DIGITAL HEALTH

Digital Health
Volume 8: 1–17
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076221131185
journals.sagepub.com/home/dhj
SAGE

Interpretable modeling and discovery of key predictors for pneumonia diagnosis in children based on electronic medical records

Jing Li[1,4,*], Yingshuo Wang[2,*], Qiuyang Sheng[3,*], Xiaoqing Liu[3], Zijian Xing[3], Fenglei Sun[3], Yuqi Wang[2], Shuxian Li[2], Yiming Li[3], Yizhou Yu[3] and Gang Yu[1,4,5]

Abstract

Background: Community-acquired pneumonia is one of the most common infectious diseases in children and is a leading cause of death among children under 5 years of age, resulting in high rates of antibiotic usage and hospitalization. It is of extremely practical significance to make full use of the existing electronic medical records to study pneumonia and to establish automatic diagnosis models for pneumonia.

Methods: We established pneumonia diagnosis models of Bayesian network using a total of 13,448 electronic medical records. We investigated learning network structure and parameter estimation and evaluated different structure learning strategies and various modeling methods. By identifying the key predictors of model, the pneumonia status was analyzed.

Results: The performance of the proposed Bayesian network was evaluated using a set of 3361 cases with a precision of 0.7861, a recall of 0.9889, and an F1-score of 0.8759. On an independent external validation set containing 4925 cases, Bayesian network achieved a precision of 0.7382, a recall of 0.9947, and an F1-score of 0.8475. Our proposed Bayesian network outperformed all other methods, including CatBoost, XGBoost, LightGBM, logistic regression, and ridge classification.

Conclusion: The appropriate feature selection improved the performance of Bayesian networks. The proposed Bayesian network had good generalizability and could be directly applied to clinical research centers. And the key predictors identified by the network demonstrated good clinical interpretability, allowing for a better understanding of pneumonia status and complications. This study had important clinical value and practical significance for the research and diagnosis of pediatric pneumonia.

Keywords

pneumonia diagnosis, interpretable modeling, knowledge discovery, Bayesian networks, electronic medical records

Submission date: 1 July 2022; Acceptance date: 20 September 2022

[1]Department of Data and Information, The Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China
[2]Department of Pulmonology, The Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China
[3]Deepwise Healthcare Artificial Intelligence Laboratory, Beijing, China
[4]Sino-Finland Joint AI Laboratory for Child Health of Zhejiang Province, Hangzhou, China
[5]Polytechnic Institute, Zhejiang University, Hangzhou, China

[*]These authors have contributed equally to this work and share first authorship.

Corresponding authors:
Gang Yu, Department of Data and Information, The Children's Hospital, Zhejiang University School of Medicine, 3333 Binsheng Road, Binjiang District, Hangzhou 310052, China.
Email: yugbme@zju.edu.cn

Yizhou Yu, Deepwise Healthcare Artificial Intelligence Laboratory, 13th Floor, Building 2, Yard 2, Xisanhuan North Road, Haidian District, Beijing, China.
Email: yizhouy@acm.org

Introduction

Community-acquired pneumonia (CAP), defined as pneumonia acquired outside of a hospital or health care setting, is a common infectious disease and a leading cause of death among children under 5 years of age, resulting in high rates of antibiotic use and hospitalization.[1,2] The annual incidence of CAP requiring hospitalization was 15.7 per 10,000 children, with the highest incidence in children under 2 years of age in the United States.[3] Pneumonia also imposes a heavy economic burden on both developed and developing countries. In the United Kingdom, potential direct medical costs for children aged 0 to 16 years hospitalized with CAP range from £12 to £18000 per year.[4] Current management strategies remain suboptimal, in part due to insufficient technology to determine etiology, triage patients, and predict their outcomes.[5]

With recent advances in machine learning technology, machine learning models have been increasingly applied to the analysis of large-scale electronic medical record (EMR) data, helping to learn effective patterns, discover knowledge, and build disease diagnosis models from the data. Prosperi et al.[6] used logistic regression, random forests, and AdaBoost to identify asthma, wheezing, and eczema. Sun et al.[7] developed pneumonia prediction models using classification and regression trees, and analyzed that older age, comorbidities, and initial presentation of lower respiratory tract infections were the main predictors of pneumonia. Giang et al.[8] attempted to build a model to predict ventilator-associated pneumonia from EMR data. Yu et al.[9] evaluated a range of machine learning methods on a dataset of 16 features from EMR, with the CatBoost model achieving the best performance. Existing methods have achieved excellent prediction ability and can well express the relationship between input and output variables, but they fail to take into consideration the underlying relationship between input variables.[10] In certain clinical scenarios, the ability to capture inherent intrinsic relationships between input variables has a far greater clinical value for disease analysis.[11]

Bayesian network modeling has attracted considerable attention in medical diagnosis due to its ability to establish probabilistic relationships between diseases and their associated symptoms.[12,13] Zhao et al.[14] proposed a hybrid neuro-probabilistic reasoning algorithm that integrated Bayesian networks with graph convolutional networks to discriminate benign and malignant pulmonary nodules in computed tomography images. Spyroglou et al.[15] evaluated the performance of a Bayesian network classifier in predicting asthma exacerbations based on multiple patient parameters, including objective measurements and medical history data. Sanders and Aronsky[16] also developed and evaluated a Bayesian network to identify patients who met asthma care guidelines using only electronically provided data at patient triage.

In this study, we established a Bayesian network pneumonia diagnosis model based on EMR data. The application of feature selection based on odds ratio (OR) values proved that the classification performance of Bayesian networks is as good as that of popular machine learning algorithms. In addition, the key predictors identified by the network demonstrated good interpretability, allowing a better understanding of pneumonia status and complications. Through independent external validation, we demonstrated that our proposed Bayesian network has good generalizability and can be directly applied to EMRs in clinical research centers. In summary, this study had incredibly important clinical value and practical significance for the research in the field of pediatric pneumonia and the rapid automated diagnosis of pneumonia.

Materials and methods

Data collection and preparation

In this study, we retrospectively collected EMRs of 33,571 consecutive patients with a mean age of 3.81 (standard deviation (SD) = 2.41) admitted to the Department of Pulmonology, Children's Hospital of Zhejiang University School of Medicine, China from 2012 to 2020, as an internal dataset for model training and validation. EMRs consisted of the *first course records*, *admission records*, *discharge records*, etc. Among them, the *discharge diagnosis* recorded the patient's final confirmed diseases, and the text in the *first course records and admission records*, such as the *history of present illness*, *physical examination*, and *auxiliary examination*, recorded the details of the clinician's inquiry and observation of the patients' statuses. We constructed an experimental dataset using texts from these domains. Specifically, the information of *history of present illness*, *physical examination*, and *auxiliary examination* were taken as feature corpus X, and the information of *discharge* diagnosis was taken as target y. Notably, statistics showed that the vast majority of patients had more than one diagnosed disease due to complications. Table 1 provides a sample English translation to illustrate the EMR types and corresponding fields used in our study.

Independent external validation data

In addition, we also collected EMR data of 6573 patients with a mean age of 2.27 years (SD = 2.16) from the Department of Pulmonary Medicine, Zhengzhou Children's Hospital, China, as an independent external validation dataset to evaluate the clinical generalization performance of the proposed Bayesian network.

Tabular dataset building

The texts of *history of present illness*, *physical examination*, and *auxiliary examination* of each patient were concatenated as characteristic corpus records. First, HanLP,[17] an open-source multilingual language processing toolkit, was used to automatically split each corpus record into a series of keywords through an entropy-based key phrase extraction method. Considering Chinese grammatical expressions, key phrases were defined in 2-gram and 3-gram forms. After filtering phrases based on the stopword list, 53,719 key phrases remained. Among them, the top 500 key phrases with the highest frequency accounted for 68.43% of the total frequency (approximately within the range of $\pm\sigma$ of the probability mass function). Based on the sigma principle, the top 500 key phrases with the highest frequency were selected. Then, a knowledge graph-driven search engine was utilized to distinguish whether key phrases were relevant to the medical domain and whether they had more canonical counterparts. For example, "tonsil infection" would be considered a term in the medical field and would be aligned with the more formal scientific name (tonsillitis). Finally, those aligned terms were confirmed through human expert review. As a result, 47 designated terms were reserved. In this study, a lookup table was used to map nonstandard terms occurring in the datasets to 47 designated terms.

Inspired by NegEx,[18] 58 regular expressions were adopted to find the positive and negative scopes of 47 designated terms in each corpus record, respectively. An example of a common negative mention pattern follows the structure of "*(no | not | there is no) key phrase (not seen | not found).*" Using regular expressions, we were able to determine the exact value of each term in each record, including positive mentions (1), negative mentions (0), and no mentions (0). According to whether pneumonia was recorded at the time of *discharge diagnosis*, cases of pneumonia were defined as positive samples (1) and cases without pneumonia were defined as negative samples (0). All of which were binary terms that should be coded as either positive (1) or negative (0), that is, bi-values are not allowed.

Finally, a tabular dataset named DataSet-PT was generated from the internal dataset containing 47 features and 1 target. By eliminating duplicate data, a total of 16,809 case records were retained, including 11,640 cases of pneumonia, accounting for 69.25% of the total, and 5169 cases of non-pneumonia, accounting for 30.75% of the total. Figure 1 illustrates the detailed statistics of DataSet-PT, in which Figure 1(a) shows the distribution of diseases included in the *discharge diagnosis* field, Figure 1(b) presents

Table 1. An example of the texts in four fields from raw EMRs in English translation.

Records	Fields	Descriptions
Admission records	History of present illness	The child had ***cough*** without obvious inducement 1 month ago, occasional ***cough***, ***sputum***, no ***barking cough*** and ***whoop***, with ***wheezing***, obvious in the morning awaking and at night. The child had no obvious ***dyspnea***, no ***fever***, no ***chill***, no ***convulsions***, and no ***diarrhea*** hence he was brought to the outpatient clinic of the local hospital. After the local hospital gave ***cefotaxime sodium intravenous*** drops to fight infection, ***methylprednisolone intravenous*** drops to fight inflammation and ***aerosol inhalation*** to relieve ***wheezing***, the symptoms were slightly improved, but there was still ***cough*** and ***wheezing***. The child has been given ongoing ***aerosol inhalation*** treatment in past 1 month. The child came to our hospital today for further treatment. At present, the child was in normal state, had a little poor ***appetite*** and normal ***sleep***. There was no abnormal in ***stool*** and ***urine***, and no significant weight gain or loss
Admission records	Physical examination	T 37°C, P 130 beats per min, R 36 breaths per min, BP 107/67 mmHg, SpO_2 96%. The *spirit* is ***conscious***, *throat* has no ***inflammation*** and breathing is smooth. There was no ***three concave sign***, no ***nodding respiration***, but ***coarse breath sound*** and ***wheezing rale*** could be heard in both lungs. The child had no ***arrhythmia***, no obvious ***pathological murmurs***, and had no ***liver***, ***spleen***, and ***subcostal enlargement*** with ***soft abdomen***. The neurological examination was negative. There was no ***rash*** and the temperature of the extremities was normal
First course records	Auxiliary examination	The chest X-ray: the color of texture in both lungs was ***thickened***; Cardiac color Doppler ultrasound: no abnormalities were observed
Discharge records	Discharge diagnosis	(1) Pneumonia (2) Mycoplasma infection (3) *Streptococcus* infection (4) Stomach twist (5) Laryngopharyngeal reflux (suspected)

Note: The *italic* **bold** phrases are the key phrases with high information entropy. BP: blood pressure; EMR: electronic medical record; P: pulse rate; R: respiration rate; SpO_2: oxygen saturation; T: temperature.

the proportion of pneumonia cases and non-pneumonia cases, and Figure 1(c) illustrates the feature distribution. Using random shuffling, DataSet-PT were further split into a training set TrainSet-PT for modeling and an internal validation set TestSet-PT for performance evaluation at a ratio of 4:1, that is, 13,448 cases in TrainSet-PT and 3361 cases in TestSet-PT.

For the independent external validation dataset, a tabular dataset named ExternalSet-PT was obtained following the same pipeline, retaining 4925 case records, including 3564 pneumonia cases (72.37% of the total) and 1361 non-pneumonia cases (27.63% of the total). The data distribution of ExternalSet-PT is shown in Figure 2, where Figure 2(a) and (b) present the disease and feature distributions, respectively.

Bayesian network modeling

A Bayesian network denoted by $N(G, P)$ consists of a direct acyclic graph (DAG), denoted by G, and a set of conditional probability distributions P. Each node of G represents a unique discrete random variable X with mutually exclusive states $x_1, \cdots, x_k$. Each node also has a conditional probability table (CPT) that quantifies the influence of the parent node (all nodes with arrows pointing to it) on it. Building a Bayesian network typically requires the following steps. First, the structure of Bayesian network is constructed through structure learning (i.e. the DAG is formed). Second, parameter estimation is used to calculate the CPT of each node (i.e. establish the strength relationship between node dependencies). Third, using methods such as variable elimination, the Bayesian network is Inferred to output prediction results. In this study, we used probabilistic graphical models using python (pgmpy)[19] to build and evaluate Bayesian networks.

Structure construction

Feature correlation. Inspired by previous works,[20,21] we used the OR values to measure the correlation between features, where an OR value is a measure of the association

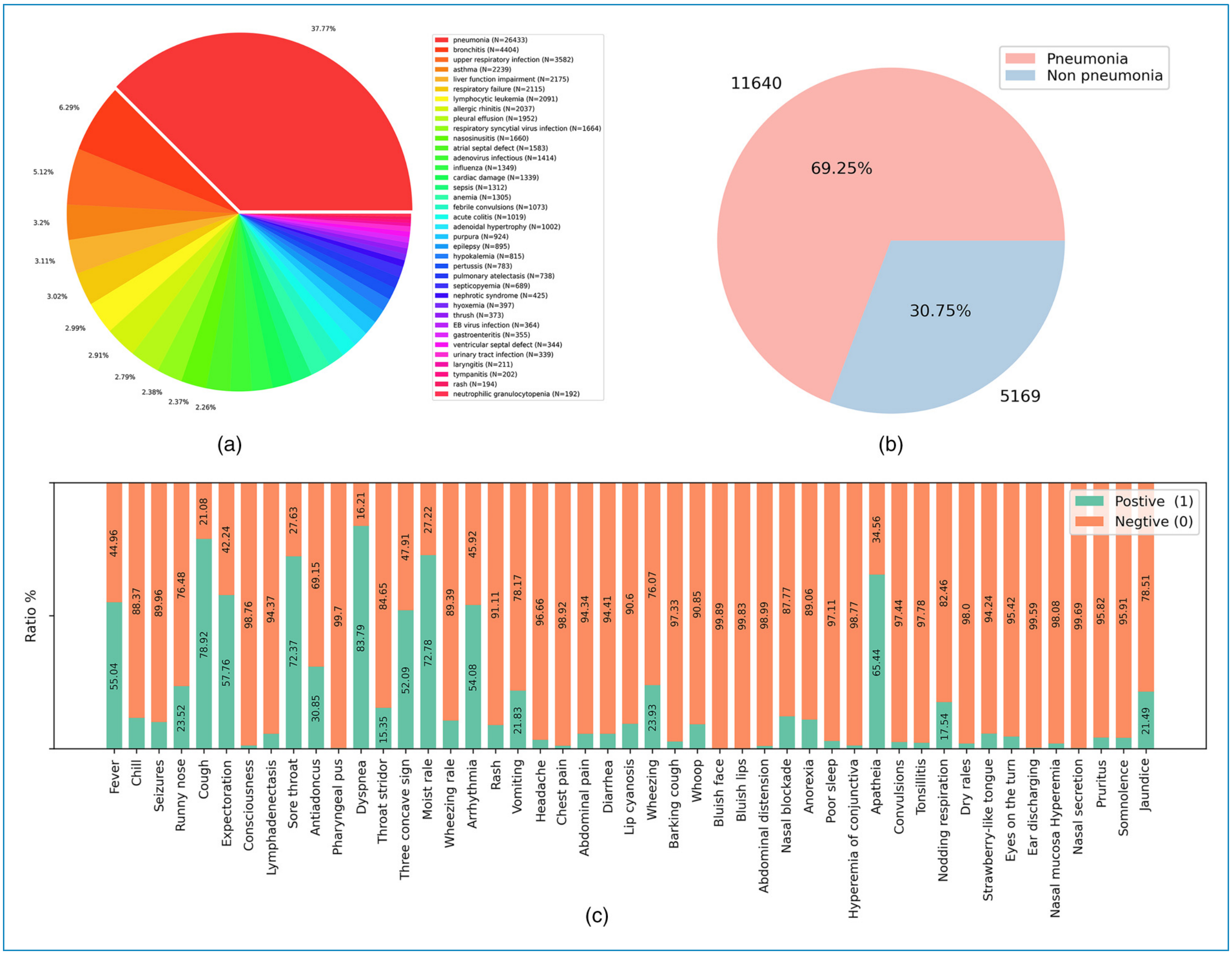


Figure 1. Data distribution. (a) Pie chart of disease distribution contained in the *discharge diagnosis* text in the electronic medical records (EMRs). There were 35 types of diseases with pneumonia being the most diagnosed disease. (b) Proportion pie chart of pneumonia cases and non-pneumonia cases in the DataSet-PT after removing duplicate records. (c) Feature distributions in the DataSet-PT. The total amount for each feature was 16,809.

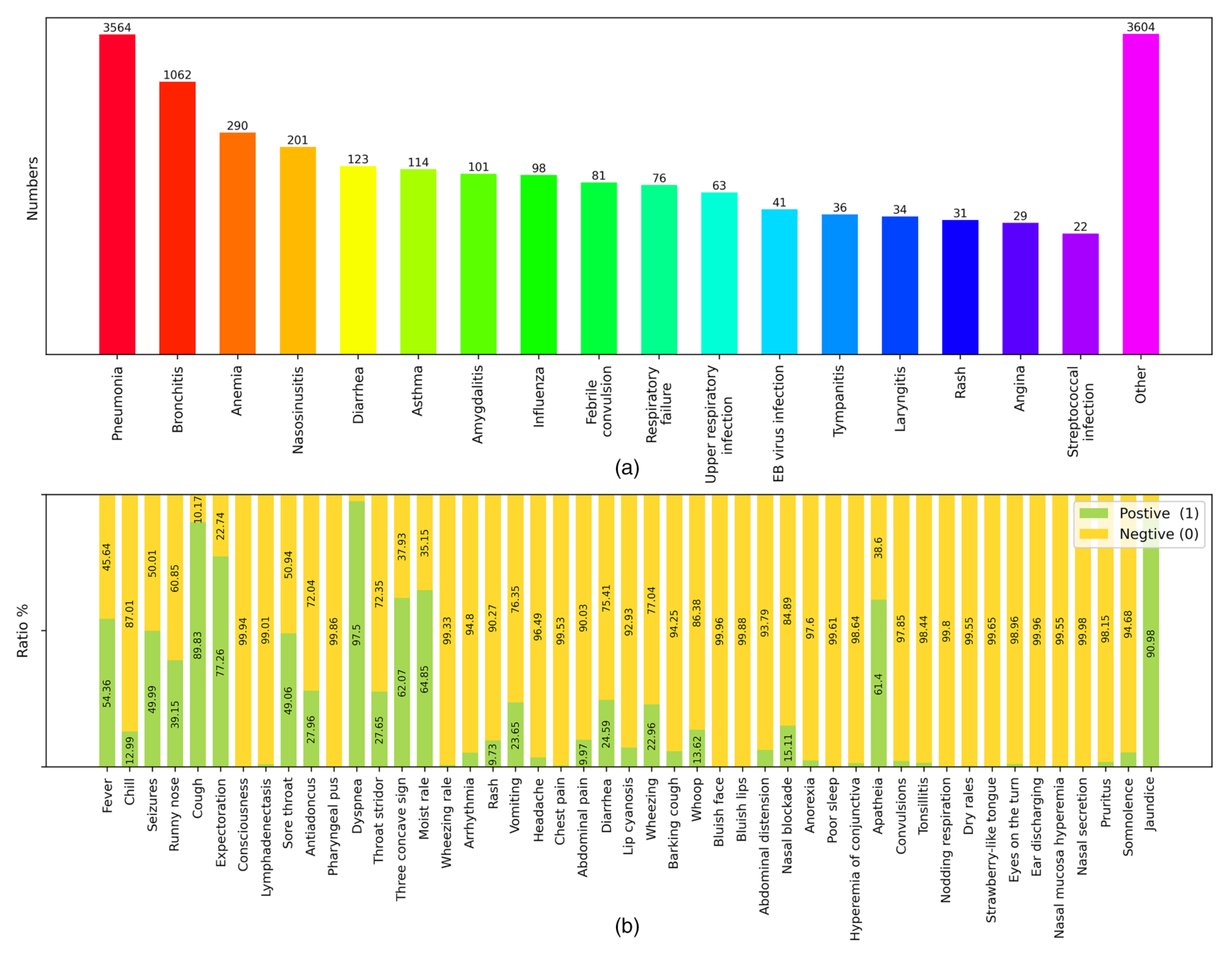


Figure 2. Distribution of independent external validation data. (a) Bar chart of the disease contained in external electronic medical records. There were 17 diseases and others. Some cases had more than one disease. (b) Feature distribution in ExternalSet-PT. The total for each feature was 4925.

between exposure and outcome. The OR value between feature N_i and N_j is calculated as follows:

$$\mathrm{OR}(N_i, N_j) = \frac{P(N_i = 1|N_j = 1) \times P(N_i = 0|N_j = 0)}{P(N_i = 0|N_j = 1) \times P(N_i = 1|N_j = 0)} \quad (1)$$

OR > 1, feature N_i is strongly correlated with N_j and OR ≤ 1, feature N_i is weakly correlated with N_j.

Features selection. Given that the purpose of our Bayesian network modeling was to predict the occurrence of pneumonia, it was necessary to determine whether a feature node should be connected to the target pneumonia node. It was also meaningful to distinguish nodes that had *direct* or *indirect connections* to the pneumonia node. This process was considered as a feature selection process, resulting in two sets of features, among which the feature set containing only *directly connected* nodes was named as Fine-Features and the feature set containing all features including both *directly* and *indirectly connected* nodes were named as All-Features.

Network structure search. The DAG was constructed using a heuristic, asymptotic, greedy, hill climbing algorithm. In the step of adding each node, three operations (namely, *adding edges*, *subtracting edges*, and *reversing edges*) were performed to reduce the score of the entire network structure, where K2 score[22] was used to measure the structure score.

Initial graph. In greedy search, the initial graph is an important factor.[23] We used two initial graph configurations, including a configuration without an initial graph and a configuration with an initial graph generated based on OR values.

Ranking strategy. The order in which nodes are added is also an important factor in greedy search. We applied four ranking strategies[21]: (1) *Random*, (2) *Global impact*, (3) *Descending*, and (4) *Ascending*.

Figure 3 shows the resulting Bayesian network architectures, where Figure 3(a) demonstrates an undirected graph with 478 edges, while Figure 3(b) and (c) shows the Bayesian network structure based on All-Features and Fine-Features, respectively. The feature correlation heatmap based on the OR value was shown in Figure 4a, where each point represented the logarithm of the OR value between the features on the *x*-axis and *y*-axis. Clearly, points with a value >0 (i.e. colors close to the red band) represented a strong correlation between the two features, while points with a value < 0 (i.e. colors close to the blue band) indicated a weaker correlation between the two features.

Parameter estimation and model inference

In the parameter estimation stage, we used a Bayesian estimator to estimate the CPT of each node. Bayesian–Dirichlet equivalence consistent prior was applied to compute an initial CPT for each node. Starting with the initial CPT, we updated each CPT using state counts from observations of TrainSet-PT. In the inference stage, the input to the Bayesian network is usually in the form of a series of observed evidence. Specifically, predicting the likelihood of pneumonia using All-Features is equivalent to computing the posterior probability of the Bayesian network $P(X_{pneu}|E = e)$, where

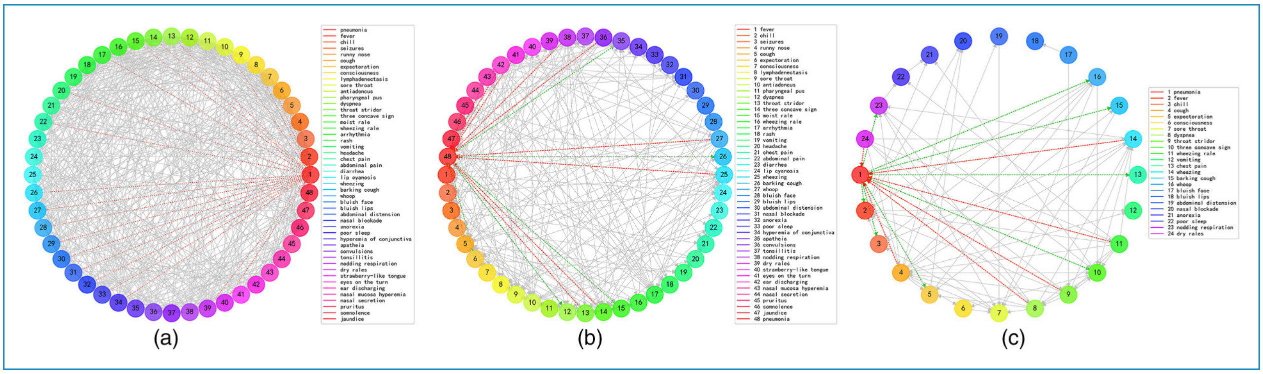
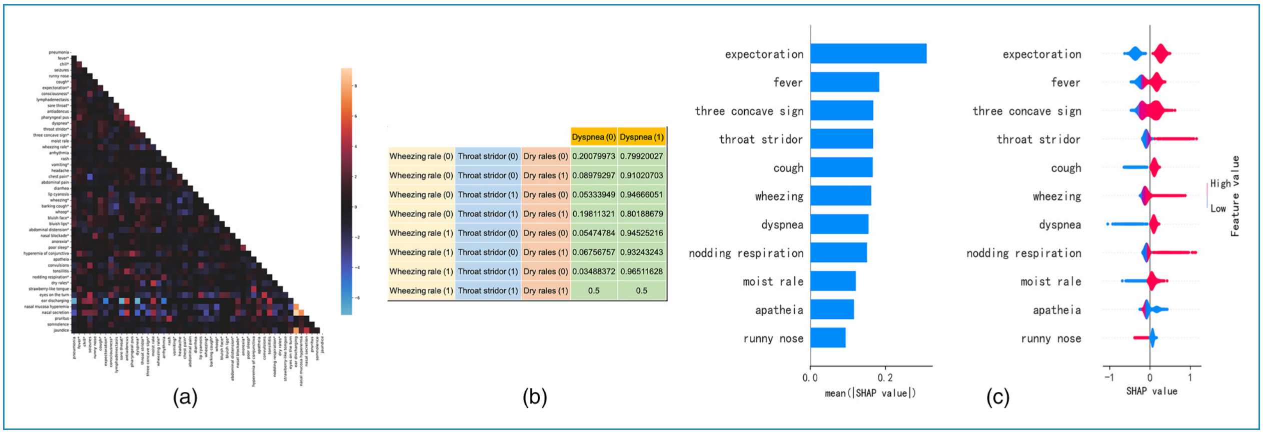


Figure 3. Bayesian network architecture. (a) Undirected graph generated based on odds ratio values. Colored circles represent specific nodes in the network. The lines between the circles represent the connections of the nodes. The red dashed lines are the connections associated with the pneumonia node, and the gray lines are the connections not associated with the pneumonia node. (b) Bayesian network generated from the All-Features set. The red dashed line represents the parent node of pneumonia, while the green dashed line represents the child node of pneumonia. Gray lines indicate relationships that are not directly related to the pneumonia node. Arrows on the lines indicate causal relationships between nodes. (b) Bayesian network generated from the Fine-Features set.



			Dyspnea (0)	Dyspnea (1)
Wheezing rale (0)	Throat stridor (0)	Dry rales (0)	0.20079973	0.79920027
Wheezing rale (0)	Throat stridor (0)	Dry rales (1)	0.08979297	0.91020703
Wheezing rale (0)	Throat stridor (1)	Dry rales (0)	0.05333949	0.94666051
Wheezing rale (0)	Throat stridor (1)	Dry rales (1)	0.19811321	0.80188679
Wheezing rale (1)	Throat stridor (0)	Dry rales (0)	0.05474784	0.94525216
Wheezing rale (1)	Throat stridor (0)	Dry rales (1)	0.06756757	0.93243243
Wheezing rale (1)	Throat stridor (1)	Dry rales (0)	0.03488372	0.96511628
Wheezing rale (1)	Throat stridor (1)	Dry rales (1)	0.5	0.5

Figure 4. Feature analysis. (a) Heatmap of correlations between features. Each point represents the logarithm of the OR value between the feature on the *x*-axis and *y*-axis. Clearly, points with a value >0 (i.e. colors close to the red band) represent a strong correlation between the two features. Points with a value <0 (i.e. colors close to the blue band) indicate a weaker correlation between the two features. Features marked with * are closely correlated with pneumonia. (b) An example of a CPT of the dyspnea node from Bayesian network using Fine-Features. For inference, the probability of dyspnea $P(Y_{dyspnea}|\ X_{wheezing\ rale} = 1,\ X_{throat\ stridor} = 1\ ,\ X_{dry\ rales} = 1)$ when the evidence is $\{X_{wheezing\ rale} = 1,\ X_{throat\ stridor} = 1,\ \ X_{dry\ rales} = 1\}$ is equivalent to the sum of the joint probability, i.e. $\sum^{P}(X_{wheezing\ rale} = 1,\ X_{throat\ stridor} = 1\ ,\ \ X_{dry\ rales} = 1)$. (c) Feature importance ranking of CatBoost in All-Features. Features were ranked according to the average absolute value of SHAP values and the SHAP values, respectively. The two rankings were consistent. SHAP: SHapley Additive exPlanations; CPT: conditional probability table; OR: odds ratio.

X_{pneu} is the pneumonia variable, $E = \{E_1, E_2, \cdots, E_{47}\}$ the 47 non-pneumonia variables, and $e = \{e_1, e_2, \cdots, e_{47}\}$ the observed values of 47 variables.

Figure 4(b) illustrates an example of CPT of the dyspnea node in the Bayesian network using Fine-Features. The Bayesian network can also be viewed as a joint probability distribution $P(X_{pneu}, E)$ consisting of a pneumonia variable and 47 non-pneumonia variables. The sum-product variable elimination algorithm can be used for inference, which eliminates the influence of non-pneumonia variables by continually summing the probabilities of the variables in the joint distribution until all non-pneumonia variables are eliminated, leaving only the marginal probability of the pneumonia variable, that is, $P(X_{pneu}) = \sum_E P(X_{pneu}, E_1, E_2, \cdots, E_{47})$. Hence, for actual inference, E is set to the exact value set $\{e_1, e_2, \cdots, e_{47}\}$, and $P(X_{pneu}|E = e) = \sum_e P(X_{pneu}, e_1, e_2, \cdots, e_{47})$.

Other modeling approaches

We applied five other machine learning models, namely CatBoost, XGBoost, LightGBM, logistic regression, and the ridge classifier[24–27] for performance comparison. A grid search strategy was used to determine model hyperparameters. For CatBoost, the number of iteration was 2000, the learning rate was 0.01, the max depth was 7, and the objective was binary log loss. For XGBoost, the number of estimators was 2000, the learning rate was 0.005, the max depth was 7 and the objective was binary log loss. For LightGBM, the number of estimators was 2000, the learning rate was 0.01, the max depth was 10, the max number of leaves was 50, and the objective was binary log loss. For logistic regression, the penalty was the L2 distance, the max iteration was 1000, and solver was L-BFGS. For the ridge classifier, the alpha was 0.5, the tolerance was $1{\times}10^{-3}$, and the solver adopted L-BFGS.

Results

Performance evaluation metrics

In this study, three metrics, including precision, recall, and F1-score, were used to evaluate the performance, and defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

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

$$\text{F1 -score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP, FP, TN, and FN are the true positive, false positive, true negative, and false negative rates, respectively. TP and TN denote correctly predicted positives and negatives with respect to the ground truth labels. FP and FN represent incorrectly predicted positives and negatives with respect to the ground truth labels. Statistical analysis of model performance was based on bootstrap and *t*-test.

Statistical analysis

Statistical *t*-test analysis between the pneumonia-positive and pneumonia-negative groups was analyzed. As shown in Table 2, most features showed statistically significant differences *($p < 0.05$) between the pneumonia-positive and pneumonia-negative groups. Specifically, 40 features showed statistically significant differences between positive and negative groups in TrainSet-PT, and 30 features showed statistically significant differences in TestSet-PT. Of these features, 30 features showed statistically significant differences in both the sets, which confirmed that the extracted features were indeed discriminative. There was no statistically significant difference between TrainSet-PT and TestSet-PT in all features (all $p > 0.05$), which was in line with common sense, indicating that the splitting of DataSet-PT was reasonable. In ExternalSet-PT, there were 25 features showed statistically significant differences between positive and negative groups. Most importantly, only seven features showed no statistically significant differences with TrainSet-PT, which indicated a large feature distribution difference between these two datasets.

Experimental results

Table 3 showed the performance of Bayesian network modeling. It is clear that models without initial graphs generally performed better than models with initial graphs. The *Ascending* strategy worked best on both feature sets, and the *Global impact* strategy achieved a better recall on All-Features. The best performances for precision, recall, and F1-score achieved with Bayesian networks were 0.7746, 0.9722, and 0.8344 with All-Features, and 0.7861, 0.9889, and 0.8759 with Fine-Features.

Table 4 demonstrated the performance comparison of different modeling approaches on TestSet-PT. Using All-Features, the CatBoost achieved the best precision and F1-score with 0.7852 and 0.8471, and the XGBoost and ridge classifier achieved the best recall of 0.9243. The Bayesian network, which without initial graph and using *Ascending* ranking strategy, outperformed other models, with the highest precision (0.7861), highest recall (0.9889), and the highest F1-score (0.8759) using Fine-Features. The metrics of other modeling approaches were shown significant differences with the performance of Bayesian network when Fine-Features was used.

Table 5 demonstrated the performance comparison of different modeling methods on ExternalSet-PT. The ridge classifier achieved the best recall and F1-score with 0.9837 and 0.8382, respectively, while the logistic regression achieved the best precision of 0.7459 with All-Features. The Bayesian network outperformed other models with the highest precision

Table 2. Statistical characteristics of TrainSet-PT, TestSet-PT, and ExternalSet-PT.

	TrainSet-PT			TestSet-PT				ExternalSet-PT			
Characteristic (*N* = 47)	Positive group (*N* = 9302)	Negative group (*N* = 4146)	*p*	Positive group (*N* = 2338)	Negative group (*N* = 1023)	*p*	*p*	Positive group (*N* = 3564)	Negative group (*N* = 1361)	*p*	*p*
Fever	0.567 ± 0.4956	0.522 ± 0.4995	<0.0001*	0.551 ± 0.4974	0.514 ± 0.4998	0.0473*	0.1785	0.543 ± 0.4982	0.546 ± 0.4979	0.8366	0.4654
Chill	0.131 ± 0.3370	0.089 ± 0.2844	<0.0001*	0.127 ± 0.3335	0.072 ± 0.2590	<0.0001*	0.2479	0.128 ± 0.3337	0.136 ± 0.3427	0.4461	0.0155*
Seizures	0.078 ± 0.2689	0.147 ± 0.3545	<0.0001*	0.088 ± 0.2828	0.138 ± 0.3447	<0.0001*	0.5808	0.513 ± 0.4998	0.465 ± 0.4988	0.0025*	<0.0001*
Runny nose	0.226 ± 0.4182	0.253 ± 0.4346	0.0008*	0.237 ± 0.4252	0.244 ± 0.4297	0.6440	0.5445	0.386 ± 0.4867	0.407 ± 0.4913	0.1682	<0.0001*
Cough	0.848 ± 0.3589	0.656 ± 0.4751	<0.0001*	0.852 ± 0.3547	0.650 ± 0.4770	<0.0001*	0.7969	0.908 ± 0.2895	0.874 ± 0.3323	0.0009*	<0.0001*
Expectoration	0.648 ± 0.4776	0.415 ± 0.4927	<0.0001*	0.661 ± 0.4734	0.405 ± 0.4908	<0.0001*	0.4900	0.792 ± 0.4062	0.723 ± 0.4475	<0.0001*	<0.0001*
Consciousness	0.012 ± 0.1110	0.012 ± 0.1092	0.8412	0.014 ± 0.1162	0.010 ± 0.0984	0.3166	0.9432	0.001 ± 0.0290	0.001 ± 0.0001	0.0833	<0.0001*
Lymphadenectasis	0.048 ± 0.2134	0.075 ± 0.2630	<0.0001*	0.052 ± 0.2224	0.068 ± 0.2525	0.0757	0.8258	0.008 ± 0.0883	0.015 ± 0.1233	0.0384*	<0.0001*
Sore throat	0.737 ± 0.4404	0.695 ± 0.4606	<0.0001*	0.726 ± 0.4459	0.717 ± 0.4507	0.5632	0.9512	0.446 ± 0.4971	0.606 ± 0.4886	<0.0001*	<0.0001*
Antiadoncus	0.292 ± 0.4546	0.348 ± 0.4762	<0.0001*	0.285 ± 0.4516	0.355 ± 0.4785	0.0001*	0.7777	0.233 ± 0.4230	0.400 ± 0.4900	<0.0001*	<0.0001*
Pharyngeal pus	0.003 ± 0.0518	0.003 ± 0.0559	0.6608	0.002 ± 0.0462	0.008 ± 0.0881	0.0516	0.3709	0.001 ± 0.0335	0.002 ± 0.0469	0.4364	0.005*
Dyspnea	0.891 ± 0.3122	0.724 ± 0.4471	<0.0001*	0.886 ± 0.3181	0.712 ± 0.4530	<0.0001*	0.3747	0.975 ± 0.1552	0.974 ± 0.1583	0.8382	<0.0001*
Throat stridor	0.189 ± 0.3913	0.076 ± 0.2657	<0.0001*	0.182 ± 0.3860	0.080 ± 0.2715	<0.0001*	0.6642	0.324 ± 0.4680	0.152 ± 0.3591	<0.0001*	<0.0001*
Three concave sign	0.583 ± 0.4930	0.381 ± 0.4856	<0.0001*	0.585 ± 0.4927	0.373 ± 0.4837	<0.0001*	0.9820	0.687 ± 0.4635	0.446 ± 0.4971	<0.0001*	<0.0001*
Moist rale	0.718 ± 0.4502	0.757 ± 0.4287	<0.0001*	0.710 ± 0.4538	0.741 ± 0.4381	0.0626	0.2282	0.670 ± 0.4701	0.591 ± 0.4916	<0.0001*	<0.0001*
Wheezing rale	0.121 ± 0.3261	0.070 ± 0.2555	<0.0001*	0.122 ± 0.3277	0.079 ± 0.2700	0.0001*	0.5155	0.006 ± 0.0801	0.007 ± 0.0854	0.7383	<0.0001*
Arrhythmia	0.543 ± 0.4981	0.546 ± 0.4979	0.7612	0.522 ± 0.4995	0.543 ± 0.4982	0.2782	0.1083	0.033 ± 0.1797	0.101 ± 0.3009	<0.0001*	<0.0001*
Rash	0.075 ± 0.2631	0.123 ± 0.3290	<0.0001*	0.066 ± 0.2488	0.128 ± 0.3342	<0.0001*	0.3815	0.086 ± 0.2797	0.128 ± 0.3339	<0.0001*	0.1391

(continued)

Table 2. Continued.

	TrainSet-PT			TestSet-PT				ExternalSet-PT			
Characteristic (*N* = 47)	Positive group (*N* = 9302)	Negative group (*N* = 4146)	*p*	Positive group (*N* = 2338)	Negative group (*N* = 1023)	*p*	*p*	Positive group (*N* = 3564)	Negative group (*N* = 1361)	*p*	*p*
Vomiting	0.230 ± 0.4205	0.194 ± 0.3952	<0.0001*	0.223 ± 0.4162	0.206 ± 0.4046	0.2787	0.9321	0.224 ± 0.4172	0.268 ± 0.4430	0.0017*	0.0036*
Headache	0.027 ± 0.1630	0.050 ± 0.2183	<0.0001*	0.019 ± 0.1374	0.053 ± 0.2236	<0.0001*	0.1393	0.027 ± 0.1619	0.057 ± 0.2310	<0.0001*	0.2503
Chest pain	0.013 ± 0.1128	0.006 ± 0.0743	<0.0001*	0.015 ± 0.1231	0.002 ± 0.0442	<0.0001*	0.7401	0.005 ± 0.0689	0.004 ± 0.0663	0.8656	0.0178*
Abdominal pain	0.045 ± 0.2081	0.084 ± 0.2780	<0.0001*	0.042 ± 0.2014	0.078 ± 0.2685	0.0001*	0.3417	0.088 ± 0.2839	0.129 ± 0.3356	0.0001*	<0.0001*
Diarrhea	0.054 ± 0.2268	0.056 ± 0.2303	0.6738	0.062 ± 0.2404	0.055 ± 0.2275	0.4302	0.3148	0.239 ± 0.4265	0.264 ± 0.4407	0.0760	<0.0001*
Lip cyanosis	0.081 ± 0.2729	0.125 ± 0.3309	<0.0001*	0.082 ± 0.2739	0.113 ± 0.3171	0.0056*	0.5516	0.072 ± 0.2587	0.067 ± 0.2498	0.5141	<0.0001*
Wheezing	0.287 ± 0.4526	0.136 ± 0.3426	<0.0001*	0.277 ± 0.4474	0.136 ± 0.3427	<0.0001*	0.4027	0.265 ± 0.4416	0.136 ± 0.3427	<0.0001*	0.0025*
Barking cough	0.033 ± 0.1784	0.011 ± 0.1047	<0.0001*	0.034 ± 0.1807	0.018 ± 0.1315	0.0036*	0.4013	0.054 ± 0.2263	0.066 ± 0.2485	0.1216	<0.0001*
Whoop	0.108 ± 0.3106	0.055 ± 0.2284	<0.0001*	0.103 ± 0.3041	0.061 ± 0.2386	<0.0001*	0.7609	0.148 ± 0.3547	0.107 ± 0.3085	0.0001*	<0.0001*
Bluish face	0.001 ± 0.0359	0.000 ± 0.0220	0.1097	0.001 ± 0.0358	0.002 ± 0.0442	0.6683	0.5355	0.001 ± 0.0167	0.001 ± 0.0271	0.5637	0.0150*
Bluish lips	0.002 ± 0.0451	0.000 ± 0.0220	0.0071*	0.002 ± 0.0413	0.004 ± 0.0624	0.3023	0.3668	0.001 ± 0.0374	0.001 ± 0.0271	0.4892	0.5689
Abdominal distension	0.011 ± 0.1031	0.008 ± 0.0915	0.1943	0.009 ± 0.0965	0.012 ± 0.1077	0.5535	0.9680	0.059 ± 0.2350	0.071 ± 0.2573	0.1150	<0.0001*
Nasal blockade	0.127 ± 0.3333	0.110 ± 0.3129	0.0037*	0.124 ± 0.3291	0.123 ± 0.3286	0.9714	0.8100	0.160 ± 0.3668	0.127 ± 0.3331	0.0025*	0.0001*
Anorexia	0.119 ± 0.3239	0.084 ± 0.2773	<0.0001*	0.121 ± 0.3262	0.098 ± 0.2970	0.0426*	0.3515	0.023 ± 0.1490	0.027 ± 0.1626	0.3790	<0.0001*
Poor sleep	0.031 ± 0.1738	0.021 ± 0.1449	0.0008*	0.034 ± 0.1807	0.027 ± 0.1632	0.3104	0.2753	0.004 ± 0.0647	0.003 ± 0.0541	0.4867	<0.0001*
Hyperemia of conjunctiva	0.011 ± 0.1041	0.017 ± 0.1306	0.0054*	0.009 ± 0.0921	0.012 ± 0.1077	0.4119	0.0779	0.013 ± 0.1117	0.016 ± 0.1261	0.3641	0.7654
Apatheia	0.636 ± 0.4812	0.698 ± 0.4591	<0.0001*	0.629 ± 0.4830	0.706 ± 0.4557	<0.0001*	0.7931	0.591 ± 0.4917	0.675 ± 0.4683	<0.0001*	0.0051*

(continued)

Table 2. Continued.

Characteristic ($N = 47$)	TrainSet-PT Positive group ($N = 9302$)	 Negative group ($N = 4146$)	 p	TestSet-PT Positive group ($N = 2338$)	 Negative group ($N = 1023$)	 p	 p	ExternalSet-PT Positive group ($N = 3564$)	 Negative group ($N = 1361$)	 p	 p
Convulsions	0.015 ± 0.1226	0.049 ± 0.2168	<0.0001*	0.012 ± 0.1068	0.056 ± 0.2294	<0.0001*	0.7884	0.015 ± 0.1210	0.039 ± 0.1935	<0.0001*	0.0571
Tonsillitis	0.018 ± 0.1347	0.031 ± 0.1723	0.0001*	0.018 ± 0.1328	0.031 ± 0.1741	0.0292*	0.9391	0.008 ± 0.0867	0.037 ± 0.1881	<0.0001*	0.0012*
Nodding respiration	0.207 ± 0.4049	0.103 ± 0.3043	<0.0001*	0.207 ± 0.4049	0.113 ± 0.3171	<0.0001*	0.6374	0.003 ± 0.0529	0.001 ± 0.0001	0.0016*	<0.0001*
Dry rales	0.023 ± 0.1492	0.014 ± 0.1164	0.0001*	0.024 ± 0.1516	0.013 ± 0.1120	0.0215*	0.9327	0.002 ± 0.0473	0.010 ± 0.1009	0.0048*	<0.0001*
Strawberry-like tongue	0.058 ± 0.2330	0.060 ± 0.2367	0.6570	0.057 ± 0.2316	0.051 ± 0.2197	0.4698	0.4720	0.003 ± 0.0502	0.006 ± 0.0764	0.1341	<0.0001*
Eyes on the turn	0.026 ± 0.1595	0.093 ± 0.2906	<0.0001*	0.022 ± 0.1461	0.088 ± 0.2833	<0.0001*	0.2175	0.006 ± 0.0801	0.021 ± 0.1420	0.0005*	<0.0001*
Ear discharging	0.000 ± 0.0104	0.013 ± 0.1123	<0.0001*	0.000 ± 0.0207	0.014 ± 0.1162	0.0003*	0.7251	0.001 ± 0.0167	0.001 ± 0.0271	0.5637	<0.0001*
Nasal mucosa hyperemia	0.011 ± 0.1031	0.038 ± 0.1920	<0.0001*	0.009 ± 0.0965	0.041 ± 0.1984	<0.0001*	0.9343	0.003 ± 0.0555	0.008 ± 0.0895	0.0548	<0.0001*
Nasal secretion	0.001 ± 0.0232	0.008 ± 0.0889	<0.0001*	0.000 ± 0.0207	0.013 ± 0.1120	0.0005*	0.2650	0.001 ± 0.0001	0.001 ± 0.0271	0.3175	0.0001*
Pruritus	0.029 ± 0.1688	0.068 ± 0.2509	<0.0001*	0.030 ± 0.1692	0.079 ± 0.2700	<0.0001*	0.3748	0.009 ± 0.0929	0.044 ± 0.2053	<0.0001*	<0.0001*
Somnolence	0.036 ± 0.1869	0.050 ± 0.2173	0.0005*	0.041 ± 0.1984	0.048 ± 0.2136	0.3833	0.4779	0.049 ± 0.2155	0.065 ± 0.2459	0.0369*	0.1866
Jaundice	0.179 ± 0.3837	0.290 ± 0.4539	<0.0001*	0.180 ± 0.3846	0.310 ± 0.4624	<0.0001*	0.4340	0.912 ± 0.2830	0.904 ± 0.2949	0.3647	<0.0001*

Note: *: $p < 0.05$, showing a statistically significant difference.

Table 3. Performance of Bayesian networks modeling on TestSet-PT.

Features set	Model	Ranking strategy	Precision	Recall	F1-score
All-Features (N = 47)	Bayesian network *w/o* initial graph	Random	0.7559	0.8011	0.7778
		Global impact	0.7239	**0.9722**	0.8299
		Descending	0.7356	0.8841	0.8030
		Ascending	**0.7746**	0.9042	**0.8344**
	Bayesian network *w/* initial graph	Random	0.7062	0.7126	0.7094
		Global impact	0.7519	0.7207	0.7360
		Descending	0.6915	0.7730	0.7300
		Ascending	0.7309	0.8191	0.7725
Fine-Features (*N* = 23)	Bayesian network *w/o* initial graph	Random	0.7730	0.9145	0.8378
		Global impact	0.7390	0.9773	0.8416
		Descending	0.7162	0.9577	0.8195
		Ascending	**0.7861**	**0.9889**	**0.8759**
	Bayesian network *w/* initial graph	Random	0.7304	0.6848	0.7068
		Global impact	0.7140	0.8969	0.7951
		Descending	0.7282	0.6647	0.6950
		Ascending	0.7386	0.8627	0.7958

Note: The numbers in bold are the best performance of models using different features in three metrics. w/ : with; w/o: without.

(0.7382), highest recall (0.9947), and the highest F1-score (0.8475) using Fine-Features. The metrics of other modeling approaches were shown significant differences with the performance of Bayesian network.

Table 6 illustrated the key predictors found by different models. For ease of comparison, we sorted the features of other models according to SHAP (sHapley Additive exPlanations)[28] values, taking the same number of features as the parent nodes of the pneumonia node, which were 11 and 6 features from All-Features and Fine-Features, respectively. Features were also ranked using SHAP to discover explanatory predictors for model interpretation as shown in Figure 4(c).

Discussion

Implications and findings

Our results revealed that feature selection based on OR values improved the performance of Bayesian networks (F1-score of 0.8759 vs. 0.8344 as shown in Table 3). However, feature selection did not show the same advantages in other methods. For example, the performance of CatBoost decreased slightly (F1-score of 0.8471 vs. 0.8351 as shown in Table 4). One possible reason is that linear models and additive tree models treat variables and outputs as directly related, which is easier than Bayesian networks to eliminate the influence of noncritical variables, so they performed roughly the same on both feature sets. However, feature selection based on OR value does reduce the search space of Bayesian network. Therefore, the selected variables have greater mutual information with the pneumonia node.

In this study, we also found that in the absence of the initial graph, the Bayesian network model outperformed models built with the initial graph. We believe this is due to redundant edges in the initial graph misleading the optimizer into local optima since the heuristic hill-climbing algorithm is very sensitive to the search starting point. The results shown in Table 3 also demonstrated that the effect ranking strategy roughly followed the following

Table 4. Performance comparison on TestSet-PT.

Features set	Model	Precision/*p*	Recall/*p*	F1-score/*p*
All-Features (*N* = 47)	Bayesian network	0.7746**/–**	0.9042/–	0.8344**/–**
	CatBoost	**0.7852**/<0.0001*	0.9196/<0.0001*	**0.8471**/<0.0001*
	XGBoost	0.7804/<0.0001*	**0.9243**/<0.0001*	0.8463/<0.0001*
	LightGBM	0.7811/<0.0001*	0.9063/<0.0001*	0.8390 <0.0001*
	Logistic regression	0.7756/<0.0001*	0.9239/<0.0001*	0.8433/<0.0001*
	Ridge classifier	0.7561/<0.0001*	**0.9243**/<0.0001*	0.8318/<0.0001*
Fine-Features (*N* = 23)	Bayesian network	**0.7861**/–	**0.9889**/–	**0.8759**/–
	CatBoost	0.7726/<0.0001	0.9085/<0.0001*	0.8351/<0.0001*
	XGBoost	0.7710/<0.0001	0.9187/<0.0001*	0.8384/<0.0001*
	LightGBM	0.7752/<0.0001	0.8879/<0.0001*	0.8278/<0.0001*
	Logistic regression	0.7621/<0.0001	0.9192/<0.0001*	0.8333/<0.0001*
	Ridge classifier	0.7547/<0.0001	0.9277/<0.0001*	0.8323/<0.0001*

Note: The numbers in bold are the best performance of models using different features in three metrics. *: $p < 0.05$, showing a statistically significant difference.

Table 5. Performance comparison on externalSet-PT.

Features set	Model	Precision/*p*	Recall/*p*	F1-score/*p*
All-Features (*N* = 47)	Bayesian network	0.7378/–	0.9346/–	0.8246/–
	CatBoost	0.7448/<0.0001*	0.9450/<0.0001*	0.8330/<0.0001*
	XGBoost	0.7447/<0.0001*	0.9501/<0.0001*	0.8349/<0.0001*
	LightGBM	0.7502/<0.0001*	0.9220/<0.0001*	0.8273/<0.0001*
	Logistic regression	**0.7459**/<0.0001*	0.9537/<0.0001*	0.8371/<0.0001*
	Ridge classifier	0.7301/<0.0001*	**0.9837**/<0.0001*	**0.8382**/<0.0001*
Fine-Features (*N* = 23)	Bayesian network	**0.7382**/–	**0.9947**/–	**0.8475**/–
	CatBoost	0.7380/<0.0001*	0.9672/<0.0001*	0.8372/<0.0001*
	XGBoost	0.7363/<0.0001*	0.9739/<0.0001*	0.8386/<0.0001*
	LightGBM	0.7440/<0.0001*	0.9588/<0.0001*	0.8377/<0.0001*
	Logistic regression	0.7341/<0.0001*	0.9790/<0.0001*	0.8390/<0.0001*
	Ridge classifier	0.7288/<0.0001*	0.9863/<0.0001*	0.8382/<0.0001*

Note: The numbers in bold are the best performance of models using different features in three metrics. *: $p < 0.05$, showing a statistically significant difference.

Table 6. Results of the key predictors discovery.

		Models					
Features set	Key predictors	Bayesian network	CatBoost	XGBoost	LightGBM	Logistic regression	Ridge classifier
All-Features (*N* = 47)	Fever		√	√	√	√	√
	Chill						√
	Runny nose		√	√	√	√	
	Cough	√	√	√	√	√	√
	Expectoration		√	√	√	√	√
	Sore throat						√
	Dyspnea		√	√	√	√	√
	Throat stridor	√	√	√	√	√	√
	Three concave sign	√	√	√	√	√	√
	Moist rale		√	√	√		
	Wheezing	√	√	√	√	√	√
	Whoop	√					√
	Apatheia		√	√	√	√	
	Convulsions	√					
	Nodding respiration	√	√	√	√	√	√
	Eyes on the turn	√					
	Ear discharging	√					
	Nasal mucosa hyperemia	√					
	Pruritus	√					
	Jaundice					√	
Fine-Features (*N* = 23)	Fever		√	√	√		
	Cough	√	√	√	√	√	√
	Expectoration		√	√	√	√	√
	Dyspnea	√		√	√	√	√
	Throat stridor	√				√	

(continued)

Table 6. Continued.

		Models					
Features set	Key predictors	Bayesian network	CatBoost	XGBoost	LightGBM	Logistic regression	Ridge classifier
	Three concave sign		√	√	√	√	√
	Wheezing rale	√					
	Wheezing	√	√	√	√	√	√
	Nasal blockade	√					
	Nodding respiration		√				√

order, *Ascending* > *Global impact* > *Descending* ≈ *Random*. We argue that for hill climbing, the later the pneumonia node is added, the easier it is to traverse and optimize the edges connected to the pneumonia node in a global view. The clearer the dependency between the pneumonia node and other nodes, the better is the prediction effect. *Global impact* is not as effective as *Ascending*, probably because its original intention is to find a reasonable global network structure rather than infer-specific nodes.

As shown in Tables 2 and 5, although the characteristics of TrainSet-PT and ExternalSet-PT have significant statistical differences, the proposed Bayesian network still exhibited strong performance and outperformed all other models with Fine-Features. Table 5 demonstrated that our Bayesian network achieved good performance in EMRs from the independent external research center (F1-score of 0.8246 with All-Features and 0.8475 with Fine-Features). The impact of feature selection based on OR values on the external validation data was also significant, and the F1-score of the Bayesian network using Fine-Features was the highest.

Clinical significance of the identified key predictors

As shown in the All-Features rows of Table 6, all models considered cough, throat stridor, three concave sign, wheezing, and nodding respiration as significant predictors. The occurrence of pneumonia is accompanied by cough, so cough is very reliable as a key predictor of confirmed pneumonia. The throat stridor is a typical clinical manifestation of laryngeal obstruction. When this symptom occurs, it often indicates that the larynx has been narrowed due to infection. Laryngeal infections are often accompanied by lower respiratory tract infections. While throat stridor is not a typical symptom of pneumonia, throat infection often accompanies infections. This may reflect some characteristics of the cases in our data, that is, a large proportion of pneumonia cases also have symptoms of throat infection, and pneumonia is likely caused by aggravation of throat infection. The three-concave sign, also known as the intercostal retraction sign, appears in patients with severe pneumonia and is a common manifestation of severe pneumonia. It is therefore not surprising that the models found three concave signs to be strong predictors of a pneumonia diagnosis. Nodding respiration is also a typical symptom of severe pneumonia, and its presence is the diagnosis of pneumonia. Wheezing is a gasping sound during the exhalation phase. It is usually caused by the stenosis below the tracheal carina. It is a typical manifestation of lower respiratory tract infection with stenosis. Wheezing not only indicates lower airway infection, but also lower airway narrowing.

In the All-Features rows of Table 6, there were several predictors (i.e. convulsions, eyes on the turn, ear discharging, nasal mucosa hyperemia, and pruritus) selected by the Bayesian network only. Convulsions are not typical symptoms of pneumonia, but in children with pneumonia, repeated high fever may cause symptoms of systemic convulsions. Eyes on the turn is a typical symptom of febrile convulsion in children. Common causes of febrile convulsions include upper respiratory tract infection, tympanitis, and pneumonia. As shown in Figure 1(a), our data included at least 1073 children with febrile convulsions. The Bayesian network identified the eyes on the turn as a key predictor, probably because in our case most of the children with febrile convulsion also had pneumonia. *Streptococcus pneumoniae* frequently causes both tympanitis and pneumonia, which maybe the reason why ear discharge is considered a key predictor. This appears to indicate insufficient pneumococcal vaccination coverage in the region, in fact, where at least 202 children were diagnosed with tympanitis as shown in Figure 1(a). Nasal mucosa hyperemia often refers to upper respiratory tract infections, caused by rhinitis, nasosinusitis, and other diseases. It is not uncommon for upper respiratory tract infection to develop into pneumonia. Pruritus is a typical symptom of rash. Figure 1(a) showed that our data contained 194 rash cases. Symptoms of skin rash, which is not a typical symptom of pneumonia, have been reported in cases of coronavirus disease 2019 (COVID-19) at times.[29] The identification of pruritus as a key predictor by the Bayesian network may be caused by these atypical cases, which may also explain why dermatology patients are referred to respiratory departments. Rash was also a feature extracted from EMR in our dataset. However, it had a low OR value with pneumonia, likely due to rash, as a disease name, being rarely mentioned in descriptive text.

In the Fine-Features rows of Table 6, the consistency of the model predictors was not as strong as the All-Features row, while coughing and wheezing were still consistently identified as key predictors. The identification results of the tree models (CatBoost, XGBoost, and LightGBM) and linear models (logistic regression and ridge classifier) showed a high degree of agreement. Figure 3(c) showed that the child nodes of the pneumonia node were expectoration, moist rales, barking cough, fever, dyspnea and apathy, respectively. A comparison with the Fine-Features of Table 6 showed that these features were also key predictors for other models, except for barking cough. The difference is that other models treat these features as causes of pneumonia, while Bayesian networks treat them as effects of pneumonia. What they have in common is that they all revealed a correlation between these features and pneumonia.

Advantages and clinical significance of the Bayesian network

Currently, the first step in screening children with suspected CAP is a rapid assessment to identify signs and make a subjective diagnosis based on expert experience. Our proposed Bayesian network will provide automated, rapid, and objective assessment while reducing the workload of specialists. CAP in children often appears as a complication of diseases, such as pertussis and influenza, and the presence of multiple symptoms can make it difficult to identify. Our proposed Bayesian network provides a second opinion, increasing the number of accurate diagnostics and yielding additional new insights. The most significant advantage of Bayesian network modeling is that it is far easier to visually understand than other common classical methods.

Existing research suggests that linear or additive tree models may yield more accurate classifications as they only consider direct relationships between input and output variables. However, the variable relationship-capture capability of the Bayesian networks has greater value for data exploration. Interpretability of tree and linear models often depends on the SHAP interpreter and SHAP value ranking. The structure of the Bayesian network is far more intuitive and easier to interpret due to causal relationships between nodes. Furthermore, Bayesian network modeling reveals relationships between various symptoms and complications in addition to direct input–output connections, which is extremely valuable for common clinical applications and research.

Limitations and future expectations

Although our experiments on the external validation set demonstrated the generalizability of the model, more validation is necessary for extrapolating to more centers. One of the future objectives is to collect more data from different individual hospitals and conduct a multicenter study. As a feasibility study, this work only included hospital admission, inpatient, and discharge records. Additional information such as epidemiology, past history of respiratory diseases, and comorbidities should also be included for a complete and accurate diagnosis.

Furthermore, the use of the proposed method is limited by the scope of the specification written by EMR. This specification means following a pre-arranged structure and using a standard vocabulary. In practice, the method relies on a pre-built lookup table to identify and align nonstandard terms to designated terms, and uses regular expressions to extract the values of 47 or 23 features from the EMR (i.e. whether the term is mentioned). Hence, future directions should inevitably include additional regular expressions to accommodate the EMR of specific centers.

For clinical application, one of the purposes of rapid assessment is to diagnose pneumonia in children as mild or severe. This diagnostic difference has critical implications for the course of treatment. Mild cases only require a prescription, whereas severe cases require routine blood tests and lung X-rays. One of our future directions is to extend our model to the diagnosis of mild and severe pneumonia.

Additionally, while models built with Fine-Features performed better in predictions, models constructed with All-Features were more interpretable. Therefore, we believe that combining large networks with higher interpretability with small networks with higher classification accuracy is another promising future approach.

Conclusions

In this study, EMR data were used to construct a Bayesian network for pneumonia diagnosis. The application of feature selection based on OR values proved that the classification performance of the Bayesian networks matches the performance of commonly used machine learning algorithms. The performance on independent external validation data demonstrated the clinical generalizability of our Bayesian network. Analysis of the key predictors identified by the network further increases our understanding of the conditions and complications of pneumonia patients. The findings of this study have important clinical value and practical significance for the study of pediatric pneumonia in the field and the rapid and automated diagnosis of pneumonia.

Acknowledgments: The authors would like to thank the Children's Hospital, Zhejiang University School of Medicine and the Zhengzhou Children's Hospital for support in this research.

Contributorship: YiW, QS, XL, and YY were involved in conception and design of the study. GY, YY, JL, YL, and YiW were part of administrative support. JL, GY, and YW provided the study materials or patients. JL, FS, YuW, and SL were involved in collection and assembly of data; QS, ZX, XL, YY, and YiW contributed toward data analysis and interpretation. All the authors were involved in manuscript writing and final approval of the manuscript.

Declaration of Conflicting Interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The authors are accountable for all the aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study approval was granted by The Institutional Review Board (IRB) of the Medical Ethics Committee of Children's Hospital, Zhejiang University School of Medicine, China (IRB approval ID: 2020-IRB-058). The procedures were performed in accordance with the Declaration of Helsinki (as revised in 2013) and relevant guidelines and regulations. Informed consent was obtained after the procedure was fully explained to all participants and their legal guardians.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported in part by grants from the National Key Research & Development Program (grant number 2019YFE0126200), the National Natural Science Foundation of China (grant number 62076218), the Zhejiang Province Research Project of Public Welfare Technology Application (grant number LGF22H180004), and the Hong Kong Research Grants Council through General Research Fund (grant number 17207722).

Guarantor: GY

ORCID iD: Qiuyang Sheng https://orcid.org/0000-0002-4140-9094

References

1. Liu L, Oza S, Hogan D, et al. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the sustainable development goals. *Lancet* 2016; 388: 3027–3035. Epub 11 Nov 2016. Erratum in: Lancet. 13 May 2017; 389(10082):1884. PMID: 27839855; PMCID: PMC5161777.
2. Perin J, Mulick A, Yeung D, et al. Global, regional, and national causes of under-5 mortality in 2000-19: an updated systematic analysis with implications for the sustainable development goals. *Lancet Child Adolesc Health* 2022; 6: 106–115.. Epub 17 Nov 2021. Erratum in: Lancet Child Adolesc Health. 2022 Jan;6(1):e4. PMID: 34800370; PMCID: PMC8786667.
3. Jain S, Williams DJ, Arnold SR, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. *N Engl J Med* 2015; 372: 835–845.. PMID: 25714161; PMCID: PMC4697461.
4. Harris M, Clark J, Coote N, et al. British Thoracic Society guidelines for the management of community acquired pneumonia in children: update 2011. *Thorax* 2011; 66: ii1–i23. PMID: 21903691.
5. Wallihan R and Ramilo O. Community-acquired pneumonia in children: current challenges and future directions. *J Infect* 2014; 69: S87–S90.. Epub 26 Sep 2014. PMID: 25264163.
6. Prosperi MC, Marinho S, Simpson A, et al. Predicting phenotypes of asthma and eczema with machine learning. *BMC Med Genomics* 2014; 7: S7.. Epub 8 May 2014. PMID: 25077568; PMCID: PMC4101570.
7. Sun X, Douiri A and Gulliford M. Applying machine learning algorithms to electronic health records to predict pneumonia after respiratory tract infection. *J Clin Epidemiol* 2022; 145: 154–163.. Epub ahead of print. PMID: 35045315.
8. Giang C, Calvert J, Rahmani K, et al. Predicting ventilator-associated pneumonia with machine learning. *Medicine (Baltimore)* 2021; 100: e26246.. PMID: 34115013; PMCID: PMC8202554.
9. Yu G, Li Z, Li S, et al. The role of artificial intelligence in identifying asthma in pediatric inpatient setting. *Ann Transl Med* 2020; 8: 1367. PMID: 33313112; PMCID: PMC7723595.
10. Lee SM and Abbott PA. Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *J Biomed Inform* 2003; 36: 389–399. PMID: 14643735.
11. Heckerman D. Bayesian networks for data mining. *Data Min Knowl Discov* 1997; 1: 79–119.
12. Aronsky D and Haug PJ. Automatic identification of patients eligible for a pneumonia guideline. In: Marc Overhage J (eds) AMIA 2000. *Proceedings of the American Medical Informatics Association Symposium*; 2000, Nov 4–8; Los Angeles, CA: American Medical Informatics Association; 2000. p. 12–16.
13. Burnside ES, Rubin DL, Fine JP, et al. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. *Radiology* 2006; 240: 666–673.. PMID: 16926323.
14. Zhao G, Feng Q, Chen C, et al. Diagnose like a radiologist: Hybrid neuro–probabilistic reasoning for attribute–based medical image diagnosis. *IEEE Trans Pattern Anal Mach Intell* 2021; 1: 1–1.
15. Spyroglou II, Spöck G, Rigas AG, et al. Evaluation of Bayesian classifiers in asthma exacerbation prediction after medication discontinuation. *BMC Res Notes* 2018; 11: 522.. PMID: 30064478; PMCID: PMC6069881.
16. Sanders DL and Aronsky D. Detecting asthma exacerbations in a pediatric emergency department using a Bayesian network. *AMIA Annu Symp Proc* 2006; 2006: 684–688. PMID: 17238428; PMCID: PMC1839558.
17. He H and Choi J D. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In: Marie-Francine M, Xuanjing H, Lucia S and Scott WY (eds) *EMNLP 2021. Proceedings of the 2021 Conference on Empirical Methods in Natural Language*; 2021 Nov 7-11; Punta Cana, Dominican Republic. Stroudsburg, PA: Association for Computational Linguistics; 2021, p. 5555–5577.

18. Chapman WW, Bridewell W, Hanbury P, et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001; 34: 301–310. PMID: 12123149.
19. Ankan A and Panda A. pgmpy: Probabilistic graphical models using python. In: Kathryn H and James B (eds) SciPy 2015. Proceedings of the 14th python in science conference; 2015 July 6-12. Austin Texas: SciPy; 2015. p. 6–11.
20. Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003; 56: 1129–1135. PMID: 14615004.
21. Shen Y, Zhang L, Zhang J, et al. CBN: constructing a clinical Bayesian network based on data from the electronic medical record. *J Biomed Inform* 2018; 88: 1–10. Epub 3 Nov 2018. PMID: 30399432.
22. Carvalho A M. Scoring functions for learning Bayesian networks. *Inesc-id Tec Rep* 2009; 12: 1–48.
23. Koller D and Friedman N. *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT press, 2009.
24. Prokhorenkova L, Gusev G, Vorobev A, et al. Catboost: Unbiased boosting with categorical features. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (eds) NeurIPS 2018. Proceedings of 32nd Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montréal, Canada. 57 Morehouse Lane, Red Hook, NY, USA: Curran Associates Inc.; 2018. p. 6638–6648.
25. Varoquaux G, Buitinck L, Louppe G, et al. Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Computing and Communications* 2015; 19(1): 29–33.
26. Chen T and Guestrin C. Xgboost: A scalable tree boosting system. In: Balaji K, Mohak S, Alex S, Charu A, Dou S and Rajeev R (eds) KDD 2016. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17. San Francisco, CA: Association for Computing Machinery; 2016. p. 785–794.
27. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. In: Ulrike VL, Isabelle G, Samy B, Hanna W and Rob F (eds) NIPS 2017. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Beach, CA: Curran Associates Inc. 2017. p. 3149–3157.
28. Lundberg SM and Lee SI. A unified approach to interpreting model predictions. In: Ulrike VL, Isabelle G, Samy B, Hanna W and Rob F (eds) NIPS 2017. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Beach, CA: Curran Associates Inc. 2017. p. 4768–4777.
29. Pagali S and Parikh RS. Severe urticarial rash as the initial symptom of COVID-19 infection. *BMJ Case Rep* 2021; 14: e241793. PMID: 33766974; PMCID: PMC8006826.