DBN-Extended: A Dynamic Bayesian Network Model Extended With Temporal Abstractions for Coronary Heart Disease Prognosis

Kalia Orphanou, Athena Stassopoulou, and Elpida Keravnou

***Abstract*—Dynamic Bayesian networks (DBNs) are temporal probabilistic graphical models that model temporal events and their causal and temporal dependencies. Temporal abstraction (TA) is a knowledge-based process that abstracts raw temporal data into higher level interval-based concepts. In this paper, we present an extended DBN model that integrates TA methods with DBNs applied for prognosis of the risk for coronary heart disease. More specifically, we demonstrate the derivation of TAs from data, which are used for building the network structure. We use machine learning algorithms to learn the parameters of the model through data. We apply the extended model to a longitudinal medical dataset and compare its performance to the performance of a DBN implemented without TAs. The results we obtain demonstrate the predictive accuracy of our model and the effectiveness of our proposed approach.**

***Index Terms*—Coronary heart disease (CHD), dynamic Bayesian networks (DBNs), medical prognostic models, temporal abstraction (TA), temporal reasoning.**

I. Introduction

TEMPORAL abstraction (TA) and dynamic Bayesian networks (DBNs) became two topics of much interest and research in clinical systems. TA is a heuristic process that provides short, informative, context-sensitive summaries of temporal data. TA techniques enable us to overcome the asynchronicity problem, and in many instances, the missing values problem which is very common in real temporal data. The derived high-level abstract concepts proved to be helpful in various clinical tasks and domains. TA techniques were utilized in various medical systems for summarizing and managing patients' data [1]–[3].

DBNs are temporal extensions of Bayesian networks, which are graphical models representing explicitly probabilistic relationships among variables [4]. They model stochastic processes in discrete time and their state changes through time in a sequence of time slices. DBNs were applied to medicine in tasks such as medical diagnosis, forecasting, and medical decision making [4]. In addition, they are often utilized as clinical prognostic models, due to their abilities of representing the temporal nature of the medical problem and of interpreting explicitly the prediction outcome. Examples of DBN prognostic models are the van Gerven model [5] that predicts the future state of patients with a carcinoid tumor and the Pittsburg cervical cancer screening model (PCCSM) [6] used for the prediction of the risk of cervical precancer and cancer for patients undergoing cervical screening. A detailed survey on TA and DBN applied to medicine can be found in our recent work in [7].

Manuscript received August 9, 2014; revised December 3, 2014 and February 15, 2015; accepted March 27, 2015. Date of publication April 6, 2015; date of current version May 9, 2016.

K. Orphanou and A. Stassopoulou are with the Department of Computer Science, University of Cyprus, Nicosia 1678, Cyprus (e-mail: korfan01@cs.ucy.ac.cy; stassopoulou.a@unic.ac.cy).

E. Keravnou is with the Department of Electrical and Computer Engineering and Computer Science, Cyprus University of Technology, Limassol 3603, Cyprus (e-mail: Rector@cut.ac.cy).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2015.2420534

A. Our Contribution

In this paper, we present a novel approach of integrating TA techniques with DBNs. Our recent review of the relevant literature [7] indicated that both these areas were largely used independently of each other in clinical domains. Our proposal is that they could be effectively integrated in the context of medical decision-support systems, and more specifically, for developing a prognostic model. We apply this integration in the medical domain of coronary heart disease (CHD), a disease generally caused by atherosclerosis—when plaque (cholesterol substances) builds up in the arteries. The proposed model, called "DBN-extended" predicts the risk of a particular patient suffering a CHD event during a particular time period t, based on the patient's past medical history (up to time $t-1$).

The CHD domain has been an attractive target for medical expert systems. A number of models have been introduced in the literature for CHD prognosis using various classification techniques such as decision trees, neural networks, naive Bayes, support vector machine, logistic regression, and regression trees to predict the CHD disease [8], [9]. Although these techniques achieve very good performance, they do not utilize the patient medical history and its influence on the disease progression over time in order to predict a potential CHD event. Contrary to such classification techniques, the key benefits of applying our proposed model to the CHD domain, are that the high degree of uncertainty inherent in the problem can be addressed, in conjunction to taking into consideration the patient history. In particular, the DBN combines prior medical knowledge with observed patient data to output a probability distribution over all classes, and not just a classification label. Moreover, the extended model is able to represent the temporal interactions among high-level temporal concepts (TAs), which represent risk factors (RFs) of CHD. Given the RFs of some patient, the model can infer the probability of a potential CHD event occurring.

More generally, the aim of our proposed approach is to enhance the functionality of a DBN by introducing TAs into its nodes, for given medical domains and tasks. Consequently, in

our experimental analysis, we compare the performance of the proposed DBN-extended model against the DBN without TAs.

In particular, we use TA methodologies to extract basic abstractions (i.e., state, single trend, and persistence). The derived concepts are then used for DBN model development and deployment. Learning parameters and inference algorithms are applied to the constructed model. In order to facilitate a comparison of our proposed model with alternative methodologies, we also implemented a DBN model representing time point data rather than high-level TAs. We evaluate the performance of both models: the proposed DBN extended and the DBN without TAs for the purpose of predicting the probability of CHD. The performance of our proposed DBN-extended model as well as the results of the comparison with the alternative approach demonstrate the effectiveness of our proposed model.

This paper is structured as follows: In Section II, we provide an overview of our approach and our evaluation dataset. The methodology for deriving the temporal basic abstractions is described in Section III and the proposed DBN-extended model is presented in Section IV. In Section V, we present the DBN model without TAs. An extensive discussion of our experiments and the comparison of the performance of both models is introduced in Section VI. We conclude in Section VII.

II. Overview

TA methods and DBNs have been successfully used independently of each other in many clinical systems, however, they are complementary to each other. TA methods are used to create high-level temporal concepts from time-point data while DBNs are used for temporal reasoning, knowledge representation, and decision making (e.g., monitoring, prognosis) under uncertainty.

Our goal is to integrate TA techniques with dynamic Bayesian networks by developing an extended prognostic DBN model. In the proposed DBN-extended model, the nodes represent TAs and the arcs temporal dependencies between the abstractions. A benchmark dataset, STULONG,[1] is used to develop the extended prognostic model in the context of CHD. CHD often develops over years and sometimes it is difficult to be diagnosed before the occurrence of a CHD event.[2] The identification of key RFs would help in detecting the disease in its early phases and prevent the occurrence of future CHD events.

A. Dataset Overview

The STULONG dataset was collected from a longitudinal study of atherosclerosis primary prevention. The target group includes 1417 middle-age men. The number of checkup examinations of a single patient ranges from 1 to 20 (1 to 24 years). The first examination of each patient includes blood pressure measurements, basic anthropometric measurements (e.g., weight and height) and ECG examination. Furthermore, on each examination, patients provided information concerning their diet, physical activity, smoking, and alcohol drinking habits, social characteristics that may affect their daily life stress (e.g., job responsibility) as well as family and personal medical history focusing around cardiovascular diseases. More specifically, the values of 244 attributes were surveyed on the entry examination of each patient, whereas on the following examinations, the values of 66 attributes were surveyed concerning physical and biomedical examination values and any disease findings. The group consisted of both men who had a CHD event in the past and men who did not have a CHD event before the beginning of the study.

B. System Overview

Our approach consists of four main phases:

1) data preprocessing and feature selection;
2) derivation of basic TAs (state, trend, and persistence TAs);
3) construction of the "DBN-extended" model;
4) application of the "DBN-extended" model for prognosis of a CHD event.

The first phase consists of the feature selection process and the selection of the temporal range of observations. We base our selection of features on the domain knowledge that we acquired from a CHD expert. The selected features that are CHD RFs include: hypertension (diastolic and systolic blood pressure), cigarette smoking status (current smoker or not), dyslipidemia levels (such as total cholesterol, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, and triglycerides levels), obesity, diabetes, history features (such as past personal history and family history), age, medicines treating high cholesterol and hypertension, diet (if they follow any diet or not), and exercise (if they regularly exercise or not). The incorporation of these features into the DBN-extended model is explained in the following two subsections.

A key issue for model construction is the choice of the total observation period for all patients, which in this study ranges from 1 to 24 years. By selecting a temporal range of 21 years, we include the majority of patients with CHD event within 19–21 years after their first examination. For patients whose total observation period is less than 21 years, all of the feature values are considered unknown for the years beyond their observation period. The target group is reduced by removing records of patients with less than two years of observations since in this study we are going to focus on the temporal aspect of the data, utilizing the advantage of long-term observations. The final target group consists of 849 individuals out of which 466 were monitored for 21 years.

III. Basic TAs

TA techniques are divided into two categories: basic and complex TAs. In this study, we are concerned with basic TA techniques such as: state, trend, and persistence. One of the assumptions used in deriving TAs (state and trend) is that the abstraction value of a variable with missing raw values at any time within the interval period, is defined to be the same as its last known value. The same applies for cases when no record is defined during the required time interval.

A. State TAs

The state abstractions determine the state of an individual parameter over a particular time period based on predefined categories. The state categories for the selected features

[1]The data resource are on: http://euromise.vse.cz/challenge2004/ [Date accessed: 12 February 2015].

[2]Examples of CHD events are: acute coronary syndrome, myocardial infarction and ischemic heart disease.

TABLE I
STATE TAS VARIABLES AND THEIR STATE VALUES

Variable	Variable Code	Value = 1	Value = 2	Value =3
Smoking	Smoking	Non-smoker	Current Smoker	
Hypertension Medicines	medBP	Taken	Not taken	
Dyslipidemia Medicines	medCH	Taken	Not taken	
Hypertension	HT	No Hypertension	Well-controlled	Poorly Controlled
Dyslipidemia	Dyslipidemia	Absent	Present	
Obesity	Obesity	Absent	Present	
Age	AGE	Normal	High	Very High
Diet	DIET	Following a Diet	Not Following a Diet	
Exercise	Exercise	Exercising	Not Exercising	

TABLE II
TREND TAS VARIABLES AND THEIR TREND VALUES IN A SORTED LIST

Variable Name	Variable Code	Value = 1	Value = 2	Value = 3	Value =4
Low-density lipoprotein cholesterol	LDL	Normal	NormalIncreasing	AbnormalDecreasing	Abnormal
Triglycerides	TRIG	Normal	NormalIncreasing	AbnormalDecreasing	Abnormal
High-density lipoprotein cholesterol	HDL	Normal	NormalDecreasing	AbnormalIncreasing	Abnormal
Total Cholesterol	TCH	Normal	NormalIncreasing	AbnormalDecreasing	Abnormal

(variables) are defined by clinical experts' rules. For example, poorly controlled and well-controlled hypertension are state TAs of systolic and diastolic blood pressure values. The hypertension variable is defined by the "Poorly Controlled" state label if the patient has a history of hypertension and his systolic or diastolic blood pressure levels are above the standard limits; and by the "Well-controlled" state label when a patient has a history of hypertension and his systolic or diastolic blood pressure levels are normal. Otherwise, it is defined by the "No Hypertension" label. Dyslipidemia is a state TA of dyslipidemia values. The Dyslipidemia variable is defined by the "Present" state label when a patient has any of the dyslipidemia values higher than the standard limits and "Absent" otherwise. State TAs for the age variable are derived using three state categories labels: "Normal" when the patient is under 50 years old, "High" when the patient is between 50 and 60 years old, and "Very High" when the patient is over 60 years old. State TAs for the rest of the variables are derived in a similar manner. All state TAs are displayed in Table I.

Trend abstractions of a feature are generated by observing the changes between their values. They are generated by comparing two or more consecutive feature values during the interval period of 3 years (1–3 examinations), and selecting the most frequent trend value for the corresponding feature for that period. When multiple values occur equally frequently, the smallest of those values is selected. The sorted list of trend values is: (1) decreasing, (2) stable, and (3) increasing. We also used a combination of trends and state abstractions in order to define the ratio of change of a particular variable based on its state value. Trend abstraction values are:

1) "Normal" when the variable state value is normal and its trend ratio is decreasing or steady;
2) "NormalIncreasing" when the variable state value is normal and its trend ratio is increasing;
3) "AbnormalDecreasing" when the variable state value is abnormal and its trend ratio is decreasing; and
4) "Abnormal" when the variable state value is abnormal and its trend ratio is increasing or steady.

It should be noted that, contrary to the rest of the variables, HDL is considered a CHD RF when its levels are low, thus its trend values are: "Normal," "NormalDecreasing," "AbnormalIncreasing," and "Abnormal." The resulting trend abstractions are displayed in Table II.

TABLE III
PERSISTENCE TAS VARIABLES AND THEIR PERSISTENCE VALUES

Variable Name	Variable Code	Value = 1	Value = 2
Diabetes	Diabetes	Present	Absent
Past Personal History	HistoryEvent	Present	Absent
Family History	FH	Present	Absent
History of Hypertension	HHT	Present	Absent

B. Persistence TAs

Persistence TA techniques derive maximal intervals for some property by applying persistence rules both backwards and forwards in time from the specific time of the given property. For example, when someone was diagnosed with diabetes at time t, diabetes remains present (persists) from time t onwards. Similarly, when someone was diagnosed with a CHD event at time t, he has a history of event from $t+1$ onwards, thus the value of HistoryEvent variable is "Present" from $t+1$ until the end of the monitoring process. The FH and HT are examples of persistence TAs for the whole representation time period, since their value does not change through time. The resulting persistence TAs are displayed in Table III.

IV. Constructing the DBN

The most popular temporal extension of a BN is the DBN [4], which represents stochastic processes using a discrete-time representation. A DBN is a network with the repeated structure of a static BN for each time slice over a certain interval. A DBN represents the change of variable states at different time points. A node in a DBN represents a stochastic variable (temporal process) and its possible states. A node can be either a hidden node, whose values are never observed, or an observed node (with a known value). Arcs represent the local or transitional dependencies among variables. Intra-slice arcs represent the dependencies within the same time slice like in an atemporal BN. Inter-slice arcs connect nodes between time slices and represent their temporal evolutions. A DBN is assumed to be time invariant, which means that the network structure per time slice and across time slices does not change. Furthermore, it is assumed that a DBN uses the Markovian property: the conditional probability distribution of each variable at time t for all $t > 1$, depends only on the parents from the same time slice or from the previous time slice but not from earlier time slices. The construction of a DBN consists of two steps: (1) building the network structure (qualitative part) and (2) learning the parameters of the network (quantitative part).

A. DBN-Extended Network Structure

The network structure, as displayed in Fig. 1, was designed by incorporating prior information elicited from medical experts and medical literature [10], [11]. The derived basic TAs described in Section III form the nodes (variables) of our DBN. The DBN framework enables us to combine all the observations of a patient and predict a probability for the hypothesis that the patient will suffer with a future CHD event, given the values of all the observed nodes.

The model consists of 19 variables, which constitute the nodes of the DBN, out of which 17 are observed and 2 are hidden. Hidden variables are: the class attribute Pred_Event, representing the occurrence of a CHD event in the last three years of the total observation period (21 years), and the Dyslipidemia node. Both of these variables take two values: "Present" and "Absent." The Pred_Event variable is a terminal node, represented outside of the temporal network and it only connects to its parents in the last time slice of the unrolled network. It represents the occurrence of a future CHD event, thus it is not repeated in every time slice but its value is inferred at the end of the inference process [12]. In order to simplify the parameter estimation process [13], Dyslipidemia is introduced as a common parent node of TCH, HDL, LDL, TRIG, which are indirect RFs to the class attribute and CurrentEvent. Dyslipidemia is a common cause of TCH, HDL, LDL, TRIG, and CurrentEvent at each time slice. At time slice $t = 5$, Dyslipidemia is a common cause of TCH, HDL, LDL, TRIG, and the class variable. If there is no evidence or information about Dyslipidemia (hidden), then the presence of one or more symptoms (TCH, LDL, HDL, TRIG) will increase the chances of Dyslipidemia, which in turn will increase the probability of the effects CurrentEvent and Pred_Event.

The variable family history (FH) is not repeated since it was modeled only as an initial condition and it is not changing over time. It is, therefore, shown in the network of Fig. 1, to be outside the temporal plate. The single digit numbers on the arcs denote the temporal delay of the influence of the cause node to the effect. For example, an arc labeled as 1 between the variables history of CHD (HistoryEvent) and itself denotes that the patient's CHD history at the current time slice t is influenced by the patient's past CHD history (at time slice $t-1$). On the other hand, the arc without label connecting the CHD RFs (hypertension, obesity, etc.) to the CHD event, denotes an instantaneous influence at the same time slice. The first time slice ($t = 0$) in the network represents the time period starting from the patient's entry examination and ending three years after their entry examination. The finest granularity in the dataset is one year. Three years is the time interval period selected for the derived abstractions, since at least two examinations are needed in order to have any abstractions and most cases do not have examinations on an annual basis.

B. DBN-Extended Learning Parameters

Having defined the structure of the network, we need to define the conditional probabilities that quantify the arcs of the network. Since the structure of a DBN is invariant for all times $t \in \{0, \ldots, 5\}$, the parameters of the network are fixed through all the time slices and divided into two categories: (1) prior probabilities for the root nodes (such as DIET and Exercise, for example) and (2) the transition probabilities for the non root nodes. For example: $\Pr(\text{Obesity}_t|\text{Obesity}_{t-1}, \text{DIET}_t, \text{Exercise}_t)$. In other words, the conditional probability distribution for Obesity depends on the current status of Diet and Exercise, as well as on the Obesity in the previous time step.

All of the parameters were learned from data using the expectation maximization (EM) algorithm [14]. The EM consists of two steps: (1) the expectation step, E, which estimates the expected likelihoods over completions of missing data based on the values of the observed data and (2) the maximization step, M, where parameters are estimated by maximizing expected log-likelihoods found in step E.

Once the network structure is defined and the network is quantified with the learned conditional probability distributions, the next step is to predict the probability of the class node: Pred_Event. Each variable in the network is instantiated with the corresponding feature value. The DBN is unrolled for six time slices $t = [0, \ldots, 5]$ in order to represent the observation period of 21 years. Then, it performs prognosis and derives the belief in the class variable Pred_Event, which represents CurrentEvent at $t = 6$. More specifically, the model derives:

$$P(\text{Pred\_Event}|\text{Smoking}_5, \text{HT}_5, \text{Dyslipidemia}_5, \text{Obesity}_5, ...)$$

that is, the probability of having a CHD event (or not) given the evidence at the last time slice.

The following section describes the DBN without TAs, constructed to facilitate a comparison with the proposed DBN-extended model.

V. Constructing a DBN without TAs

Contrary to the DBN-extended model, a standard DBN model without TAs is developed to represent time-points rather than

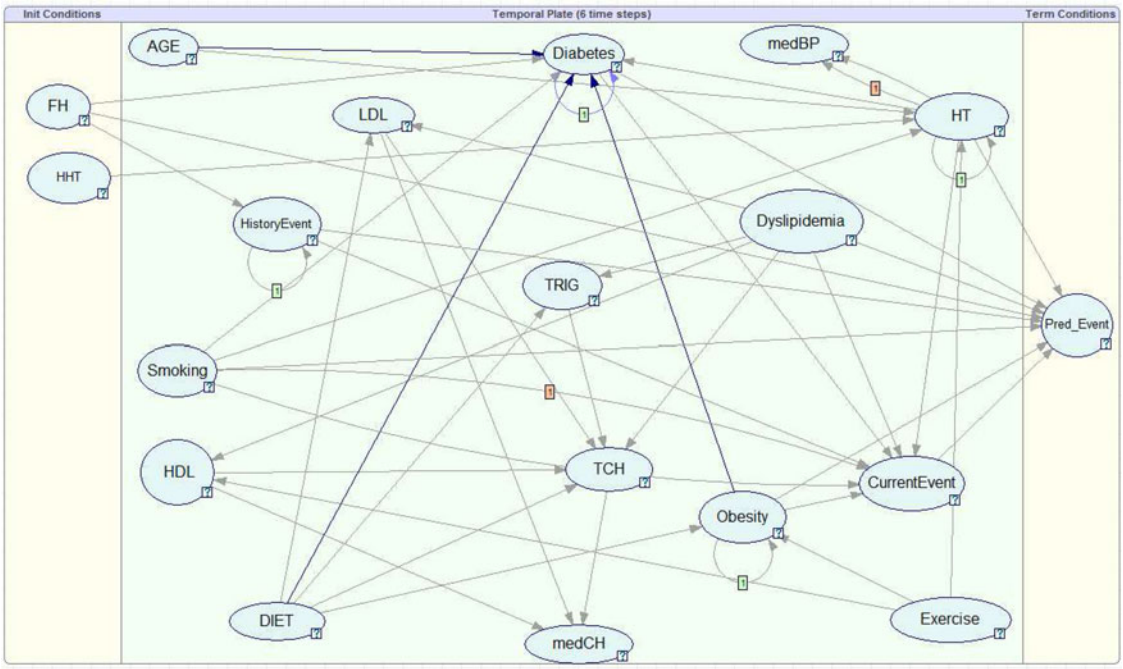

Fig. 1. Structure of the DBN-extended model representing only the basic TAs. An arc labeled as 1 between the variables denotes an influence that takes one time step

high-level TAs. Applying this model to the STULONG dataset, each time slice of the network represents a patient examination. Although the total number of examinations of all patients is 22, most of the patients had a CHD event during or before their 13th examination. Consequently, the chosen total temporal range for this network is 13 examinations. The total number of patients who had been monitored for 13 or more examinations is 478 over which 24 had a CHD event on their 13th examination. We use the EM algorithm to handle such missing data. The data preprocessing phase involves three steps:

1) feature selection;
2) handling missing values;
3) discretization process.

The selected features are RFs of CHD, as with the DBN-extended system. The main difference is the removal of features that represent high-level concepts such as hypertension, history of CHD event and obesity, and the addition of new features representing time-point events such as body mass index (BMI), systolic and diastolic blood pressure.

To deal with missing values, the patient's health condition is assumed to remain stable during any time period that their examination results are unknown. We also discretized the values of the continuous features in the dataset, using a discretization technique for clinical domains, called "domain-based discretization." Using a domain-based technique, the values of each continuous attribute are divided into a particular number of bins based on clinical domain knowledge [11], [15]. All the selected features and the resulting discretization features values of the continuous features are displayed in Table IV.

As displayed in Fig. 2, the model consists of 19 variables, out of which 17 are observed and 2 are hidden. The hidden nodes are the Pred_Event and Dyslipidimea as with the DBN-extended network. The parameters of the network are learned from data using the EM algorithm [14]. The DBN is unrolled for 12 time slices: $t = [0, 1, \ldots, 11]$. It performs prognosis and derives the belief in the class variable given all the observed data (evidence) from the last time slice, i.e.

$$P(\text{Pred\_Event}|\text{Smoking}_{11}, \text{SystBP}_{11}, \text{DiastBP}_{11}...)$$

VI. Experimental Results and Analysis

In this section, we present the experiments performed in order to apply our methodology and evaluate the performance of our "DBN-extended" model against the performance of a DBN model without TAs using the tenfold cross-validation method. Cross-validation methods [16] are widely used to evaluate the performance of predictive models using statistical analysis. In k-fold cross validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the testing data for testing the model, and the remaining $k-1$ subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the testing data. The k results from the folds can then be averaged to produce a single estimation of the model accuracy.

A. Training in the Presence of Class Imbalance

One important problem in the data mining field is to deal with imbalanced datasets. The datasets present a class imbalance when there are many more examples of one class (majority class) than of the other (minority class). It is usually the case that this latter class, i.e., the minority class, is the class of interest. Because this minority class is rare among the

TABLE IV
FEATURES SELECTED FOR THE NON-TA MODEL AND DISCRETIZATION OF CONTINUOUS FEATURES

Variable	Variable Code	Value = 1	Value = 2	Value = 3
Smoking	Smoking	Non-smoker	Current Smoker	
Medicines for Reducing Cholesterol	medCH	Taken	Not taken	
Medicines for Reducing Blood Pressure	medBP	Taken	Not taken	
Systolic Blood Pressure	SystBP	Normal: <120	Prehypertension: [120–140]	High Blood Pressure: >140
Diastolic Blood Pressure	DiastBP	Normal: < 80	Prehypertension: [80–90]	High Blood Pressure: >90
Dyslipidemia	Dyslipidemia	Absent	Present	
Disease	Current_Event	Absent	Present	
Predict CHD	Pred_Event	Absent	Present	
Diabetes	Diabetes	Absent	Present	
Family History	FH	Absent	Present	
History of CHD	HHD	Absent	Present	
Body Mass Index	BMI	Normal Weight: < 25	Overweight:[25–30]	Obesity: >30
Low-density lipoprotein cholesterol	LDL	Normal: < 100 mg	High: [100–160]mg	Very High: >160 mg
Triglycerides	TRIG	Normal: <150 mg	High:[150–200]mg	Very High: > 200 mg
High-density lipoprotein cholesterol	HDL	Normal: < 40 mg	High: [40–60]mg	Very High: > 60 mg
Total Cholesterol	TCH	Normal: < 200 mg	High:[200–240]mg	Very High: >240 mg
Age	AGE	< 50 years	[50–60] years	> 60 years
Diet	DIET	Following a Diet	Not Following a Diet	
Exercise	Exercise	Exercising	Not Exercising	

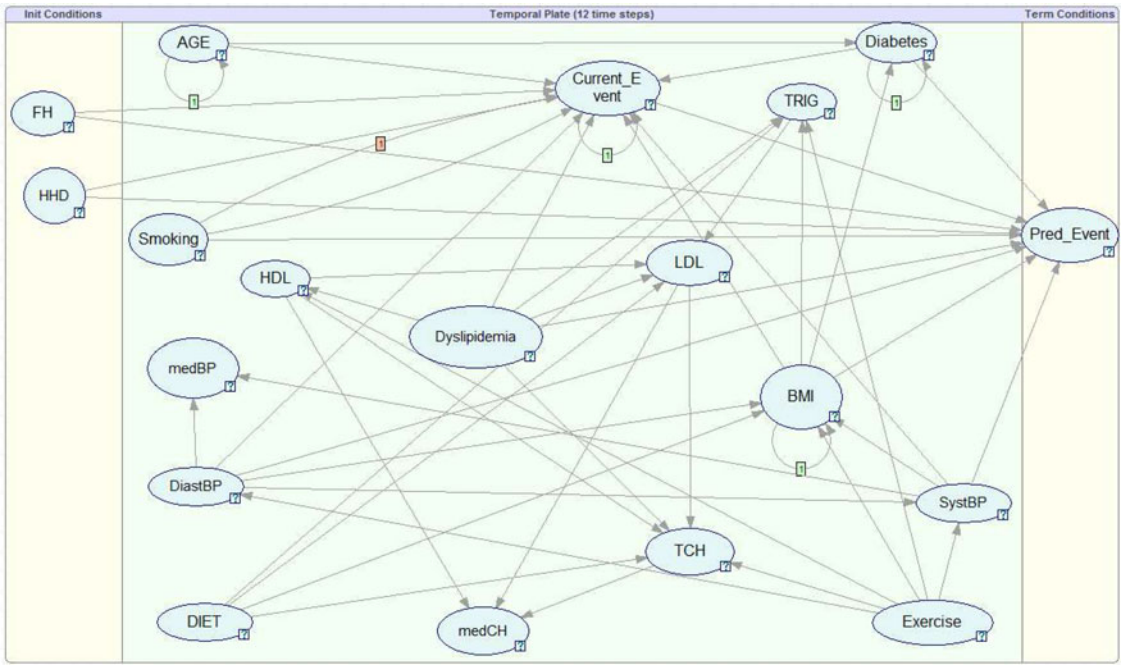

Fig. 2. DBN model representing low-level data (without TAs)

general population, the class distributions are highly skewed. This causes the classification methods to perform poorly on the minority class examples. In the current dataset, individuals who did not suffer a CHD event (majority class) at a particular time period are many more than those who suffered a CHD event (minority class).

One approach to tackle the problem of an imbalanced training dataset is to use resampling to modify the training dataset [17]. This is achieved by either removing examples from the majority class (undersampling) or adding more examples to the minority class (oversampling) or a combination of both. In our system, we evaluate our classifier by applying two oversampling methods as well as a combination of oversampling with undersampling on the training dataset. More specifically, we apply the following resampling methods: (1) SMOTE-N (Synthetic Minority Oversampling Technique for nominal features), which generates synthetic examples to be added to the minority class, (2) random oversampling where minority cases are randomly chosen for duplication, (3) a combination of SMOTE-N oversampling with clustering undersampling [18]. Clustering undersampling uses the *k*-means algorithm to divide the majority class samples into K ($K > 1$) clusters, and then, make K subsets of majority class samples, where each cluster is considered to be one subset of the majority class. All the subsets of the

TABLE V
NO OF RECORDS FOR EACH CLASS FOR ONE FOLD OF CROSS VALIDATION ON EACH OF THE TRAINING DATASETS OF DBN-EXTENDED MODEL

Number of Records with Class Value	Dataset D1	Dataset D2 & D3	Dataset D4 & D5
Present	54	366	183
Absent	366	366	183
Missing	383	383	383
Total	803	1115	749

TABLE VI
NO OF RECORDS FOR EACH CLASS FOR ONE FOLD OF CROSS VALIDATION ON EACH OF THE TRAINING DATASETS OF DBN-WITHOUT TAS MODEL

Number of Records with Class Value	Dataset D1	Dataset D2 & D3	Dataset D4 & D3
Present	22	409	145
Absent	409	409	145
Missing	371	371	371
Total	802	1189	661

TABLE VII
PERFORMANCE FOR ALL FIVE DBN-EXTENDED MODELS CORRESPONDING TO THE FIVE TRAINING DATASETS

Evaluation Metrics	C1	C2	C3	C4	C5
Precision	0.8555	0.5966	0.5057	0.5090	0.7207
Recall	0.1167	0.6667	0.5167	0.6000	0.75
F_1 score	0.2053	0.6297	0.5111	0.5508	0.7351

majority class are separately combined with the minority class samples to make K different training datasets. In our implementation, K is chosen to be equal to 2, and finally, (4) Random oversampling on the minority class combined with clustering undersampling the majority class.

We obtain the following training datasets for each classifier:

1) *Training dataset 1 ($D1$):* No resampling. In this experiment, the training data are not altered.
2) *Training dataset 2 ($D2$):* Random oversampling of the minority cases until we got 1:1 ratio of the minority class to the majority class.
3) *Training dataset 3 ($D3$):* Oversampling using the SMOTE-N technique. We applied oversampling to the minority class cases until we got 1:1 ratio of the minority class to the majority class.
4) *Training dataset 4 ($D4$):* Oversampling using the SMOTE-N technique and undersampling. We applied clustering undersampling to the majority class cases, and then, we applied SMOTE-N until we got 1:1 ratio.
5) *Training dataset 5 ($D5$):* Oversampling using random oversampling technique and clustering undersampling. We applied clustering undersampling, and then, we applied random oversampling to the minority class cases until we got 1:1 ratio.

We constructed five DBN-extended networks, one for each experiment. The networks had the same structure but differed in their parameters, i.e., prior probabilities and the conditional probability tables according to the respective dataset: each time a new training dataset was introduced, new network parameters were derived using training on the new set. Throughout the remaining of this paper, we will refer to the five models as: $C1$, $C2$, $C3$, $C4$, and $C5$ corresponding to the five datasets: $D1$, $D2$, $D3$, $D4$, and $D5$, respectively.[3] The resulting number of records for each training dataset of the DBN-extended model are displayed in Table V.

TABLE VIII
PERFORMANCE FOR ALL FIVE DBN WITHOUT TAS MODELS CORRESPONDING TO THE FIVE TRAINING DATASETS

Evaluation Metrics	M1	M2	M3	M4	M5
Precision	0.9265	0.5115	0.4329	0.5502	0.6350
Recall	0.0833	0.6250	0.3750	0.2917	0.333
F_1 score	0.1529	0.5626	0.4019	0.3812	0.4372

For the DBN without TAs model, we obtain the same training datasets, as for the DBN extended and we construct five networks, one for each experiment, which we will refer to as: $M1$, $M2$, ..., $M5$. The resulting number of records for each training dataset is displayed in Table VI.

B. Testing and Evaluation

In order to evaluate the prognostic performance of our developed models, we adopted metrics that are commonly applied to imbalanced datasets: precision, recall, and the F_1 score as defined in (1)–(3), respectively. The F_1 score summarizes the two metrics into a single number in a way that both metrics are given equal importance. Recall and precision should be close to each other, otherwise the F_1 measure yields a value closer to the smaller of the two.

$$\text{Precision}(P) = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}} \tag{1}$$

$$\text{Recall}(R) = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}} \tag{2}$$

where

$$\text{TP} = \text{\# positive examples correctly predicted as positive}$$
$$\text{FP} = \text{\# negative examples wrongly predicted as positive}$$
$$\text{FN} = \text{\# positive examples wrongly predicted as negative}$$
$$F_1 \text{ score} = 2 \times \frac{P \times R}{P + \text{R}} \tag{3}$$

In our domain, positive classification is a CHD event to be present at $t = 6$, which is when the derived probability of this event is higher than 0.5, i.e.

$$P(\text{Pred\_Event} = \text{Present}|E) > 0.5$$

[3]The models presented in this paper were created and tested using the SMILE inference engine and GeNIe available at: https://dslpitt.org/genie/ [Date accessed: 12 February 2015].

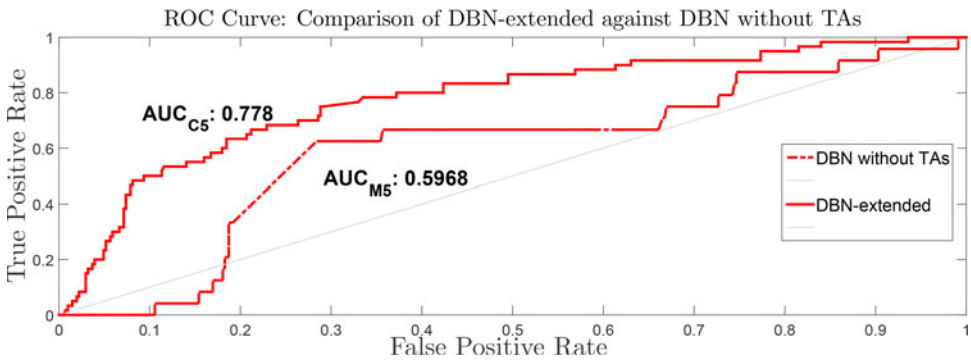

Fig. 3. ROC curves comparing the two prognostic models on dataset $D5$

Applied to our problem, the aforementioned equations are translated into

$$\text{Precision} = \frac{\#\text{ patients correctly predicted to have CHD event}}{\#\text{ patients predicted to have CHD event}}$$

$$\text{Recall} = \frac{\#\text{ patients correctly predicted to have CHD event}}{\#\text{ patients with an actual CHD event}}$$

To avoid or minimize skew-biased estimates of performance, we normalize the performance scores of precision and F_1 score by normalizing, to a given degree of skew (target skew ratio = 1), the true negative (TN) and the false positive (FP) [19]. TN corresponds to the number of negative examples correctly predicted (4), whereas FP corresponds to the number of negative examples wrongly predicted as positive (5).

$$\text{TN} = \text{TN} * \frac{\text{Target Skew Ratio}}{\text{Original Skew Ratio}} \tag{4}$$

$$\text{FP} = \text{FP} * \frac{\text{Target Skew Ratio}}{\text{Original Skew Ratio}} \tag{5}$$

where $\text{Original Skew Ratio} = \frac{\#\text{of negative instances}}{\#\text{positive instances}}$

The values of precision, recall and F_1 score obtained from the evaluation of DBN-extended model for each of the five training datasets are given in Table VII. As it can be seen from Table VII, our proposed system yields promising results with precision reaching as high as 72%, and recall reaching as high as 75% and a combined F_1 score of 74%. These are our best results derived by training the system with dataset $D5$, which applied random oversampling combined with clustering undersampling. The lowest F_1 score is obtained when we train the system with the original, highly imbalanced dataset $D1$ (no resampling). As expected, this yields a low recall due to the high false negative (FN), as the classifier fails to recognize the minority cases (CHD present) and are instead classified as majority cases (CHD absent). In many medical domains, such as CHD, recall is more important than precision since the false negative diagnosis (i.e., failure to predict a CHD event) has much more serious consequences on the patient than a false positive.

The values of precision, recall and F_1 score obtained from the evaluation of the DBN model without TAs, for each of the five training datasets are given in Table VIII. As it can be seen from Table VIII, our proposed extended DBN model outperforms the prognostic DBN model without TAs, and further supports our belief that DBNs can be effectively integrated with TAs.

C. Receiver Operating Characteristic (ROC) Curves

We also use ROC curves to compare graphically the predictive performance of our DBN-extended model against the performance of the DBN without TAs model. ROC [20] is a 2-D graph in which the true positive rate (TPR) is plotted on the *y*-axis and the false positive rate (FPR) on the *x*-axis. TPR is the fraction of positive examples predicted correctly (i.e., same as recall defined above), whereas FPR is the fraction of negative examples predicted as positive. It displays graphically the tradeoff between TPR and FPR of the predictor. A point (*x*, *y*) on the ROC curve represents an (FPR, TPR) pair associated with the prediction based on a given discrimination threshold. The threshold refers to the cutoff value above which a record is predicted as positive and it corresponds to the posterior probabilities generated by the DBN prognostic model. The lowest threshold is the lowest posterior probability of the class attribute $P(\text{Pred\_Event} = \text{Present}|\text{Obesity}, \text{FH}, \text{CurrentEvent}...)$ given to a record. By varying the threshold, we produce different points on the ROC curve (i.e., different (FPR,TPR) pairs). Fig. 3 shows the ROC curves of each classifier developed with the dataset with the best performance, $D5$. A good prediction model is located as close as possible to the upper left corner of the diagram, i.e., point (TPR = 1, FPR = 0).

The area under the ROC curve (AUC) [20] provides another approach of evaluating which model is better. If the model is perfect, its AUC would equal 1, whereas if the model performs random guessing, then its AUC would be equal 0.5. A classifier that is much better than another would have a larger AUC. As displayed in Fig. 3, DBN-extended model has larger AUC than the DBN without TAs model, which proves that the prognostic performance of our proposed model is much better than a standard DBN model.

VII. Conclusion and Future Work

In this paper, we represented an extended DBN whose nodes represent TAs. We utilize the proposed model in the context of CHD prognosis. The benefits of applying our approach to the CHD prognosis are that this extension can handle incomplete data values in predicting disease outcomes and in dealing with

uncertainty that are the most common challenges in the domain of CHD. The interpretation of the obtained probabilistic results can give insight as to how causal dependences and temporal relationships between abstract concepts may influence the risk of a future CHD event.

During our training and evaluation stages, we addressed the class imbalance problem on both training and testing datasets. We used four techniques of resampling to deal with imbalance in the training dataset: random oversampling, SMOTE-N, a combination of SMOTE-N with clustering undersampling, and a combination of random oversampling with clustering undersampling. We compared the performance of our proposed model with the standard DBN, by developing five different networks for each prognostic model. Based on the aforementioned four resampling techniques, the higher predictive results of the DBN-extended model compared to a DBN without TAs prove the effectiveness of our proposed methodology and provide a promising direction for future work.

In addition, we are currently investigating the introduction of complex TAs to the DBN-extended model and the need to use irregular time intervals may arise. Another issue for future work is to assess the robustness of the proposed model against the chosen cutoff values for deriving the TAs.

REFERENCES

[1] N. Lavrač, I. Kononenko, E. Keravnou, M. Kukar, and B. Zupan, "Intelligent data analysis for medical diagnosis using machine learning and temporal abstraction," *AI Commun.*, vol. 11, no. 3,4, pp. 191–218, 1998.

[2] M. Kahn, L. Fagan, and L. Sheiner, "Combining physiologic models and symbolic methods to interpret time-varying patient data," *Methods Inform. Med.*, vol. 30, no. 3, pp. 167–178, Aug. 1991.

[3] Y. Shahar, and M.A. Musen, "Résumé: A temporal-abstraction system for patient monitoring," *Comput. Biomed. Res.*, vol. 26, pp. 255–273, 1993.

[4] T. Charitos, L. C. van der Gaag, S. Visscher, K. A. M. Schurink, and P. J. F. Lucas, "A Dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 1249–1258, Mar. 2009.

[5] M. van Gerven, P. Lucas, and T. van der Weide, "A generic qualitative characterization of independence of causal influence," *Int. J. Approximate Reasoning*, vol. 48, no. 1, pp. 214–236, 2008.

[6] R. Austin, A. Onisko, and M. Druzdzel, "The Pittsburgh cervical cancer screening model: A risk assessment tool," *Archives Pathol. Lab. Med.*, vol. 134, no. 5, pp. 744–750, 2010.

[7] K. Orphanou, A. Stassopoulou, and E. Keravnou, "Temporal abstraction and temporal Bayesian networks in clinical domains: A survey," *Artif. Intell. Med.*, vol. 60, no. 3, pp. 133–149, 2014.

[8] V. S. H. Rao and M. N. Kumar, "Novel approaches for predicting risk factors of atherosclerosis," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 1, pp. 183–189, Jan. 2013.

[9] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 366–374, 2008.

[10] P. W. Wilson, R. B. DAgostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.

[11] R. Conroy, K. Pyörälä, A. E. Fitzgerald, S. Sans *et al.*, "Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project," *Eur. Heart J.*, vol. 24, no. 11, pp. 987–1003, 2003.

[12] J. Hulst, "Modeling physiological processes with Dynamic Bayesian networks," Master's thesis, Delft University of Technology, Delft, Netherlands, 2006.

[13] F. V. Jensen, *An Introduction to Bayesian Networks*, vol. 210. London, U.K.: Univ. College London Press, 1996.

[14] T. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996.

[15] J. Gennest and P. Libby, "Lipoprotein disorders and cardiovascular disease," *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine, 9th ed. Philadelphia, PA, USA: Saunders Elsevier*, 2011.

[16] M. W. Browne, "Cross-validation methods," *J. Math. Psychol.*, vol. 44, no. 1, pp. 108–132, 2000.

[17] A. Stassopoulou and M. D. Dikaiakos, "Web robot detection: A probabilistic reasoning approach," *Comput. Netw.*, vol. 53, no. 3, pp. 265–278, 2009.

[18] M. M. Rahman and D. Davis, "Cluster based under-sampling for unbalanced cardiovascular data," presented at the World Congr. Engineering., London, U.K., vol. 3, 2013.

[19] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Proc. IEEE Humaine Assoc. Conf. Affective Comput. Intell. Interaction*, 2013, pp. 245–251.

[20] T. Fawcett, "An introduction to ROC analysis," *Pattern Recog. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

Kalia Orphanou is a Ph.D. candidate at the Department of Computer Science, University of Cyprus, Nicosia, Cyprus. She received an M.Eng. degree in Computer Science with Artificial Intelligence from the University of Southampton, U.K., in 2009.

Her main research interests include temporal reasoning, artificial intelligence in medicine, probabilistic graphical models, dynamic Bayesian networks, temporal abstraction, machine learning and data mining.

Athena Stassopoulou received the B.Sc. degree in Computer Science and Mathematics (joint Hons.) from the University of Manchester, Manchester, U.K. in 1992 and the Ph.D. degree in Artificial Intelligence, from the University of Surrey, Centre for Vision Speech and Signal Processing, U.K., in 1996.

She is a Professor of Computer Science at the University of Nicosia where she served as the founding Head of the CS Department (2001–2008). She worked as a Postdoctoral Researcher and as a Senior Research Associate at the Center for Mapping, The Ohio State University, USA. She has more than 20 years of research experience and has worked in projects funded by NASA (Image Understanding Initiative), the National Imagery and Mapping Agency in the USA, the European Union and the Research Promotion Foundation of Cyprus. She has been publishing in international refereed journals and conferences in the following areas which constitute her research interests: uncertain reasoning, Bayesian networks, geographic information systems, computer vision and image understanding, machine learning, neural networks and more recently, in artificial intelligence applications for the web.

Elpida Keravnou-Papailiou received the B.Tech. degree in Computer Science and the Ph.D. degree in Cybernetics from Brunel University, West London, in 1982 and 1985, respectively.

She is the first Rector of the Cyprus University of Technology, assuming her duties on January 4th 2012. She started her academic career from the Department of Computer Science, University College London as a Lecturer in 1985–1992, a Senior Lecturer 1991–1992, and the Director of the M.Sc. course in computer science. In 1992, she took up an academic position in the Department of Computer Science, University of Cyprus as an Associate Professor from 1992 to 1996, and a Professor since 1996. At the University of Cyprus, she served as a Vice-Rector for Academic Affairs (2002–2006), as the Dean of the School of Pure and Applied Sciences (1999–2002), and as the first Chairperson of the Department of Computer Science (1994–1998). She also served as the President of the Governing Board of the Cyprus University of Technology (2009–2010). She is currently the Vice-Chair of the Evaluation Committee for Private Universities in Cyprus and a Member of the Governing Board of the European Institute of Innovation and Technology and its Executive Committee (www.eit.europa.eu). She has carried out research in the areas of knowledge engineering, expert systems, deep knowledge models, diagnostic reasoning, temporal reasoning, artificial intelligence in medicine, intelligent data analysis in medicine and hybrid decision support systems. She is an Associate Editor of the scientific journal *Artificial Intelligence in Medicine* (Elsevier) since the launch of the journal in 1989. During the period 2003–2005, she served as the Chairperson of the Artificial Intelligence in Medicine Europe Board.