Lin *et al. BMC Medical Research Methodology* (2023) 23:249
https://doi.org/10.1186/s12874-023-02070-9

BMC Medical Research Methodology

RESEARCH Open Access

A Bayesian network perspective on neonatal pneumonia in pregnant women with diabetes mellitus

Yue Lin[1], Jia Shen Chen[1], Ni Zhong[1], Ao Zhang[1] and Haiyan Pan[1*]

Abstract

Objective To predict the influencing factors of neonatal pneumonia in pregnant women with diabetes mellitus using a Bayesian network model. By examining the intricate network connections between the numerous variables given by Bayesian networks (BN), this study aims to compare the prediction effect of the Bayesian network model and to analyze the influencing factors directly associated to neonatal pneumonia.

Method Through the structure learning algorithms of BN, Naive Bayesian (NB), Tree Augmented Naive Bayes (TAN), and k-Dependence Bayesian Classifier (KDB), complex networks connecting variables were presented and their predictive abilities were tested. The BN model and three machine learning models computed using the R bnlean package were also compared in the data set.

Results In constraint-based algorithms, three algorithms had different presentation DAGs. KDB had a better prediction effect than NB and TAN, and it achieved higher AUC compared with TAN. Among three machine learning modes, Support Vector Machine showed a accuracy rate of 91.04% and 67.88% of precision, which was lower than TAN (92.70%; 72.10%).

Conclusion KDB was applicable, and it can detect the dependencies between variables, identify more potential associations and track changes between variables and outcome.

Keywords Bayesian networks, Neonatal pneumonia, Naive Bayes network, Tree Augmented Naive Bayes model, K-Dependence Bayesian Classifier

Introduction

Gestational diabetes mellitus (GDM) is a common chronic disease of pregnancy that affects the health of tens of millions of women worldwide each year [1, 2]. During pregnancy, women experience disturbances in insulin secretion, leading to abnormal glucose metabolism, persistently elevated blood glucose levels and ultimately gestational diabetes mellitus, which is usually associated with adverse pregnancy outcomes [3, 4]. As well as affecting the mother's own health, GDM can cause adverse pregnancy outcomes such as neonatal pneumonia [5, 6]. According to WHO and the Maternal Child Epidemiology Estimation (MCEE) group, a child will die from pneumonia every 43 s in 2020. Neonatal pneumonia is a serious threat to the health of newborn babies, and the disease can easily progress to respiratory failure or sepsis and other conditions, ultimately leading to neonatal death [7]. Therefore, it is very important to diagnose neonatal pneumonia in pregnant women with diabetes [8]. The current research focuses on the effect

*Correspondence:
Haiyan Pan
GDMU-PH@outlook.com
[1] School of Public Health, Guangdong Medical University, Dongguan 523808, China

of neonatal ventilators on neonatal pneumonia, and the prediction models used are mainly logistic regression model [9].

How to classify efficiently and accurately has always been a problem in disease prediction. Common classification algorithms include Bayesian network [10], K-nearest neighbour algorithm [11] and decision tree [12]. Researchers take advantage of Bayesian network and machine learning methods to predict amyotrophic lateral sclerosis, and Bayesian network produced the best results [13]. However, by applying these models, it is difficult to reveal the potential information of neonatal pneumonia of gestational diabetes, which is a complex disease affected by multiple factors. Although many scholars at home and abroad have studied the effects of gestational diabetes on maternal and infant perinatal outcomes, the method of Bayesian network (BN) model has not been applied to gestational diabetes complicated with neonatal pneumonia.

Bayesian network is an effective hotspot method that has been applied to disease data mining research in recent years. Bayesian network has several advantages that make it a promising tool for these purposes. It is an uncertain causal relationship model that organically combines directed acyclic graph with probability theory, which represents directly and intuitively. Integrating data into the model in the form of conditional probability can deal with various uncertainties and incomplete information. In addition, the Bayesian network can predict the effectiveness of an intervention strategy by introducing new evidence, which is an important and unique advantage over other methods [14].

As it is shown [13, 15], BN is very useful for prediction and diagnosis, which is very important in disease interventions because they are usually expensive and their effects can only be observed in the long term. BN has the properties to be very useful in predicting the effectiveness of different strategies and selecting the best among them. Currently there are many advances in Bayesian classifiers such as Naive Bayes (NB), TAN and so on. Naive Bayes (NB) is one of the simplest and most efficient BNs due to the independence of hypothetical features. The independence hypothesis between features is usually not true, so relaxing the independence hypothesis and expanding the dependency of the Bayesian network has become the main improvement of Naive Bayes, among which the more successful algorithms are Tree-Augmented Naive Bayes (TAN) and k-Dependence Bayesian Classifier (KDB) [16].

In this study, three Bayesian network models, namely Naive Bayes, Tree Augmented Naive Bayes and k-Dependence Bayesian Classifier are used to predict the risk of neonatal pneumonia in pregnancy diabetes. The prediction accuracy and recall rate are compared internally and externally with three machine learning models such as Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM). At the same time, Bayesian networks known as Directed Acyclic Graphs (DAGs) are analysed to find the advantages, disadvantages and scope of Bayesian network models.

Materials and methods

Data

A total of 2008 pregnant women with diabetes who gave birth at Shunde Women and Children's Hospital of Guangdong Medical University between June 2019 and June 2021 were included: 305 pregnant mothers whose babies had neonatal pneumonia (case group), and 1703 who did not have neonatal pneumonia (control group). Written informed consent was obtained from all recruited participants or their legal guardians, and this study was approved by the Ethics Committee of Foshan Women's and Children's Hospital of Guangdong Medical University (SDFYMC001).

Univariate analysis of variables

The chi-squared test was performed on the categorical variables. The 13 variables included in the model were age (age), two-hour postprandial glucose (pbg), parity (p), gestational hypertension (hdop), preterm birth (ptb), preterm rupture of membranes (prom), macrosomia (ms), neonatal respiratory distress syndrome (nrds), neonatal jaundice (nnj), postpartum haemorrhage (pph), neonatal asphyxia (na) and neonatal growth restriction (ngr). Mann–Whitney U test was performed on the rank variables and 4 variables were included in the model: number of pregnancies (g), amniotic fluid volume (afv), amniotic fluid cleanliness (afc) and C-reactive protein (crp).

Software and programs

In the study, our first step is to perform a BN analysis to build a causal graphical model in the form of a directed acyclic graph (DAG), which represents the relationships between all the variables of interest [17]. Bayesian networks can be constructed using structure learning algorithms, which can be categorised into two main groups: constraint-based and score-based methods. In this article, three types of constraint-based learning algorithms were used with the R package bnlearn [18]. In order to ensure that the resulting network was stable, we performed bootstrapping by extracting 1000 samples with replacement, computing a network for each sample, and then averaging them to obtain the resulting network. The study then performed an analysis of the intercorrelation between the feature variables before selecting an appropriate Bayesian method. As expected, Naive Bayes (NB), Tree-augmented Naive Bayes

(TAN) and K-Dependence Bayesian Classifier (KDB) were then selected to build the BN. However, NB is used in the study for parameter learning due to its hypothesis [19, 20]. At the same time, its performance is compared with three types of machine learning methods, Random Forest, SVM and DT. The Netica software package developed by Norsys Software Corporation was used to perform TAN [21, 22]. A schematic diagram of methodology is shown in Fig. 1:

Bayesian network

Bayesian network (BN)

.Bayesian networks (BNs), also known as probabilistic directed acyclic graphs (DAGs), are directed networks accompanied by probabilistic links between edges. A graph is a DAG if all the links (edges) have directions, but none of the nodes go directly to itself or through a path to itself (a circle) [23]. Bayesian networks are able to connect probability distributions on a finite set of random variables. Directed edges represent statistical or causal dependencies between variables [24]. For example, given an edge X → Y, X is the parent node of Y and Y is the child node. Each node, e.g. Xi, has a conditional probability distribution that quantifies the effect of the parent on the child node. In general, the joint conditional probability distribution of any combination of random variables is simplified to formula (1).

$$P(x_1, x_2, x_n) = \prod_{i=1}^{n} P(x_i|Parents(x_i)) \quad (1)$$

The parent nodes of a particular node are its immediate predecessors within the network. These parent nodes of a particular node are its immediate predecessors within the network. These parent nodes are the variables that directly influence the associated node. In Bayesian networks, the term "parent node set" refers to the aggregation of all the parent nodes that influence a particular node. A node has several parents, which together form a set of parents. Ancestor nodes, on the other hand, include all parent nodes, their parent nodes, and so on, forming a lineage of dependencies tracing back to the most distant nodes in the network. The concept of a "Markov Blanket" is the minimal set of nodes that contains all the information necessary to specify the conditional probability distribution of a given node, given its parent and child nodes. The high dimensionality of the data has led to the development of several learning algorithms that focus on reducing computational complexity while still learning the correct network. On the one hand, among several structure learning algorithms [25], constraint-based learning algorithms consist of growth-shrink, fast.iamb and mmpc etc. and it provides a free implementation of some of these structure learning algorithms along with the conditional independence tests and network scores used to construct the Bayesian network. The resulting models are often interpreted as causal models.

Naive Bayes

Naive Bayesian is a simple, stable, easy-to-implement Bayesian algorithm with better classification efficiency based on the assumption that each feature condition

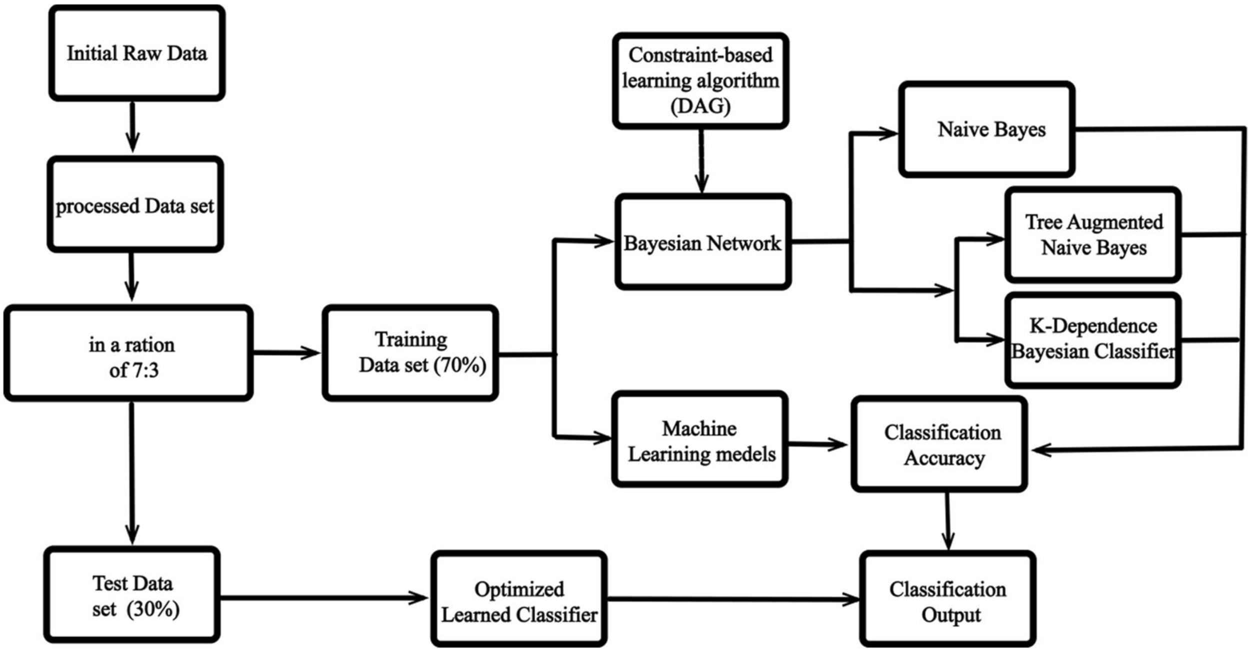


Fig. 1 A schematic diagram of methodology

is independent of each other [19]. Its algorithm is as follows:

① Supposing that $x = \{a_1, a_2, \cdots, a_m\}$ is an item to be classified, and each a_i is a characteristic attribute of x;
② A set of categories $C = \{\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_\mathrm{n}\}$;
③ Calculate the conditional probability of each feature, namely:
$P(y_1|x), P(y_2|x), \cdots, P(y_n|x)$;
④ Take the maximum conditional probability:
$P(y_k|x) = max\{P(y_1|x), P(y_2|x), \cdots, P(y_n|x)\}$;
then $x \in y_k$.

The calculation process of conditional probability is as follows:

① Establish a sample data set as a training sample set.
② Calculate the conditional probability of each eigenvalue under each category
$P(a_i|y_j)(1 \leq i \leq m, 1 \leq j \leq n)$
③ Assuming that all attributes are independent of each other, then according to Bayes' theorem we have drawn formula (2):

$$P(y_1|x) = \frac{P(x|y_i)P(y_i)}{P(x)} \tag{2}$$

Since the denominator is a constant, the largest numerator is needed taken, adding to that each attribute is independent of each other, there are shown in formula (3):

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i)\cdots P(a_m|y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i) \tag{3}$$

When attribute events are independent of each other, the accuracy of Naive Bayesian classification is very good. Figure 2 shows the structure of NB. In reality, each feature variable is often not conditionally independent, but has a dependency relationship, which limits Naive Bayesian classification ability.

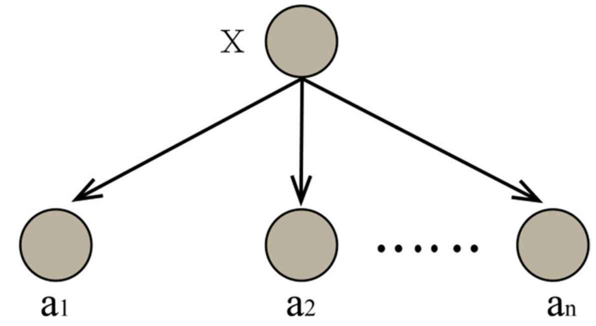


Fig. 2 The structure of NB

Tree Augmented Naive Bayes (TAN)

Tree Augmented Naive Bayes (TAN) is a type of Bayesian network that is an improvement on NB. It assumes that the relationships between attribute variables conform to a qualified tree structure. The basic concept is to break the independence assumption of NB and allow dependencies between categorical variables, but a categorical variable is allowed to have a dependency with at most one other categorical variable. This dependency is represented by a tree structure [26]. Figure 3 shows the structure of TAN.

Construction of Tree Augmented Naive Bayes Network (TAN) contains two parts, structure learning and parameter learning:

① X_j provides information for X_i when C is known, represented by mutual information. Calculate X_i of attribute C according to the data set of training set.

$$I(X_i, X_j|C) = \sum_{x_i,x_j,c} P(x_i, x_j, c) \times log\frac{P(x_i, x_j|c)}{P(x_i|c)P(x_j|c)} \tag{4}$$

In formula (4), $P(x_i, x_j, c)$ represents the probability of occurrence of features x_i 、 x_j, and category c, and logarithms are used to avoid numerical issues. The formula $\log(P(x_i, x_j|c)/P(x_i|c)P(x_j|c))$ is used to calculate the mutual information between features x_i and x_j, which measures their dependence. If features x_i and x_j are independent, then $P(x_i, x_j|c)$ equals $P(x_i|c) \times P(x_j|c)$, and the mutual information is 0. If features x_i and x_j are dependent, then $P(x_i, x_j|c)$ is not equal to $P(x_i|c) \times P(x_j|c)$, and the mutual information is greater than 0.

② The maximum weight span tree is established by using the conditional mutual information of the node pair as the weight of the edge. First, sort the edges according to their weights, and then select the edges in order. After each edge is selected, check whether the tree contains cycles. If it contains cycles, delete the edge and select the next edge. The resulting tree is the maximum weight span tree. Finally, take a node

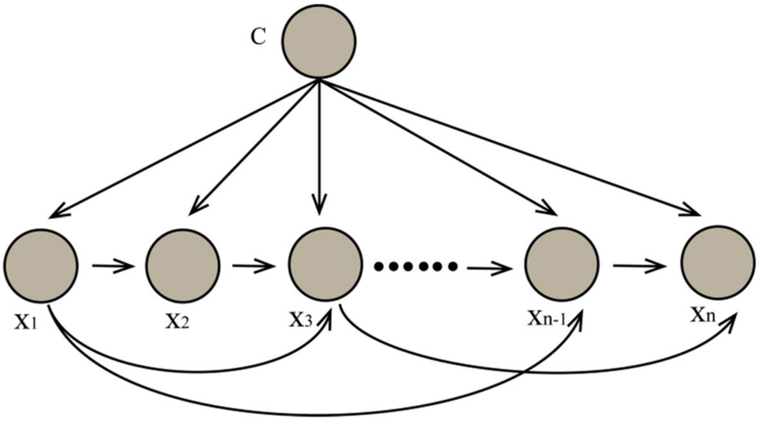


Fig. 3 The structure of TAN

as the starting point, and the direction of leaving the node is used as the direction of the edge in the tree;

③ Add class node as parent node for all attribute nodes;

④ Calculate the joint probability of each classification node to get the TAN classifier.

K-Dependence Bayesian Classifier(KDB)

KDB is also an improvement on the assumption of conditional independence of the Naive Bayes [27, 28]. Compared with TAN and NB, KDB allows high-order conditional dependencies between attributes, further alleviating dependencies between attributes. KDB sort the attributes based on MI $(Xi; Y)$, then add them to the network in turn, and according to CMI $(Xi; X_i|Y)$ selects K attributes as its parent node. As a result, KDB can make better use of the information in the dataset and can perform better. KDB further frees up the limitations of TAN, allowing an attribute to have up to k attributes as its parent node. In general, k(1≦k≦n-1) is determined before building the model, and the structure diagram is shown in Fig. 4, k = 2. However, k can be freely adjusted, which makes KDB extremely flexible and malleable. Assume that the attribute order is $\{X_1,..., X_n\}$, by comparing MI, Xi will choose min(i − 1,k) features with the highest CMI values from the first i − 1 candidates. The joint probability for KDB proves to be formula (5):

$$\boldsymbol{P}_{KDB}(\boldsymbol{x},\ c) = P(c)\prod_{i=1}^{n}P(x_i|c,\pi_{x_i}) \qquad (5)$$

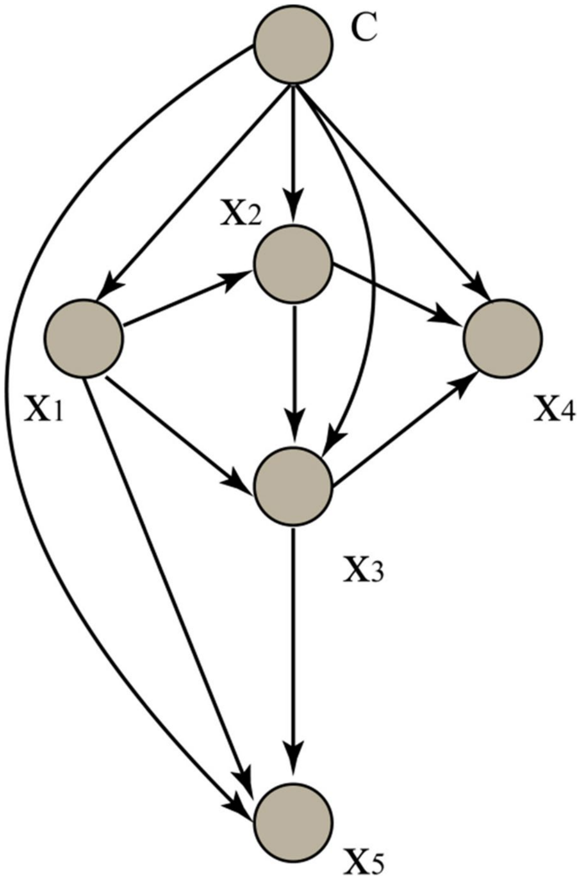


Fig. 4 The structure of KDB (k = 2)

Model evaluation

Firstly, this study uses the changes of node parameter values in two Bayesian network models to evaluate their applicable conditions. Secondly, each sample can be divided into four cases: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) according to the combination of its real category and the prediction category of each model, so that TP, FP, TN, FN represent their corresponding sample cases, and the "confusion matrix" of the classification results can be obtained, as shown in the Table 1.

The authenticity of each model was evaluated by evaluation indicators such as accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Among it, accuracy = (TP + TN)/(TP + FN + FP + TN), sensitivity = TP/(TP + FN), specificity = TN/(FP + TN), positive predictive value = FP/(TP + FP) and negative predictive value = FN/(TN + FN).

Table 1 Confusion matrix

	True value	
	Positive	**Negative**
Predicted value		
Positive	TP	FN
Negative	FP	TN

Results

The results of the constraint-based learning algorithm

Constraint-based learning algorithm. Even while all the algorithms produce network structures that are remarkably similar and agree on the arc direction (as shown in Fig. 5), there are notable differences: The graph's layout shows that mechanics and variables are crucial to the overall assessment of the test, in all models the analysis and statistics scores are conditionally independent of each other. Fast.iamb has 23/14 of directed/undirected arcs while Growth-Shrink has 29/8 of directed/undirected arcs; mmpc learns the underlying structure of the Bayesian network with all the arcs undirected. The red line depicts the variation between fast.iamb and growth-shrink. However, demonstrating the real connections between variables is difficult.

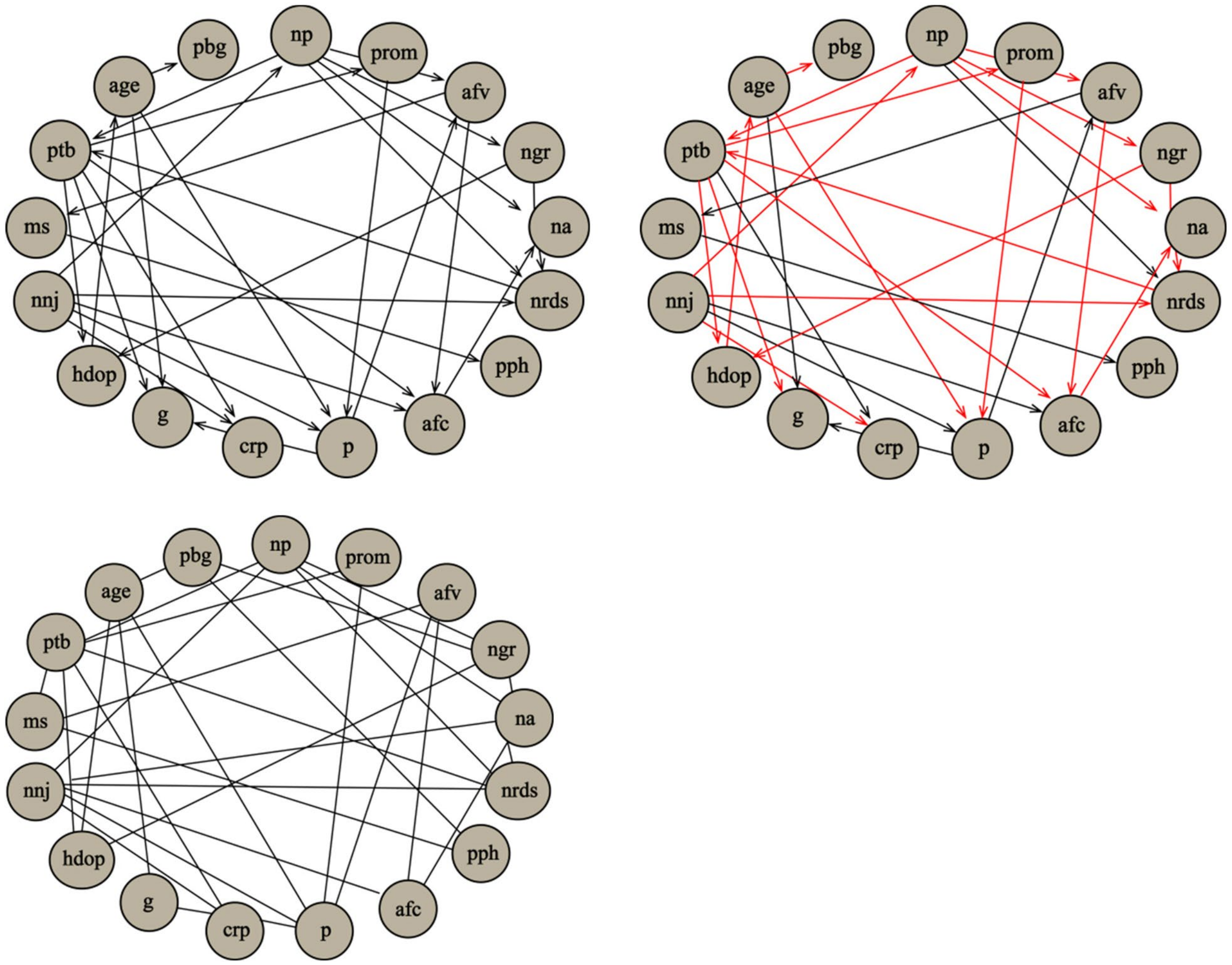


Fig. 5 The Growth-Shrink network structure (top left) and the network structures learned by fast.iamb (top right), mmpc (bottom left)

Building a Naive Bayesian network model

Multivariate correlation analysis was conducted on all the varibales. Then we developed separate NB and evaluate the contributions of ptb, crp, nrds, ngr, afv, and prom in neonatal pneumonia. Factual or reference status is introduced to show the counterfactual condition with or without a patient-specific factor regarding its impact on neonatal pneumonia (node "np") respectively. In the complex correlation analysis model ($np \sim ptb + crp + nrds + ngr + afv + prom$, $R^2 \geqq 0.6$), indicating the goodness of the model as well as np [13]. The relationship among patient-specific factors and neonatal pneumonia in an NB is fixed, and among them nodes represent these factors and neonatal pneumonia, directed arcs denote dependent relationship between each factor and the stage. Although NB is based on the assumption of independence among all features, we build a NB in this dataset by R. When the test set samples (30% of the data) are imported to verify the prediction performance of the model, Naive Bayesian has accuracy rate of 92.20% and achieves the performance with ROC of 94.64%. The results are shown in Table 2. Relatively simple structure may result from underfitting rather than overfitting may help to improve the classification performance of learning algorithm.

Build Tree Augmented Naive Bayes (TAN)

The data was divided into the training set and the test set in a ratio of 7:3 and the Tree Augmented Naive Bayes was learned and tested on Netica software, as shown in Fig. 6.

As we can see, TAN allows a dependency between an attribute variable, so more information can be obtained by looking at the conditional probability table of the node. Take the example of neonatal jaundice (nnj), as

Table 2 Prediction performance of Naive Bayesian

	Prediction					
	Neonatal pneumonia					
Measured	**sick**	**normal**	**recall rate(%)**	**precision rate(%)**	**accuracy rate(%)**	**F1-score(%)**
Neonatal pneumonia						
sick	61	25	73.4	70.9	92.20	72.12
normal	22	495				

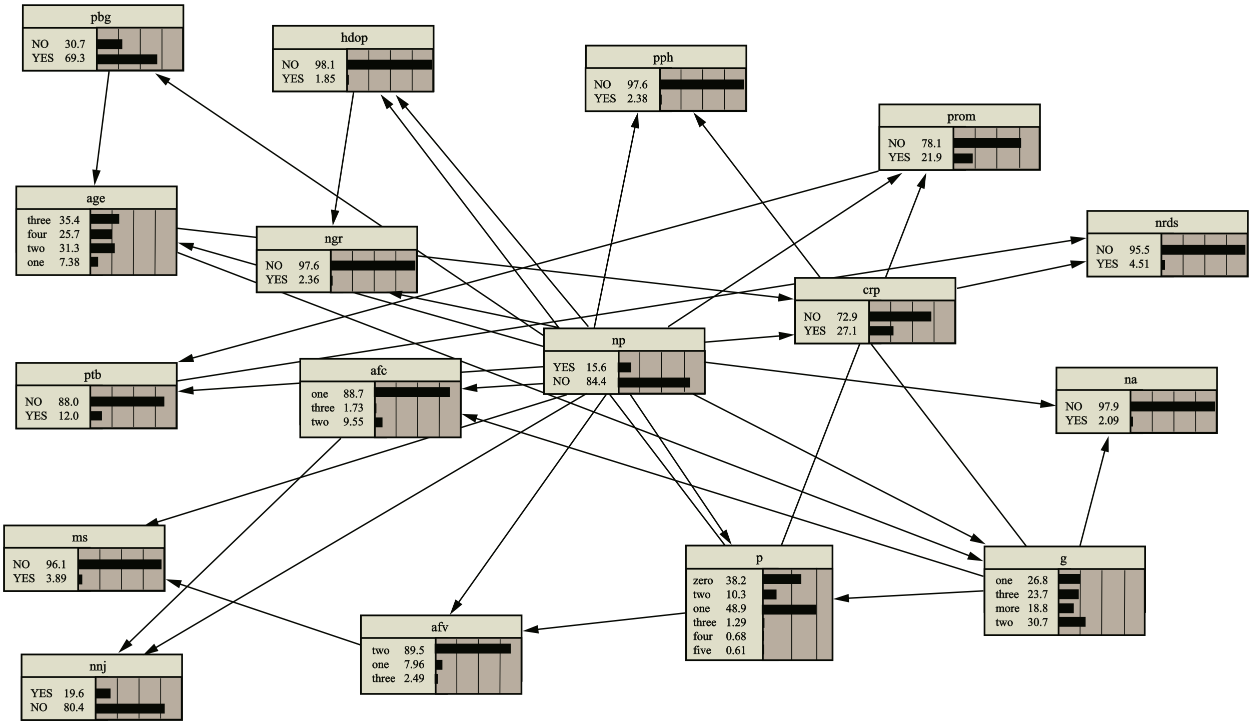


Fig. 6 Tree Augmented Naive Bayes (TAN)

shown in Fig. 7. The data in Fig. 6 represent the corresponding changes in the parameters of neonatal jaundice (nnj) under different conditions of neonatal pneumonia (np) and amniotic fluid cleanliness (afc) in the two parent nodes connected to it. If a patient has neonatal pneumonia (np) and amniotic fluid cleanliness (afc) is the first level, the probability of occurrence of neonatal jaundice (nnj) is 89.7% and the probability of non-occurrence is 10. 3%; if the patient does not develop neonatal pneumonia (np), but amniotic fluid cleanliness (afc) is the third level, the probability of neonatal jaundice is 27.27% and the probability of non-occurrence is 72.72%. Thus, neonates from neonatal jaundice populations are more likely to have amniotic fluid cleanliness and neonatal pneumonia.

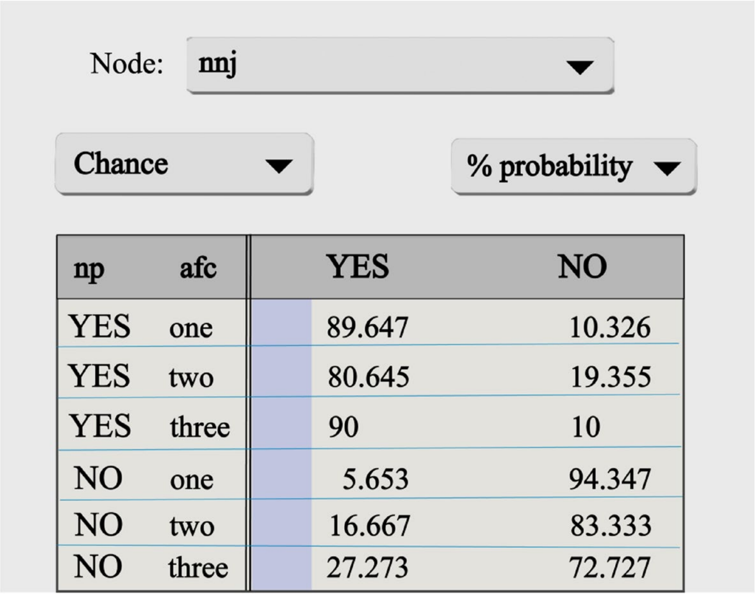
The test set are imported into the model in the TAN to verify the prediction performance. The results are shown in Table 3. TAN has accuracy rate of 92.7%, a recall rate of 68.6% and a precision rate of 74.68%. Besides, TAN achieves the performance with ROC of 93.78%.

Build K-Dependence Bayesian Classifier (KDB)

The algorithms KDB were developed in C++using the NetBeans IDE compiler+GCC. The data was divided into the training set and the test set in a ratio of 7:3. KDB treats training set as a target and build general BN. When $k=2$, KDB can represent $0+1+2\cdots+2=33$ conditional dependencies, while TAN only needs to represent 16 conditional dependencies, shown in Fig. 8. In this data, KDB (0.95 ± 0.09) achieves higher AUC compared with TAN (0.95 ± 0.10). As a result, the KDB does fit the testing instance much better than TAN.

The comparison of modes performance

The data were split 7:3 into training and test sets. Data analysis was performed in the R environment (R Foundation for Statistical Computing, Beagle Scouts, version 4.3.1., https://cran.r-project.org/src/base/R-4, using the KlaR, rRpart, randonForest, pROC and e1071 libraries). Three machine learning methods were used to construct the prediction model for neonatal pneumonia and the predictive performance of these three models is



Node: nnj

Chance

% probability

np	afc	YES	NO
YES	one	89.647	10.326
YES	two	80.645	19.355
YES	three	90	10
NO	one	5.653	94.347
NO	two	16.667	83.333
NO	three	27.273	72.727

Fig. 7 Conditional probabilities of neonatal jaundice (nnj)

Table 3 Prediction performance of TAN

	Prediction					
	Neonatal pneumonia					
Measured	sick	normal	recall rate(%)	precision rate(%)	accuracy rate (%)	F1-score(%)
Neonatal pneumonia						
sick	62	24	75.60	72.10	92.70	73.81
normal	20	497				

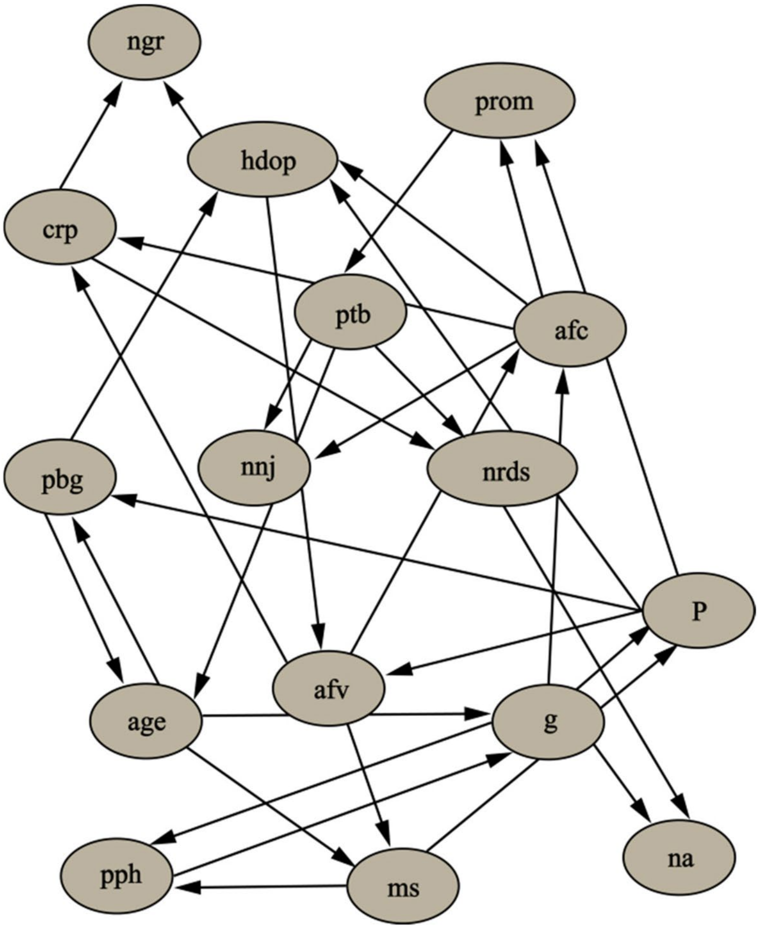


Fig. 8 Conditional dependencies between attributes are shown (KDB)

compared. Table 4 describes the performance of the prediction models considered. The Support Vector Machine (AUC, 0.957) achieved the accuracy rate of 91.04%. However, there is no difference between Decision Tree and Random Forest (AUC, 0.951).

Discussion

In this study, we first applied structure learning algorithms and created a directed acyclic graph (DAG) for neonatal pneumonia. Although it appeared that neonatal pneumonia was more likely to affect linked variables, it provided a new way to understand the aetiology and generate hypotheses about potential causal symptom structures and identify factors that may bridge neonatal pneumonia. Secondly, the influence of the pregnant population on the outcome of neonatal pneumonia was investigated. The naive Bayesian learnt that preterm birth, C-reactive protein, neonatal growth restriction, neonatal respiratory distress syndrome, amniotic fluid volume and premature rupture of membranes have a greater impact on the outcome of neonatal pneumonia. As a result, it has a predictive performance of 92.20% for neonatal pneumonia. On the other hand, Tree Augmented Naive Bayes found that age was associated with

Table 4 Model performance

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
Decision Tree	90.71	69.51	64.77	65.59
Support Vector Machine	91.04	67.88	90.29	77.49
Random Forest	90.71	75.27	67.96	71.43

pregnancies and the index of C-reactive protein affecting neonatal pneumonia. At the same time, pregnancies affect parity and amniotic fluid volume, circularly increasing the risk of macrosomia. The performance of KDB appears to be slightly higher than that of TAN, but the difference is very small. In addition, we were interested in building machine learning models to predict the occurrence of neonatal pneumonia and make a comparison with Bayesian network models. Although the support vector machine showed a better area under the ROC curve, it is still lower than the Bayesian network model [29].

As for the factors of neonatal pneumonia, studies have reported that the blood glucose level of pregnant women can affect fetal development, leading to adverse pregnancy outcomes such as macrosomia and premature rupture of membranes [30–33], as shown in KDB (pbg → age → ms). It can be seen that if the mother continues to have high blood glucose levels, this will affect the health of the newborn. This is because high blood glucose levels can reduce leukocyte phagocytosis and chemotaxis in pregnant women, which increases the risk of urinary tract and reproductive tract infections, thereby increasing the risk of premature rupture of membranes (age → pbg → prom) in the Bayesian network of KDB [34]. Studies have shown that in the gestational diabetes population, the rate of neonatal pneumonia is lower in the well-controlled group than in the poorly controlled group [29]. Specifically, pregnant women and fetuses have a special relationship in which any physical changes in the mother will affect the fetus and even affect its future growth and development (pbg → hdop → ngr in KDB). Neonatal gestational diabetes affects neonatal lung development, neonatal lung maturation and associated lung diseases (np → pbg; np → hdop in TAN), with a high incidence of neonatal pneumonia, neonatal respiratory distress syndrome and bronchopulmonary dysplasia in gestational diabetes patients [34, 35].

Compared with the traditional regression model, the Bayesian network model can handle large sample data. While compared with machine learning with poor interpretability [36], the Bayesian networks built on the coefficients reveal some patterns of disease variables, which have the potential to help diagnose the disease [23].The Bayesian network returns a DAG that identifies the direction of prediction and potential causal influence among factors in the absence of a randomised controlled experiment, but cross-sectional data cannot confirm to the true situation. Furthermore, in a DAG, activation flows in only one direction, never returning to the node of origin. Therefore, there are no important variables influencing associations between risk factors that have been omitted from the DAG. Furthermore, such instability in the direction of the edge may suggest bidirectionality of influence, considering that in 51% of the bootstrapped samples the edge points from factor X to factor Y, and in 49% of the bootstrapped samples from factor Y to factor X [27]. In research, TAN can observe the dependencies between attribute variables and reveal the strength of their dependencies, although the naive Bayesian has higher restrictions that require variables to be independent of each other.

When it comes to K-dependence Bayesian classifiers, it has been proposed to mine dependency relationships from the data. For KDB, all features are indiscriminately conditionally dependent on at most k parent features, even if the conditional dependencies are very weak. KDB provides the "average network" to express significant dependencies, so it cannot apply to all cases. At the same time, KDB cannot accurately describe the dependency relationships in different patient records [37]. However, KDB has satisfactory classification accuracy when dealing with large samples. In addition, KDB uses a single parameter k to determine the number of parents for each feature, thus controlling the complexity of the structure [38]. In the study, the full network structure with 16 attributes is too complex (32 arcs or conditional dependencies) to explain, so we only select a substructure to clarify. The disadvantage of KDB in terms of extensibility is obvious. As shown in Fig. 8, the value afv (amniotic fluid volume) is a precondition of the value afc (amniotic fluid cleanliness). If they appear as co-parents of some other attribute, e.g. crp (C-reactive protein), the conditional probability P(crp|afv, afc, np) will approximate the estimate of P(crp|afc, y) and afv (amniotic fluid volume) cannot provide valuable information on amniotic fluid cleanliness.

The limitation of this study is that only 16 variables were included into the model, and many variables that were not statistically significant were excluded. In fact, many influencing factors must be considered in the application. At the same time, these variables may have collinearity problems, and subsequent research can also consider combining principal component analysis or factor analysis to reduce the dimensionality of variables. Some continuous variables are discretized, which leads to the waste of data information. In this paper, constraint-based learning algrithms may be useful to compare different network structures for the same data, to verify the goodness of fit of the learned network with respect to a particular score function where our future work will focus on. Since the structures is not evident in the Max–Min Hill-Climbing (MMHC) hybrid algorithm, and some hybrid algorithms such as MMPC-Tabu, Fast.iamb-Tabu and Inter.iamb-Tabu. However, we should look at it in a dialectical way where the

ML models in another set of datasets is to test model accuracies, whether it is consistent or not relies on one dataset [38]. Our next step is to assess their performance and make a comparison between the widely used Bayesian network algorithm on more generated datasets, and explore a new hybrid Bayesian network method.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-023-02070-9.

Additional file 1: Figure S1. Bayesian network with MMHC[1]. **Figure S2.** Bayesian network with Fast.iamb-Tabu[1]. **Figure S3.** Bayesian network with Inter.iamb-Tabu[1]. **Figure S4.** Bayesian network with MMHC.Tabu[1]. **Figure S5.** Bayesian network (hill climbing, directed acyclic graph)[2]. **Figure S6.** Bayesian network (Scutari & Nagarajan's (2013) method)[2-4].

Acknowledgements

In carrying out this research project, we sincerely acknowledge all the support from clinical doctors of Foshan Women's and Children's Hospital of Guangdong Medical Universitywho gave me a lot of explanation of clinial treatment and experience.

Authors' contributions

LIN Yue designed the study, led the writing of the article and made analysis of data. CHEN JiaShen and Zhong Ni collected the data. Zhang ao revised the version the article. Pan Haiyan is charge of supervision. All authors critically revised the article and approved the final version of the article.

Authors' information

Lin Yue, famale, 26 years old by far, a Master majored in Epidemiology and Health Statistics of public health in GuangDong Medical University. She has been working on Chronic epidemiology and applying programmes of chronic diseases like diabetes and Chronic Obstructive Pulmonary Disease.

Funding

General Project of Humanities and Social Science Research (NO: 21YJC910007); Dongguan Social Development Science and Technology Key Project (NO: 20231800936252); Funding for innovation strategies for college students (NO: pdjh2023b0240).

Availability of data and materials

All data generated or analysed during this study are included in this published article.

Declarations

Ethics approval and consent to participate

All procedures performed in studies involving human participants were in accordance with the the Ethics Committee of Foshan Women's and Children's Hospital of Guangdong Medical University. Informed consent was obtained from all subjects and/or their legal guardian(s) and this study was approved by the Ethics Committee of Foshan Women's and Children's Hospital of Guangdong Medical University (SDFYMC001).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 6 April 2023 Accepted: 13 October 2023
Published online: 25 October 2023

References

1. Zhu Y, Zhang C. Prevalence of gestational diabetes and risk of progression to type 2 diabetes: a global perspective. Curr Diab Rep. 2016;16:7.
2. Saravanan P, Diabetes in Pregnancy Working Group, Maternal Medicine Clinical Study Group, Royal College of Obstetricians and Gynaecologists, UK. Gestational diabetes: opportunities for improving maternal and child health. Lancet Diabetes Endocrinol. 2020;8:793–800.
3. Hartling L, Dryden DM, Guthrie A, Muise M, Vandermeer B, Donovan L. Benefits and harms of treating gestational diabetes mellitus: a systematic review and meta-analysis for the U.S. Preventive Services Task Force and the National Institutes of Health Office of Medical Applications of Research. Ann Intern Med. 2013;159:123–9.
4. McIntyre HD, Catalano P, Zhang C, Desoye G, Mathiesen ER, Damm P. Gestational diabetes mellitus. Nat Rev Dis Primers. 2019;5:47.
5. Balsells M, García-Patterson A, Gich I, Corcoy R. Maternal and fetal outcome in women with type 2 versus type 1 diabetes mellitus: a systematic review and metaanalysis. J Clin Endocrinol Metab. 2009;94:4284–91.
6. Murphy HR, Steel SA, Roland JM, et al. East Anglia Study Group for Improving Pregnancy Outcomes in Women with Diabetes (EASIPOD). Obstetric and perinatal outcomes in pregnancies complicated by Type 1 and Type 2 diabetes: influences of glycaemic control, obesity and social disadvantage. Diabet Med. 2011;28:1060–7.
7. Farrar D, Simmonds M, Bryant M, et al. Hyperglycaemia and risk of adverse perinatal outcomes: systematic review and meta-analysis. Obstet Anesthes Dig. 2017;37:64–5.
8. Ye W, Luo C, Huang J, Li C, Liu Z, Liu F. Gestational diabetes mellitus and adverse pregnancy outcomes: systematic review and meta-analysis. BMJ. 2022;377:e067946.
9. Omran A, Ali Y, Abdalla MO, El-Sharkawy S, Rezk AR, Khashana A. Salivary interleukin-6 and C-reactive protein/mean platelet volume ratio in the diagnosis of late-onset neonatal pneumonia. J Immunol Res. 2021;18(2021):8495889.
10. Kline JA, Novobilski AJ, Kabrhel C, Richman PB, Courtney DM. Derivation and validation of a Bayesian network to predict pretest probability of venous thromboembolism. Ann Emerg Med. 2005;45(3):282–90.
11. Zhu M, Chen W, Hirdes JP, Stolee P. The K-nearest neighbor algorithm predicted rehabilitation potential better than current clinical assessment protocol. J Clin Epidemiol. 2007;60(10):1015–21.
12. Suner A, Çelikoğlu CC, Dicle O, Sökmen S. Sequential decision tree using the analytic hierarchy process for decision support in rectal cancer. Artif Intell Med. 2012;56(1):59–68.
13. Karaboga HA, Gunel A, Korkut SV, Demir I, Celik R. Bayesian network as a decision tool for predicting ALS disease. Brain Sci. 2021;11(2):150 Published 2021 Jan 23.
14. Stewart GB, Mengersen K, Meader N. Potential uses of Bayesian networks as tools for synthesis of systematic reviews of complex interventions. Res Synth Methods. 2014;5(1):1–12.
15. Seixas FL, Zadrozny B, Laks J, Conci A, MuchaluatSaade DC. A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer׳s disease and mild cognitive impairment. Comput Biol Med. 2014;51:140–58.
16. Sahami M. Learning limited dependence Bayesian classifiers[C]. In: Proceedings of knowledge discovery and data mining (International Conference). 1996;96(1):335-338.
17. Chattopadhyay S, Sahu SK. "A predictive stressor-integrated model of suicide right from one's birth: a Bayesian approach." J Med Imaging Health Inform. 2012;2(2):125–31.
18. Fuster-Parra P, Yañez AM, López-González A, Aguiló A, Bennasar-Veny M. Identifying risk factors of developing type 2 diabetes from an adult population with initial prediabetes using a Bayesian network. Front Public Health. 2023;10:1035025 Published 2023 Jan 12.
19. Luo Y, Carretta H, Lee I, LeBlanc G, Sinha D, Rust G. Naïve Bayesian network-based contribution analysis of tumor biology and healthcare factors to racial disparity in breast cancer stage-at-diagnosis. Health Inf Sci Syst. 2021;9(1):35.
20. Peng Y, Cheng L, Jiang Y, Zhu S. Examining Bayesian network modeling in identification of dangerous driving behavior. PLoS ONE. 2021;16(8):e0252484.
21. Jing C, Gang T, Yong L, et al. Bayesian network based Netica for respiratory diseases. 2018.

22. Zhang H, Huang X, Han S, et al. Gaussian Bayesian network comparisons with graph ordering unknown. Comput Stat Data Anal. 2021;157:107156.
23. McNally RJ, Mair P, Mugno BL, Riemann BC. Co-morbid obsessive-compulsive disorder and depression: a Bayesian network approach. Psychol Med. 2017;47(7):1204–14.
24. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco: Morgan Kaufmann; 1988.
25. Long Y, Wang L, Sun M. Structure extension of tree-augmented naive bayes. Entropy (Basel). 2019;21(8):721.
26. Liu Y, Wang L, Sun M. Efficient heuristics for structure learning of k-dependence Bayesian classifier. Entropy (Basel). 2018;20(12):897 Published 2018 Nov 22.
27. Wang L, Liu Y, Mammadov M, Sun M, Qi S. Discriminative structure learning of bayesian network classifiers from training dataset and testing instance. Entropy (Basel). 2019;21(5):489.
28. Chattopadhyay S, Davis RM, Menezes DD, Singh G, Acharya UR, Tamura T. "Application of Bayesian classifier for the diagnosis of dental pain." J Med Syst. 2012;36:1425–39. https://doi.org/10.1007/s10916-010-9604-y.
29. Zhao E, Zhang Y, Zeng X, Liu B. Association between maternal diabetes mellitus and the risk of congenital malformations: a meta-analysis of cohort studies. Drug Discov Ther. 2015;9:274–81.
30. Chen L, Yang T, Chen L, et al. Risk of congenital heart defects in offspring exposed to maternal diabetes mellitus: an updated systematic review and meta-analysis. Arch Gynecol Obstet. 2019;300:1491–506.
31. He XJ, Qin FY, Hu CL, Zhu M, Tian CQ, Li L. Is gestational diabetes mellitus an independent risk factor for macrosomia: a meta-analysis? Arch Gynecol Obstet. 2015;291:729–35.
32. Tabrizi R, Asemi Z, Lankarani KB, et al. Gestational diabetes mellitus in association with macrosomia in Iran: a meta-analysis. J Diabetes Metab Disord. 2019;18:41–50.
33. Farrar D, Simmonds M, Bryant M, et al. Hyperglycaemia and risk of adverse perinatal outcomes: systematic review and meta-analysis. BMJ. 2016;354:i4694.
34. Li Y, Wang W, Zhang D. Maternal diabetes mellitus and risk of neonatal respiratory distress syndrome: a meta-analysis. Acta Diabetol. 2019;56:729–40.
35. Phan LT, Oh C, He T, Manavalan B. A comprehensive revisit of the machine-learning tools developed for the identification of enhancers in the human genome. Proteomics. 2023;23(13–14):e2200409.
36. Ma Y, Shen J, Zhao Z, et al. What can facial movements reveal? Depression recognition and analysis based on optical flow using Bayesian networks [published online ahead of print, 2023 Aug 15]. IEEE Trans Neural Syst Rehabil Eng. 2023;PP. https://doi.org/10.1109/TNSRE.2023.3305351.
37. Lou H, Wang L, Duan D, Yang C, Mammadov M. RDE: A novel approach to improve the classification performance and expressivity of KDB. PLoS ONE. 2018;13(7):e0199822.
38. Chattopadhyay S, Rajput SS, Prajesh AR. "Testing Bayesian classifiers on adult depression data: a study to handle uncertainty related to its grading." J Med Imaging Health Inform. 2013;3(4):607–16.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.