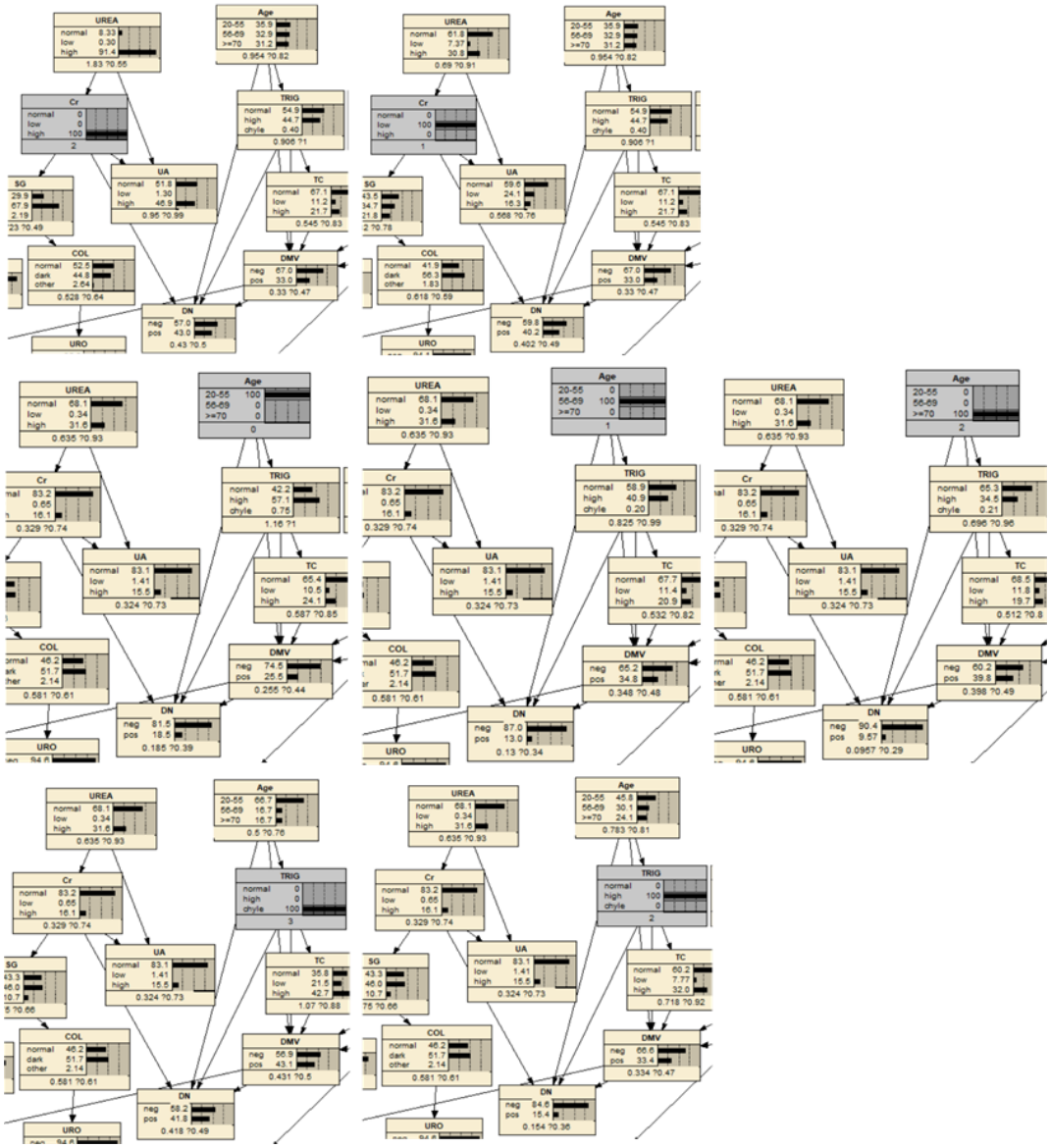

Figure 10: Part of the T2DM complications model focus on DN variable and its MB.

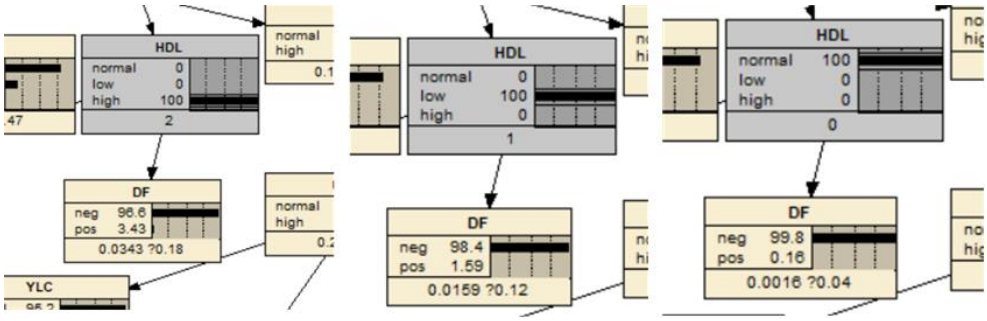

Figure 11: Part of the T2DM complications model focus on DF variable and its MB.

Note that for DMV, there is a probability of 39.8% in the positive state of DMV in the people aged over 70 years old, which is the group of the highest probability. Furthermore, if TC, Na, HDL variables are in High state and TRIG variable is in Chyle state respectively, there is always a rise in the probability of DMV. As we can see in Figure 12, the state of DN variable has no impact on DMV when the TC, Na, HDL variables are in High state, TRIG variable is in Chyle state, and Age variable is in the category of 56-69 state or over 70 whereas DK in the positive condition could have negative relationship with the DMV variable. In other words, it is just a little possibility for an elder to have DK and DMV at the same time when TC, Na, HDL, TRIG are all in a poor state. Diabetic patients who are in the condition would not focus on the DN and Cr variables. Then, both of U-GLU and U-KET variables could have a positive influence on DMV when DK is in negative condition.

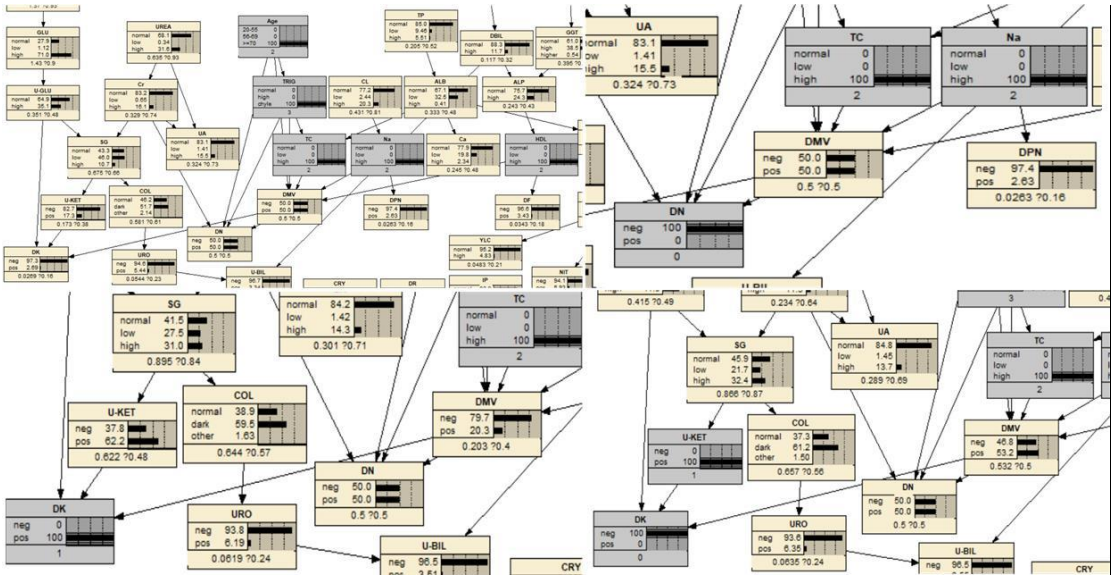

Figure 12: Part of the T2DM complications model focus on DMV variable and its MB.

For the DK variable, U-KET is the most important factor that the probability of DK increases from 3.24% to 11.6% when the state of U-KET is in positive. U-GLU also has a positive impact on DK variables (refer to Figure 13).

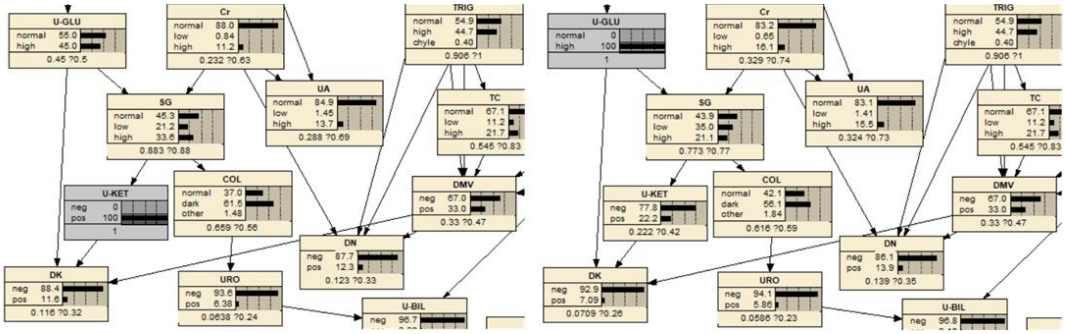

Figure 13: Part of the T2DM complications model focus on DK variable and its MB.

The data we used to build the model is from the inpatients who have already been diagnosed with various kinds of T2DM complications followed the principle of authenticity and professional characteristic, which means that the warning factors in BN model might be applied to self-monitoring of T2DM patients and the assistant treatment of T2DM complications. The model built in this paper aims to predict the condition of complications with warning features. Therefore, the analysis for warning factors of T2DM complications may be helpful to avoid or limit as much as possible two situations: (i) patients worried about T2DM complications may have access to ambulatory visit and health care services unnecessarily too often, and (ii) patients with no realization wait too long before they go to hospital and complications may occur.

The one probable application scenario is that patients with T2DM use warning factors to predict the probability of T2DM complications. There is a certain threshold for each outcome variable in the model to distinguish the positive state and negative state, which means the highest accepted negative value. If the conditional probability of one of the complications is higher than the threshold when patients instantiate the warning factors, they are more likely to have had complications. It is advised for patients to see the doctor and do the relevant examination so that they could detect the complications or treat the disease as early as possible.

In addition, it might be beneficial for diabetic patients to focus on the warning features and use warning features to monitor their physical condition related to

complications before entering the hospital. If the warning features are at a normal level but close to the threshold value of the high or low, taking actions to control them around the average normal level might be beneficial. A diabetic patient, for example, finds that his HDL is 1.5mmol/L close to the threshold of high state. It might be good for him to do some exercise or have a more healthy diet to decrease the level of HDL which is related to DMV.

Furthermore, the warning features could be obtained easily in medical examinations, which is convenient for diabetic patients to do self-management in daily life.

5. Conclusions

Analysis of warning factors of T2DM complications is necessary and significant. In this study, the warning factors of T2DM complications including urine test data, glycated hemoglobin (HbA1c) and biochemical test data with MB based on a BN model were found out and analyzed, and T2DM complications were predicted in a T2DM cohort. It was learned from the dataset related to complications of T2DM provided by PLAGH. According to the missing value, we did multiple imputations with PMM and split the dataset into the training set and testing set with 10-fold cross-validation. For different variables, we set two, three or four thresholds.

Based on the prior knowledge, the structure of the BN model was built with Bootstrap and Tabu search algorithm merging data information and expert knowledge, which made a strong foundation of warning factors analysis. In addition, parameters of the model are learned with MLE and 10-fold cross-validation was used to learn 10 times. The MB is used to select the warning features.

We also compared the performance of BN model and BN without prior information model using warning factors with the Naïve Bayes model, the Random Forest model, and C5.0 Decision Trees model using all other variables in prediction from the perspective of AUC, 95%CI, sensitivity and specificity. Finally, the two BN models

predicting the warning factors of outcome variables was proved the best. Then the comparison with other experiments was also carried out, the result of which indicated that the prediction in DN, DF, DMV and DK variables with warning factors was practically significant based on the dataset. Moreover, we made inferences of outcome variables with warning factors and reported the context of potential clinical assistant treatment of T2DM complications.

Due to the limitation of sources and ways of collecting data, we could not get the further dataset on which we could make a prediction and perform the assessment. Besides, the method used in imputation which created correlations between samples leading to independent folds could be prompted. In terms of future directions, we intend to use more T2DM complication cases to test our model and train our model again and again to analyze warning factor more deeply, and improve our health management system in predicting tasks with these warning factors. The methods of processing missing data also needed to be explored further to make model building and assessment more reasonable. In addition, more features that are easy to be accessed to will be considered in our model to make predictions with warning features more convenient and reliable.

Conflicts of interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (key program) (Grant number 71532002, 61702023), a major project of the National Social Science Foundation of China with grant number 18ZDA086, and a key project of Beijing Social Science Foundation Research Base (Grant number 18JDGLA017).

REFERENCES

[1] J. Singh, D. Kumar, D. Baleanu, On the analysis of fractional diabetes model with exponential law, Adv. Differ. Equ-ny. 2018 (2018).

[2] Y. Seino, K. Nanjo, N. Tajima, T. Kadowaki, A. Kashiwagi, E. Araki, C. Ito, N. Inagaki, Y. Iwamoto, M. Kasuga, T. Hanafusa, M. Haneda, K. Ueki, Report of the Committee on the Classification and Diagnostic Criteria of Diabetes Mellitus, J. Diabetes Invest. 1 (2010) 212-228.

[3] H.Y. Chen, H.C. Pan, Y.C. Chen, Y.C. Chen, Y.H. Lin, S.H. Yang, J.L. Chen, H.T. Wu, Traditional Chinese medicine use is associated with lower end-stage renal disease and mortality rates among patients with diabetic nephropathy: a population-based cohort study, BMC Complem. Altern. M. 19 (2019).

[4] C.M. Steppan, S.T. Bailey, S. Bhat, E.J. Brown, R.R. Banerjee, C.M. Wright, H.R. Patel, R.S. Ahima, M.A. Lazar, The hormone resistin links obesity to diabetes, Nature, 409 (2001) 307-312.

[5] I. Blesneac, A.C. Themistocleous, C. Fratter, L.J. Conrad, J.D. Ramirez, J.J. Cox, S. Tesfaye, P.R. Shillo, A.S.C. Rice, S.J. Tucker, D.L.H Bennett, Rare Nav1.7 variants associated with painful diabetic peripheral neuropathy, Pain, 159 (2018) 469-480.

[6] P.M. Seferović, M.C. Petrie, G.S. Filippatos, et al. Type 2 diabetes mellitus and heart failure: a position statement from the Heart Failure Association of the European Society of Cardiology, Eur. J. Heart. Fail. 20 (2018) 853-872.

[7] L. Wang, T. He, A.D. Fu, et al., Hesperidin enhances angiogenesis via modulating expression of growth and inflammatory factor in diabetic foot ulcer in rats, Eur. J. Inflamm. 16 (2018).

[8] A. Ramaesh, Incidence and long-term outcomes of adult patients with diabetic ketoacidosis admitted to intensive care: A retrospective cohort study, J. Intens. Care Soc. 17 (2016) 222-233.

[9] P. Ghosh, A.D. Kandhare, K.S. Raygude, V.S. Kumar, A.R. Rajmane, M. Adil, S.L. Bodhankar, Determination of the long term diabetes related complications and cardiovascular events using UKPDS risk engine and UKPDS outcomes model in a

representative western Indian population, Asian Pac. J. Trop. Dis. 2 (2012) S642-S650.

[10] A.P. Hills, R. Arena, K. Khunti, et al, Epidemiology and determinants of type 2 diabetes in south Asia, Lancet Diabetes Endo. 6 (2018) 966-978.

[11] D. Chapman, R. Foxcroft, L. Dale-Harris, H. Ronte, F. Bidgoli, S. Bellary, Insights for Care: The Healthcare Utilisation and Cost Impact of Managing Type 2 Diabetes-Associated Microvascular Complications, Diabetes Ther. 10 (2019) 575-585.

[12] F.R. Mancini, A. Affret, C. Dow, B. Balkau, H. Bihan, F. Clavel-Chapelon, M.C. Boutron-Ruault, F. Bonnet, G. Fagherazzi, Educational level and family structure influence the dietary changes after the diagnosis of type 2 diabetes: evidence from the E3N study, Nutr. Res. 44 (2017) 9-17.

[13] R.B. Pachori, M. Kumar, P. Avinash, K. Shashank, U.R. Acharya. An improved online paradigm for screening of diabetic patients using RR-interval signals, J. Mech. Med. Biol. 16 (2016).

[14] X. Gómez-Batiste, M. Martínez-Muñoz, C. Blay, J. Espinosa, J.C. Contel, A. Ledesma, Identifying needs and improving palliative care of chronically ill patients: a community-oriented, population-based, public-health approach, Curr. Opin. Support Pa. 6 (2012) 371-378.

[15] S.Z. Cui, D.J. Wang, Y.Z. Wang, P.W. Yu, Y.C. Jin, An improved support vector machine-based diabetic readmission prediction, Comput. Meth. Prog. Bio. 166 (2018) 123-135.

[16] M. Tayefi, M. Tajfard, S. Saffar, P. Hanachi, A.R. Amirabadizadeh, H. Esmaeily, A. Taghipour, G.A. Ferns, M. Moohebati, M. Ghayour-Mobarhan, hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm, Comput. Meth. Prog. Bio. 141 (2017) 105-109.

[17] D. Yin, Y. Yan, N. Xu, Y. Hui, G.J. Han, N. Ma, C.H. Yang, G.F. Wang, Predictive values of obesity categories for cardiovascular disease risk factors in Chinese adult population, J. Cell. Biochem. 120 (2019) 7276-7285.

[18] B.H. Cho, H. Yu, K.W. Kim, T.H. Kim, I.Y. Kim, S.I. Kim, Application of irregular

and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods, Artif Intell Med. 42 (2008) 37-53.

[19] S. Ananthi, V. Bhuvaneswari, Prediction of heart and kidney risks in diabetic prone population using fuzzy classification, Coimbatore: ICCCI. (2017).

[20] R.K.K. Leung, Y. Wang, R.C.W. Ma, A.O.Y. Luk, V. Lam, M. Ng, W.Y. So, S.K.W. Tsui, J.C.N. Chan, Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype warning patterns in diabetic kidney disease: A prospective case-control cohort analysis, BMC Nephrol. 14 (2013) 1-9.

[21] K. Chauhan, D.A. Verghese, V. Rao, L.L. Chan, C.R. Parikh, S.G. Coca, G.N. Nadkarni, Plasma endostatin predicts kidney outcomes in patients with type 2 diabetes, Kidney Int. 95 (2019) 439-446.

[22] H.M. El-Ashmawy, H.S. Roshdy, Z. Saad, A.M. Ahmed, Serum endostatin level as a marker for coronary artery calcification in type 2 diabetic patients, J. Saudi Heart Assoc. 31 (2019) 24-31.

[23] R. Jiménez-Lucena, O.A. Rangel-Zuniga, J.F. Alcala-Diaze, et al. Circulating miRNAs as Predictive Biomarkers of Type 2 Diabetes Mellitus Development in Coronary Heart Disease Patients from the CORDIOPREV Study, Mol. Ther – Nucl. Acids, 12 (2018) 146-157.

[24] S.B. Lee, Y.S. Kim, J.H. Kim, K. Park, J.S. Nam, S. Kang, J.S. Park, S. Shin, C.W. Ahn, Use of RBC deformability index as an early marker of diabetic nephropathy, Clin. Hemorheol. Micro. (2018).

[25] Y.H. Tang, S.M.C. Pang, M.F. Chan, G.S.P. Yeung, V.T.F. Yeung, Health literacy, complication awareness, and diabetic control in patients with type 2 diabetes mellitus, J. Adv. Nurs. 62 (2010) 74-83.

[26] D. Heckerman, Bayesian Networks for Data Mining, Data Min. Knowl. Disc. 1 (1997) 79-119.

[27] S. Nadkarni, P.P. Shenoy, A Bayesian network approach to making inferences in causal maps, Eur. J. Oper. Res. 128 (2001) 479-498.

[28] J. Tian, B. Yu, D. Yu, et al. Missing data analyses: a hybrid multiple imputation

algorithm using Gray System Theory and entropy based on clustering, J. Appl. Intell. 40 (2014) 376-388.

[29] B. Pester, T. Lehmann, L. Leistritz, H. Witte, C. Ligges, Influence of imputation strategies on the identification of brain functional connectivity networks, J. Neurosci. Meth. 309 (2018) 199-207.

[30] N. Bhakta, Q. Liu, K.K. Ness, et al. The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE), Lancet, 390 (2017) 2569-2582.

[31] H. Heitjan, R.J.A. Little, Multiple imputation for the Fatal Accident Reporting System, J. R. Stat. Soc. Series C-Appl. 40 (1991) 13-29.

[32] B.F.M. Wijnen, S.L. Schat, K.R.J.A. De Kinderen, A.J. Colon, P.P.W. Ossenblok, S.M.A.A. Evers, Burden of disease of people with epilepsy during an optimized diagnostic trajectory: costs and quality of life, Epilepsy Res. 146 (2018) 87-93.

[33] A. Marshall, D.G. Altman, R.L. Holder, Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study, BMC Med. Res. Methodol. 10 (2010).

[34] D.B. Rubin, Statistical matching using file concatenation with adjusted weights and multiple imputations, J. Bus. Econ. Stat. 4 (1986) 87-94.

[35] R.J.A. Little, Missing-data adjustments in large surveys, J. Bus. Econ. Stat. 6 (1988) 287-296.

[36] T. Fushiki, Estimation of prediction error by using K-fold cross-validation, Stat. Comput. 21 (2011) 137-146.

[37] F. Sambo, B.D. Camillo, A. Franzin, A. Facchinetti, L. Hakaste, J. Kravic, G. Fico, J. Tuomilehto, L. Groop, R. Gabriel, T. Tuomi, C. Cobelli, A Bayesian Network analysis of the probabilistic relations between risk factors in the predisposition to type 2 diabetes, Milan: EMBC. (2015) 2119-2122.

[38] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian Network Classifiers, Mach. Learn. 29 (1997) 131-163.

[39] F.V. Jensen, Bayesian Networks and Decision Graphs, Technometrics, 45 (2007) 178-179.

[40] P. Fusterparra, P. Tauler, M. Bennasar-Veny, A. Ligeza, A.A. Lopez-Gonzalez, A. Aguilo, Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk, Comput. Meth. Prog. Bio. 126 (2016) 128-142.

[41] S.L. Lauritzen, N.A. Sheehan, Graphical Models for Genetic Analyses, Stat. Sci. 18 (2003) 489-514.

[42] C. Bielza, P. Larranaga, Discrete Bayesian Network Classifiers: A Survey, ACM Comput. Surv. 47 (2014) 1-43.

[43] J.R. Vergara, P.A. Estevez, A review of feature selection methods based on mutual information, Neural Comput. Appl. 24 (2014) 175-186.

[44] K. Yu, X.D. Wu, W. Ding, Y. Mu, H. Wang, Markov Blanket Feature Selection Using Representative Sets, IEEE T. Neur. Net. Lear. 28 (2017) 2775-2788.

[45] M.D. Sharma, J.A. Farmer, A. Garber, Type 2 diabetes and cardiovascular risk factors, Curr. Med. Res. Opin. 27 (2011) 1-5.

[46] N.M. Doliba, A.M. Babsky, M.D. Osbakken, The Role of Sodium in Diabetic Cardiomyopathy, Front. Physiol. 9 (2018).

[47] L.L. Yang , Q.M Li, X.M. Liu, S.G. Liu, Roles of Voltage-Gated Tetrodotoxin-Sensitive Sodium Channels NaV1.3 and NaV1.7 in Diabetes and Painful Diabetic Neuropathy, Int. J. Mol. Sci. 7 (2016).

[48] G.P. Fadini, A. Solini, M.L. Manca, G. Zatti, I. Karamouzis, A. Di Benedetto, L. Frittitta, A. Avogaro, Phenotyping normal kidney function in elderly patients with type 2 diabetes: a cross-sectional multicentre study, Acta Diabetol. 55(2018) 1121-1129.

[49] F. Glover, M. Laguna, Tabu Search, Gen. Inform. 106 (1997) 221-225.

[50] S. Marini, E. Trifoglio, N. Barbarini, F. Sambo, B. Di Camillo, A. Malovini, M. Manfrini, C. Cobelli, R. Bellazzi, A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes, J. Bio. Inform. 57 (2015) 369-376.

[51] Z.W. Ji, Q.B. Xia, G.M. Meng, A Review of Parameter Learning Methods in Bayesian Network, Fuzhou: ICIC. 2015.

[52] G.M. Huang, K.Y. Huang, T.Y. Lee, J.T.Y. Weng, An interpretable rule-based

diagnostic classification of diabetic nephropathy among type 2 diabetes patients, BMC Bioinformatics, 16 (2015).

[53] S. Dubrava, J. Mardekian, A. Sadosky, E.J. Bienen, B. Parsons, M. Hopps, J. Markman, Using Random Forest Models to Identify Correlates of a Diabetic Peripheral Neuropathy Diagnosis from Electronic Health Record Data, Pain Med. 18 (2017) 107-115.

[54] M. Toussi, J.B. Lamy, P. Le Toumelin, A. Venot, Using data mining techniques to explore physicians therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes, BMC Med. Inform. Decis. 9 (2009).

[55] J. Myerson, L. Green, M. Warusawitharana, Area under the curve as a measure of discounting, J. Exp. Anal. Behav. 76 (2013) 235-243.

[56] Ling C X, Jin H, Zhang H. AUC: a better measure than accuracy in comparing learning algorithms, Halifax: Canadian Society for Computational Studies of Intelligence Conference on Advances in Artificial Intelligence, 2671 (2003) 329-341.

[57] X. Song, L.R. Waitman, Y. Hu, A.S.L. Yu, D. Robins, M. Liu, Robust clinical marker identification for diabetic kidney disease with ensemble feature selection, J. Am. Med. Inform. Assn. 26 (2019) 242-253.

[58] H. Sone, S. Tanaka, S. Tanaka, et al. Comparison of Various Lipid Variables as Predictors of Coronary Heart Disease in Japanese Men and Women With Type 2 Diabetes: Subanalysis of the Japan Diabetes Complications Study, Diabetes Care, 35 (2012) 1150-1157.

[59] X.P. Lin, L.L. Xu, D.Q. Zhao, Z.Y. Luo, S.Y. Pan, Correlation between serum uric acid and diabetic peripheral neuropathy in T2DM patients, J. Neurol. Sci. 385 (2018) 78-82.

[60] I. Sanz-Corbalán, J.L. Lázaro-Martínez, E. García-Morales, R. Molines-Barroso, F. Alvaro-Afonso, Y. Garcia-Alvarez, Advantages of early diagnosis of diabetic neuropathy in the prevention of diabetic foot ulcers, Diabetes Res. Clin. Pr. 146 (2018) 148-154.