


Article

Healthcare Cost Prediction Based on Hybrid Machine Learning Algorithms

Shujie Zou ¹, Chiawei Chu ^{1,*} , Ning Shen ² and Jia Ren ³¹ Faculty of Data Science, City University of Macau, Macau 999078, China; d21092100231@cityu.mo² Department of Innovation, Technology and Entrepreneurship, United Arab Emirates University, Al Ain P.O. Box 15551, United Arab Emirates; ningshen@uaeu.ac.ae³ School of Information and Communication Engineering, Hainan University, Haikou 570100, China; renjia@hainanu.edu.cn

* Correspondence: cwchu@cityu.mo

Abstract: Healthcare cost is an issue of concern right now. While many complex machine learning algorithms have been proposed to analyze healthcare cost and address the shortcomings of linear regression and reliance on expert analyses, these algorithms do not take into account whether each characteristic variable contained in the healthcare data has a positive effect on predicting healthcare cost. This paper uses hybrid machine learning algorithms to predict healthcare cost. First, network structure learning algorithms (a score-based algorithm, constraint-based algorithm, and hybrid algorithm) for a Conditional Gaussian Bayesian Network (CGBN) are used to learn the isolated characteristic variables in healthcare data without changing the data properties (i.e., discrete or continuous). Then, the isolated characteristic variables are removed from the original data and the remaining data used to train regression algorithms. Two public healthcare datasets are used to test the performance of the proposed hybrid machine learning algorithm model. Experiments show that when compared to popular single machine learning algorithms (Long Short Term Memory, Random Forest, etc.) the proposed scheme can obtain similar or higher prediction accuracy with a reduced amount of data.

Keywords: healthcare costs; CGBN; regression algorithm; hybrid algorithm**MSC:** 68T09

Citation: Zou, S.; Chu, C.; Shen, N.; Ren, J. Healthcare Cost Prediction Based on Hybrid Machine Learning Algorithms. *Mathematics* **2023**, *11*, 4778. <https://doi.org/10.3390/math11234778>

Academic Editor: Jonathan Blackledge

Received: 20 October 2023
Revised: 20 November 2023
Accepted: 25 November 2023
Published: 27 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the birth of various advanced medical technologies, the safety of human life has been greatly guaranteed; however, this brings with it larger medical expenses, which represent a great challenge for many patients [1]. Data collected by the Centers for Medicare and Medicaid Services shows that the U.S. spent a larger share of its gross domestic product on healthcare in 2018, increasing by 4.6% from the previous year [2]. Nonetheless, even very large healthcare expenditures may not provide appropriate and affordable healthcare for patients [3]. If the healthcare expenditures can be foreseen in advance, more precise services and treatments can be provided to patients. Thus, predicting healthcare costs can provide protection for patients while assisting healthcare organizations, e.g., drug manufacturers.

Currently, the study of healthcare costs is receiving attention from many researchers. For example, Kharat [4] used descriptive statistics to study the trend of chronic kidney disease in diabetic and non-diabetic patients from 2002 to 2016 and combined it with the associated quality of life to derive the healthcare expenditure of the patients. Yassine [5] used a cross-sectional study to analyze the healthcare expenditures of Moroccan basic health insurers from 2009 to 2014. Zhang studied data on the medical expenditures of lung cancer patients in thirteen provinces in China from 2002 to 2011, deducing that medical and

chemotherapy fees were the main factors that increased patients' healthcare cost [6]. Ma [7] collected data on the healthcare expenditures of selected middle-aged and elderly people in Beijing and analyzed the collected data using chi-square tests, t-tests, multivariate analysis, and linear regressions. Gong collected data on the healthcare expenditures of hemophiliacs from China's national insurance database from 2014 to 2016, compared the healthcare cost of employees and residents using the Kolmogorov–Smirnov test, and finally speculated on the factors affecting the healthcare expenditures of hemophiliacs using quantile regression [8]. Yang [9] collected data on medical expenditures of strabismus patients in the First Affiliated Hospital of Harbin Medical University, China and analyzed anesthesia as a major factor influencing medical expenditure of strabismus patients. Wang used Markov and two-part models to analyze healthcare expenditure for the elderly in China from 2011 to 2015 and make predictions about healthcare expenditures for the elderly in China from 2020 to 2060 [10]. Han [11] collected data from the 2018 Peking University Chinese Household Panel Study, used a Heckman sample selection model to analyze the data, and speculated on the extent to which the internet influences personal healthcare expenditures.

The above literature shows good results obtained in the study of healthcare expenditure; however, research in this area requires a large amount of data and extensive expert experience, and the algorithms used encounter difficulty when learning the information contained in the data [2]. With the rapid development of computer technology, researchers have begun to use complex machine learning algorithms to analyze healthcare expenditures. For example, Morid [12] used multiple supervised learning algorithms to learn from a large amount of healthcare cost data and make cost predictions, concluding that artificial neural networks can realize superior performance in the prediction of healthcare costs. Kaushik analyzed time-series healthcare cost data using LSTM (Long Short-Term Memory), CNN (Convolutional Neural Network), and Ensemble Learning to predict the average weekly spending of patients on two pain medications [13–15]. Yang [16] used machine learning algorithms to predict future healthcare expenditures and analyze the temporal correlation of patients' healthcare expenditures, concluding that more historical data leads to better predictive performance on the part of machine learning algorithms. Kuo [17] used machine learning algorithms such as Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest to analyze the healthcare expenditure data of spinal fusion patients in Taiwan from 2021 to 2023 and predict the healthcare expenditures of patients, with Random Forest showing optimal predictive performance. Zeng [18] built a multi-layer self-attention model to learn the relationship between medical codes and medical visits in order to predict future medical expenditures and diseases. The above studies have obtained better results in predicting healthcare cost; however, most of them focus on time-series healthcare expenditure data and require a large amount of data and expertly selected characteristic variables. Meanwhile, many researchers are using advanced single algorithms and expert experience to obtain high prediction accuracy with little consideration of the need to reduce data dimensionality. Irrelevant characteristic variables may represent a kind of "noise" in the dataset that can affect the prediction accuracy of the prediction model. This paper studies non-temporal and small amounts of data that contain partial information (age, gender, previous disease history, etc.) about each individual. More importantly, it focuses on identifying irrelevant characteristic variables in healthcare cost data as a way to reduce the amount of data and the amount of time required by the regression model to analyze the data while improving the prediction accuracy of regression models, an approach that has received little attention from researchers.

2. Related Work

Currently, researchers are using CGBN to study healthcare expenditures. For example, Wang [19] used CGBN to analyze data on the healthcare expenditures of lung cancer patients in Taiwan and predict healthcare expenditures based on different disease levels of lung cancer patients. In addition, researchers have used CGBN in other fields of research; for example, Hu [20] used CGBN to process seismic data with a mixture of discrete and

continuous variables, then applied CGBN to the prediction of earthquakes in Canterbury from 2010 to 2011; the experimental results showed that the prediction performance of CGBN was better than that of algorithms such as neural networks and support vector machine. Liu in [21] used CGBN to mine gene loci for carotenoid components of maize, finding that CGBN exhibited better performance than other algorithms in the experiment. In this paper, CGBN is used to learn isolated feature variables (variables that do not affect the target variable) from healthcare expenditure data, then regression algorithms are used to learn the data with the isolated variables removed. CGBN plays a key role in filtering data (reducing the amount of data) in the research presented in this paper.

3. Prediction Model

In this paper, CGBN is combined with regression algorithms to form hybrid machine learning models used to predict healthcare cost. First, the multiple structural learning algorithms of CGBN are used to learn the information in the dataset in order to build multiple network structures. Then, the number of occurrences of each isolated node in all network structures is counted and the feature variables corresponding to the isolated nodes with the highest number of occurrences are removed from the original dataset. Finally, the regression algorithms are used to learn the processed data and make predictions. The workflow block diagram of the proposed prediction model is shown in Figure 1. The modules in the dotted box in Figure 1 are the steps in which the hybrid machine learning model analyses and processes the data. The processing of the Analysis Module in Figure 1 is shown in Figure 2. The dots in the left three boxes of Figure 2 represent isolated nodes learned by network structure algorithms. Next, the structure learning algorithms used to construct the CGBN are presented.

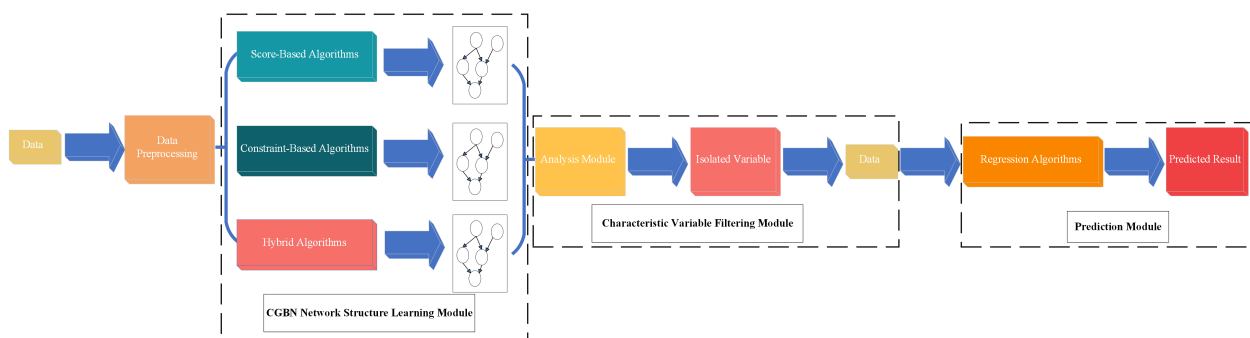


Figure 1. Block diagram of the workflow of the hybrid algorithm model.

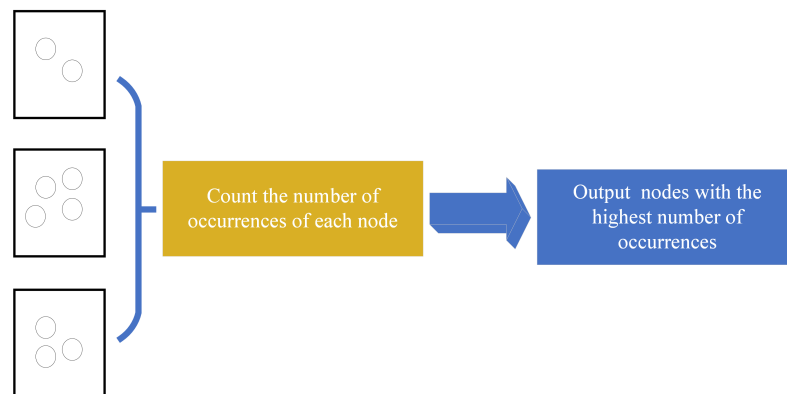


Figure 2. Block diagram of the workflow of the analysis module.

3.1. Network Structure Learning Algorithm

The attributes of the feature variables (continuous or discrete) are not changed when constructing the CGBN, which does not result in the loss of information in the dataset. In addition, the network structure of the CGBN has the following rules:

1. The nodes corresponding to discrete variables can only have nodes constructed by discrete variables as their parents.
2. The nodes corresponding to continuous variables can have either nodes constructed by discrete variables or nodes constructed by continuous variables as their parents.

Based on the properties of CGBN, in this paper we select two commonly used algorithms from each network structure algorithm to learn the network structure. Next, all network structure learning algorithms used in this paper are introduced in turn.

3.1.1. Score-Based Algorithms

In this paper, two score-based algorithms (the Hill Climbing and Tabu algorithms) for learning network structures are used. The BIC (Bayesian Information Criterion) algorithm for evaluating the structure of the network constructed by the Hill Climbing algorithm and the Tabu algorithm is

$$Score(G, D) = \sum_{i=1}^{k_D} (\log P(v_i|pa(v_i)) - \frac{d_{v_i}}{2} \log N) + \sum_{j=1}^{k_C} (\log f(v_j|pa(v_j)) - \frac{d_{v_j}}{2} \log N) \quad (1)$$

where D is the dataset, G is the directed acyclic graph, N is the number of data, $P(\cdot)$ is the conditional probability, $f(\cdot)$ is the probability density function, d_{v_i} and d_{v_j} are the number of network node parameters, k_D is the number of discrete network nodes, k_C is the number of continuous network nodes, $pa(v_i)$ and $pa(v_j)$ are the parents of v_i and v_j , respectively, and v_i and v_j are network nodes, where $v_i, v_j \in V = \{v_1, v_2, \dots, v_n\}$, $k_D + k_C = n$ (with n being the total number of nodes or the number of feature variables), and V is the set of network nodes. The first term on the right side of the equation represents the formula for calculating the scores of the nodes corresponding to discrete variables, while the second term represents the formula for calculating the scores of the nodes corresponding to continuous variables. It is worth noting that when there are nodes of discrete variables in the parent nodes of continuous variables, the expression for the second term on the right side of the equation can be further expressed as follows:

$$\sum_{j=1}^{k_C} (\log f(v_j|pa(v_j)) - \frac{d_{v_j}}{2} \log N) = \sum_{j=1}^{k_C} (\sum_{c \in v_d} \log f(v_j|v_d = c, pa(v_{j \neq d})) - \frac{d_{v_j,c}}{2} \log N) \quad (2)$$

where v_d (discrete variable) represents the parent of node v_j , c is the value of node v_d , and $d_{v_j,c}$ is the number of network node parameters. Next, the fundamentals of the Hill Climbing algorithm and Tabu algorithm are introduced.

The basic principle of both the Tabu and Hill Climbing (HC) algorithms is to start the search from the network structure (usually an empty graph), then add, delete, or reverse arcs on the network structure until $Score(G, D)$ cannot be improved any more. The Tabu algorithm overcomes the shortcomings of the HC algorithm to a certain extent, and has a better ability to learn the network structure [22].

3.1.2. Constraint-Based Algorithms

Two constraint-based network structure learning algorithms (the PC algorithm and Grow–Shrink algorithm) are used as well. The basic principle of the PC and the Grow–Shrink (GS) algorithms is to construct a Bayesian network structure using conditional independent testing. The GS algorithm differs from the PC algorithm in that the GS algorithm learns the Markov blanket of each node. The conditional independence test is based on the principle that, given any two nodes $v_p, v_q \in V (p \neq q)$, finding the subset

$V_S \subset V(v_p, v_q \notin V_S)$ enables the nodes v_p, v_q to be independent given a subset V_S or an arc that exists between the nodes v_p, v_q if no subset V_S exists. Because CGBNs are established in this paper, mutual information is used to test the conditional independence between nodes.

3.1.3. Hybrid Algorithms

Two hybrid algorithms (the Max–Min Hill-Climbing (MMHC) and Restricted Maximization (rsmx2) algorithms) are used to learn the network structure of CGBN. These hybrid algorithms combine the benefits of constraint algorithms and score algorithms to learn the network structure of the CGBN. The MMHC algorithm combines Max–Min Parents and Children (the MMPC constraint algorithm) and HC score algorithm. First, the MMPC algorithm learns the candidate parent nodes for each node $v_m \in V$ to form the set C_m , then searches for the network structure that maximises the BIC score under the constraints of the set of parent nodes C_m . In this paper, the combination of rsmx2 is set to be the same as that of MMHC, with the difference that the rsmx2 algorithm can repeat the network structure learning process of the MMHC algorithm until convergence.

With all of the network structure learning algorithms for CGBN used in this paper described above, we next turn to the regression algorithms.

3.2. Regression Algorithms

3.2.1. Linear Regression (LR)

The multiple linear regression algorithm used in the paper is a simple and practical machine learning algorithm that plays an important role in regression research. LR algorithms continue to occupy an important place in practical research. Taking the i -th instance as an example, the expression for the LR is

$$\overline{y_{l,i}} = w_1x_{i,1} + w_2x_{i,2} + \dots + w_nx_{i,n} + w_0, \tag{3}$$

where $w_{i \neq 0}$ is the partial regression coefficient, w_0 is the constant term, $\overline{y_{l,i}}$ is the predicted value, $x_{i,n}$ is the value of the n -th feature variable for the i -th instance, and n is the number of feature variables.

3.2.2. Support Vector Regression (SVR)

Support Vector Machine algorithms are classical machine learning algorithm that use appropriate kernel functions (linear kernel function, Gaussian kernel function, sigmoid kernel function, polynomial kernel function, etc.) and parameters for the analysis of classification or regression. SVR is a part of Support Vector Machine algorithms, and can be used for the prediction of continuous data. SVR builds a hyperplane that tries to keep all sample points from the hyperplane as small as possible. Taking the i -th instance as an example, the objective function for finding the hyperplane can be described as follows:

$$\min_{b,W} \frac{\|W\|^2}{2} + B \sum_{i=1}^n (\overline{\xi}_i + \xi_i) \tag{4}$$

$$\overline{\xi}_i = \begin{cases} y_i - \overline{y_{s,i}} - \varepsilon & , \quad y_i > \overline{y_{s,i}} + \varepsilon \\ 0 & , \quad \text{other} \end{cases} \tag{5}$$

$$\xi_i = \begin{cases} \overline{y_{s,i}} - \varepsilon - y_i & , \quad y_i < \overline{y_{s,i}} - \varepsilon \\ 0 & , \quad \text{other} \end{cases} \tag{6}$$

$$s.t. \begin{cases} y_i - \overline{y_{s,i}} \leq \varepsilon + \overline{\xi}_i \\ \overline{y_{s,i}} - y_i \leq \varepsilon + \xi_i \\ \overline{\xi}_i, \xi_i \geq 0 \end{cases} \tag{7}$$

$$\overline{y_{s,i}} = w_{s,1}x_{i,1} + w_{s,2}x_{i,2} + \dots + w_{s,n}x_{i,n} + b \tag{8}$$

where $W = (w_{s,1}, w_{s,2}, \dots, w_{s,n})$ is the vector of coefficients, $\overline{y_{s,i}}$ is the predicted value of the i -th instance, y_i is the true value of the i -th instance. $\xi_i, \overline{\xi}_i$ are the slack variables, B is the regularisation constant, ε is the tolerance deviation, and b is a constant term. The correlation algorithm is used to solve the parameters of the above objective function to obtain the hyperplane of the SVR.

3.2.3. Backpropagation Neural Networks (BPnet)

Based on the superior performance of neural networks and improvements in computing power, neural networks can approximate the complex nonlinear relationships between variables. Therefore, neural networks provide better results in regression analysis. The simplest neural network architecture consists of three layers: an input layer, a hidden layer, and an output layer. In healthcare cost prediction, BPnet first learns the feature information of patients in the training dataset, then assigns appropriate weights to each feature variable and establishes the relationship between the feature variables and healthcare cost. Based on this, information about the characteristics of the patients in the test dataset can be input into the neural network to predict patients' healthcare costs.

3.2.4. Random Forest (RF)

RF consists of multiple decision trees, and belongs to the class of ensemble learning in machine learning. Because RF has better anti-interference ability, many researchers use RF to perform regression analysis research. The basic principle of RF for regression analysis is that the subset of data and the subset of features are randomly selected from the healthcare cost data to build each decision tree of the RF. The test set is input to the trained RF, which averages all the decision tree outputs to output a prediction.

3.2.5. Long Short-Term Memory (LSTM)

LSTM is a neural network with good ability to handle sequential data. It is an optimization of the RNN (Recursion Neural Network) model, and has the quality features of RNN. The LSTM structure contains input gates, forgetting gates, and output gates that determine the loss or preservation of information in the data to achieve forgetting and remembering, which overcomes the drawbacks of the single memory overlay approach of RNNs. LSTM currently plays an important role in many research areas, and many researchers have achieved good results using LSTM to predict healthcare costs.

4. Dataset

The datasets studied in the paper are all mixed datasets containing both discrete and continuous variables. In order to show the superior performance of the predictive model, two datasets are used to test the predictive ability of the hybrid model. Both datasets collected in this paper are from Kaggle. The first dataset is from a health insurance company and has 986 instances, each containing ten feature variables and one healthcare cost variable. A specific description of the first dataset is shown in Table 1. The second dataset has a total of 1338 instances, each containing six feature variables and one healthcare cost variable. A specific description of the second dataset is shown in Table 2. There are no missing values in either dataset. Before analyzing the data, characters in the dataset are replaced with numerical values (e.g., 1 for male and 0 for female).

Table 1. Description of the first dataset. × represents that the variable does not have the relevant attribute.

Variable	Description	Attribute	Min	Max	Mean	Standard Deviation
Age	Age of the patient	Continuous	18	66	41.75	13.96337
Diabetes	Whether the patient has diabetes	Discrete	×	×	×	×
Blood Pressure Problems	Whether the patient has blood pressure disease	Discrete	×	×	×	×
Any Transplants	Whether the patient has undergone transplant surgery	Discrete	×	×	×	×
Any Chronic Diseases	Whether the patient has any chronic diseases	Discrete	×	×	×	×
Height	Patient height	Continuous	145	188	168.2	10.09815
Weight	Patient weight	Continuous	51	132	76.95	14.2651
Known Allergies	Whether the patient has any allergies	Discrete	×	×	×	×
History of Cancer in Family	Whether there is any history of cancer in the patient’s family	Discrete	×	×	×	×
Number of Major Surgeries	The number of major surgeries the patient has undergone	Discrete	×	×	×	×
Charges	Healthcare cost	Continuous	15,000	40,000	24,337	6248.184

Table 2. Description of the second dataset. × represents that the variable does not have the relevant attribute.

Variable	Description	Attribute	Min	Max	Mean	Standard Deviation
Age	Age of the patient	Continuous	18	64	39.21	14.04996
Sex	Gender of the patient	Discrete	×	×	×	×
BMI	Body Mass Index of patient	Continuous	15.96	53.13	30.66	6.098187
Children	The number of children covered under the medical insurance	Discrete	×	×	×	×
Smoker	Whether the patient is a smoker or not	Discrete	×	×	×	×
Region	The geographic region of the patient	Discrete	×	×	×	×
Charges	Healthcare cost	Continuous	1122	63,770	13,270	12,110.01

5. Evaluation Method

To accurately and effectively test the predictive performance of the proposed model, MRE (Mean Relative Error), MSE (Mean Square Error), RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and SMAPE (Symmetric Mean Absolute Percentage Error) are used to test the model performance, all of which play important roles in regression analysis in many fields. Below, the expressions for these evaluation methods are provided in turn:

$$MRE = \frac{\sum_{i=1}^L \frac{|\bar{y}_i - y_i|}{y_i}}{L} \quad (9)$$

$$MSE = \frac{\sum_{i=1}^L (\bar{y}_i - y_i)^2}{L} \quad (10)$$

$$MAE = \frac{\sum_{i=1}^L |\bar{y}_i - y_i|}{L} \quad (11)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^L (\bar{y}_i - y_i)^2}{L}} \quad (12)$$

$$SMAPE = \frac{\sum_{i=1}^L \frac{|\bar{y}_i - y_i|}{(\bar{y}_i + y_i)/2}}{L} \quad (13)$$

where L is the number of instances in the test set, \bar{y}_i represents the predicted value, and y_i represents the true value.

6. Experimental Analysis

Rstudio was used to build the hybrid models used to predict healthcare cost; 80% of the number of instances of the healthcare data were used for the training set and 20% for the test set. The network structure algorithm for CGBN came from Rstudio's `bnlearn` package, and the LSTM was built using Rstudio's `keras` package. The dataset with ten feature variables and one target variable is defined as dataset A and the dataset with six feature variables and one target variable as dataset B. For simplicity of description, the hybrid models built by CGBN with the various regression algorithms are abbreviated as CGBN + RF, CGBN + SVR, CGBN + BPnet, CGBN + LR, and CGBN + LSTM.

6.1. Dataset A

The three classes of CGBN structure learning algorithm were first used to learn dataset A. Then, the multiple network structures learned from the CGBN network structure algorithm were analyzed to obtain isolated nodes. After analysis, there was one isolated node with the highest number of occurrences in multiple network structures. The results obtained by the hybrid algorithms after deleting the isolated node corresponding to the feature variables in dataset A are shown in Figure 3. The single models (e.g. RF, SVR, etc.) analyzed the original dataset with no reduction in feature variables. Figure 3 shows the prediction graphs for CGBN + RF, CGBN + SVR, CGBN + BPnet, CGBN + LR, and CGBN + LSTM. Each graph contains the results predicted by the hybrid model, the results predicted by the single model, and the true values in Figure 3. As can be seen from Figure 3, the trend of the prediction curves of the hybrid and single models in Figure 3a,c are very close to the trend of the true curves, which indicates that CGBN + RF, RF, CGBN + BPnet, and BPnet have a better prediction ability on dataset A. The prediction curve of CGBN + LR in Figure 3d almost overlaps with the prediction curve of LR, which indicates that the prediction ability of CGBN + LR is the same as that of LR in the case of reduced data. Moreover, Figure 3b,e exhibit better predictive power than Figure 3d. Overall, the prediction curves of the hybrid models with reduced data volume are similar to those of the single models, which reflects the reasonableness of the hybrid models.

In order to better show the performance advantages of the proposed model, MRE, MSE, RMSE, MAE, and SMAPE were used to evaluate the prediction results of both the hybrid and single models. The error analysis between the prediction results and the true values of the hybrid models and single models is shown in Table 3. As can be seen in Table 3, the error analyses of the various evaluation methods is not necessarily consistent for the hybrid models and single models. For example, the MRE of CGBN + RF is lower than that of RF, while the MRSE of CGBN + RF is slightly higher than that of RF. In general, the MRE better reflects the predictive power of the models in healthcare cost forecasting. As can be seen from Table 3, CGBN + RF has the lowest MRE, followed by CGBN + BPnet, which is in line with the trend of the predicted curves in Figure 3a,b. The MRE of CGBN + LR is essentially the same as that of LR, which is in line with the trend of the predicted curves in Figure 3d. More importantly, Table 3 shows that the MREs and SMAPEs of the hybrid models are lower than the MREs and SMAPEs of the corresponding single models in all cases where the amount of data is reduced, which fully reflects the superior predictive performance of the hybrid models. MSE, RMSE, and MAE can be inconsistent in their evaluation of the hybrid models and single models; however, each evaluation algorithm calculates similar error values for the hybrid and single models, which further reflects the validity of the hybrid models.

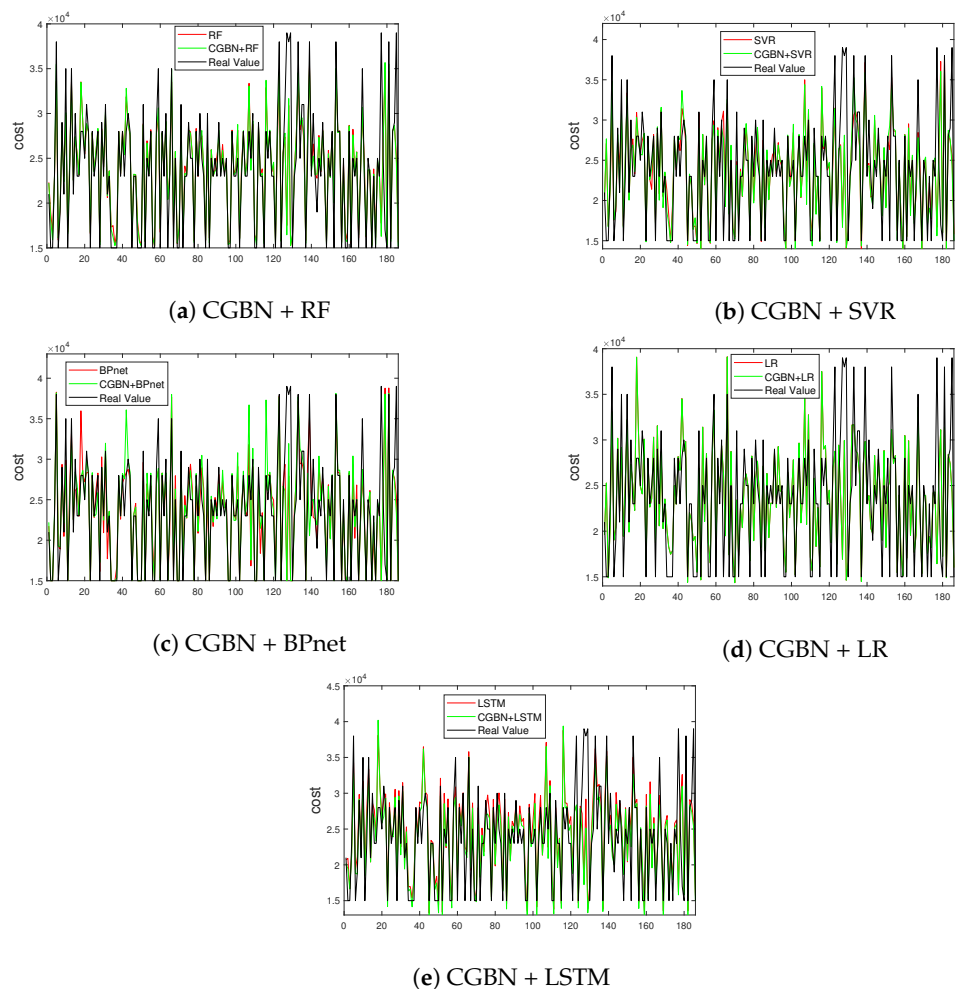


Figure 3. Predictive results for hybrid and single models; the horizontal axis represents the number of instances and the vertical axis represents the healthcare cost.

Table 3. Evaluation results for hybrid and single models.

Method Model	MRE	MAE	MSE	RMSE	SMAPE
CGBN + RF	0.071	1.79×10^3	1.49×10^7	3.86×10^3	0.072
CGBN + SVR	0.113	2.65×10^3	2.15×10^7	4.64×10^3	0.110
CGBN + BPnet	0.074	2.018×10^3	1.98×10^7	4.45×10^3	0.075
CGBN + LR	0.129	3.0865×10^3	2.28×10^7	4.77×10^3	0.127
CGBN + LSTM	0.118	2.83×10^3	2.11×10^7	4.59×10^3	0.119
RF	0.078	1.92×10^3	1.45×10^7	3.80×10^3	0.077
SVR	0.116	2.70×10^3	2.16×10^7	4.65×10^3	0.112
BPnet	0.077	2.047×10^3	1.82×10^7	4.26×10^3	0.078
LR	0.130	3.0869×10^3	2.28×10^7	4.77×10^3	0.127
LSTM	0.123	2.87×10^3	2.08×10^7	4.56×10^3	0.119

6.2. Dataset B

In order to test the generality of the hybrid models, dataset B was analyzed using the hybrid models. Similarly to dataset A, the three classes of CGBN structure learning algorithms were used to learn dataset B, then the multiple network structures learned from the CGBN network structure algorithm were analyzed to obtain isolated nodes. After analysis, there were two isolated nodes with the highest number of occurrences in multiple network structures. The prediction results obtained by the hybrid models after deleting the feature variables in dataset B corresponding to the isolated nodes are shown in Figure 4. The single models analyzed the original dataset with no reduction in feature variables. As can be seen from Figure 4, the trend of the prediction curves of the hybrid and single models in Figure 4b,e are very close to the trend of the true curves, which suggests that CGBN + SVR, SVR, CGBN + LSTM, and LSTM have better prediction ability on dataset B. Compared to the prediction curves shown in the other figures, the trends of the prediction curves of the hybrid and single models in Figure 4d deviate from the trend of the true curves more, indicating that the prediction performances of CGBN + LR and LR on dataset B are poor. Similar to the case of Figure 3, the prediction curves of the hybrid models with reduced data volumes are demonstrated in Figure 4 to be similar to that of the single models, further demonstrating the superior performance of the hybrid models.

Again, MRE, MSE, RMSE, MAE, and SMAPE were used to evaluate the prediction results of hybrid models and single models in order to highlight the prediction performance of each model. The error analysis between the prediction results and the true values of the hybrid models and single models is shown in Table 4. As can be seen from Table 4, the MRE of CGBN + SVR is the lowest among the hybrid models and the MRE of LSTM is the lowest among the single models, which is in line with the trend of the curves in Figure 4b,e. The MRE of CGBN + LR is the highest among the hybrid models, while the MRE of LR is the highest among the single models, which indicates that LR cannot analyze dataset B as well as the other regression algorithms. Although the prediction performance of CGBN + LSTM is lower than LSTM, the predictive performance of the other hybrid models is similar to or better than that of the corresponding single models. Overall, the hybrid models can obtain better prediction results with reduced data volume.

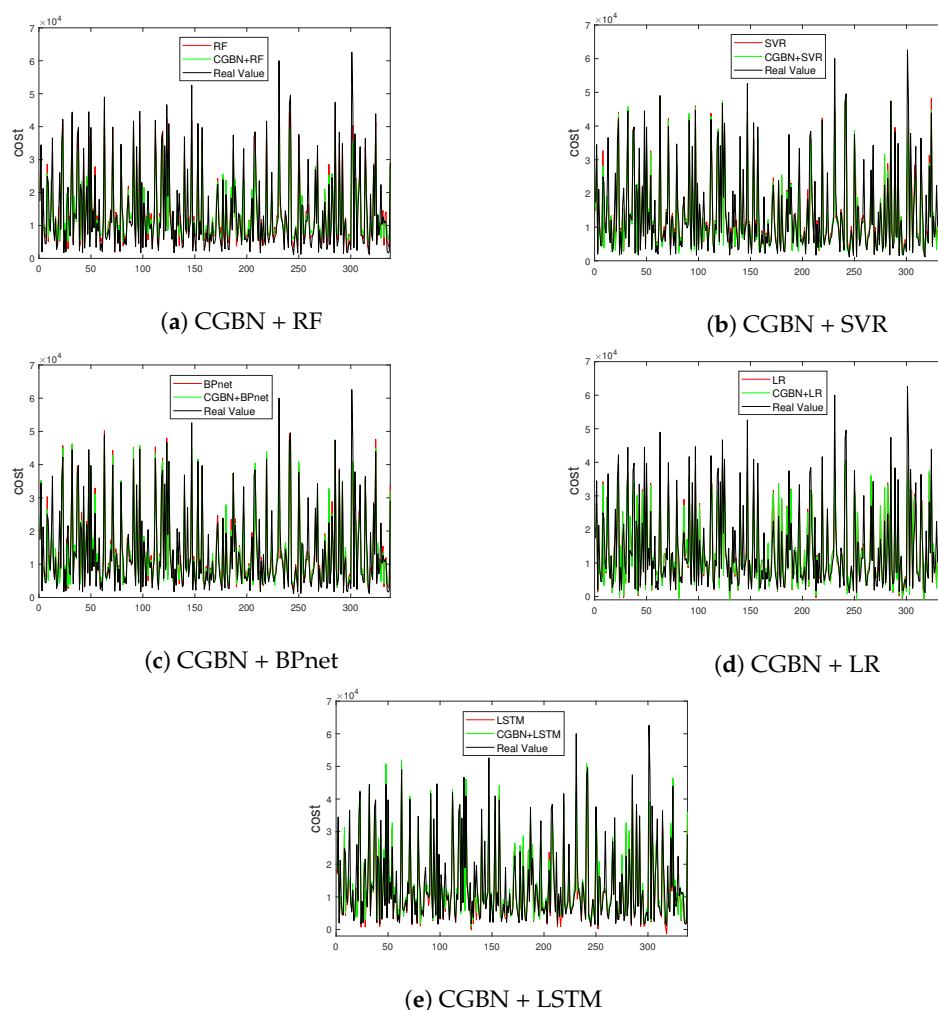


Figure 4. Predictive results for hybrid and single models. The horizontal axis represents the number of instances and the vertical axis represents the healthcare cost.

Table 4. Evaluation results for hybrid and single models.

Method Model	MRE	MAE	MSE	RMSE	SMAPE
CGBN + RF	0.27	3.91×10^3	3.57×10^7	5.97×10^3	0.27
CGBN + SVR	0.21	2.60×10^3	2.79×10^7	5.28×10^3	0.22
CGBN + BPnet	0.24	2.86×10^3	2.62×10^7	5.12×10^3	0.25
CGBN + LR	0.32	4.52×10^3	4.34×10^7	6.58×10^3	0.34
CGBN + LSTM	0.25	3.05×10^3	2.81×10^7	5.30×10^3	0.26
RF	0.24	3.28×10^3	2.93×10^7	5.41×10^3	0.25
SVR	0.25	2.91×10^3	2.77×10^7	5.26×10^3	0.26
BPnet	0.24	3.02×10^3	2.73×10^7	5.23×10^3	0.25
LR	0.32	4.46×10^3	4.29×10^7	6.55×10^3	0.34
LSTM	0.19	2.75×10^3	3.03×10^7	5.51×10^3	0.16

Although the addition of the CGBN algorithm increases the complexity of the overall prediction model, only the network structure learning algorithms of CGBN are used in the paper. During the experiment, the proposed model spends little computation and time in processing the data. Meanwhile, the feature variables filtered out by the proposed model

can provide a reference for staff and facilitate the collection and processing of data which cannot be done by a single algorithm.

7. Conclusions

This paper proposes combining CGBN with various regression algorithms to form hybrid models to predict healthcare costs. First, CGBN network structure learning algorithms reduce the amount of data in the dataset by deleting irrelevant information, then regression algorithms are used to learn the rest of the dataset, and finally regression algorithms make predictions. The predictive performance of the hybrid models was tested separately on two healthcare cost datasets, with the hybrid models obtaining better predictive results than the single models. Therefore, the proposed hybrid models can obtain better prediction performance with reduced data volumes, which can provide a corresponding reference for related workers and reduce the workload of data collection and processing. Although the CGBN structure learning algorithms can accurately identify the irrelevant variables in the dataset, it is difficult for the current CGBN network structure learning algorithms to independently learn the legitimate network structure from medical cost data. In future studies, we intend to optimise the CGBN network structure algorithm in order to enable them to learn legitimate network structures without relying on expert experience. In addition, time-series healthcare cost data will be investigated to further improve the applicability of the hybrid models.

Author Contributions: Conceptualization, S.Z. and C.C.; methodology, S.Z.; software, S.Z.; validation, S.Z. and C.C.; formal analysis, S.Z.; investigation, S.Z.; resources, S.Z. and C.C.; data curation, S.Z.; writing—original draft preparation, S.Z.; writing—review and editing, J.R. and N.S.; visualization, S.Z.; supervision, J.R. and N.S.; project administration, C.C.; funding acquisition, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: MOST-FDCT Projects: 0058/2019/AMJ Research and Application of Cooperative Multi-Agent Platform for Zhuhai–Macao Manufacturing Service.

Data Availability Statement: The dataset A is available at <https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction> (accessed 20 August 2023). The dataset B is available at <https://www.kaggle.com/datasets/harshsingh2209/medical-insurance-payout> (accessed 20 August 2023).

Conflicts of Interest: We declare that we have no financial and personal relationships with other people or organizations that could inappropriately influence our work, and no professional or other personal interests of any nature or kind in any product, service, and/or company that could be construed as influencing the position presented in or the review of the manuscript entitled “Healthcare Cost Prediction Based on Hybrid Machine Learning Algorithms”.

References

1. Kane, J. Health costs: How the US compares with other countries. *PBS Newshour* **2012**, *22*, 1–32.
2. Zeng, X.; Lin, S.; Liu, C. Multi-view deep learning framework for predicting patient expenditure in healthcare. *IEEE Open J. Comput. Soc.* **2021**, *2*, 62–71. [[CrossRef](#)]
3. Kaushik, S.; Choudhury, A.; Natarajan, S.; Pickett, L.A.; Dutt, V. Medicine Expenditure Prediction via a Variance- Based Generative Adversarial Network. *IEEE Access* **2020**, *8*, 110947–110958. [[CrossRef](#)]
4. Kharat, A.A.; Muzumdar, J.; Hwang, M.; Wu, W. Assessing trends in medical expenditures and measuring the impact of health-related quality of life on medical expenditures for US adults with diabetes associated chronic kidney disease using 2002–2016 medical expenditure panel survey data. *J. Pharm. Health Serv. Res.* **2020**, *11*, 365–373. [[CrossRef](#)]
5. Yassine, A.; Hangouche, A.J.; El Malhouf, N.; Maarouf, S.; Taoufik, J. Assessment of the medical expenditure of the basic health insurance in Morocco. *Pan Afr. Med. J.* **2020**, *35*, 115. [[CrossRef](#)] [[PubMed](#)]
6. Zhang, X.; Shi, J.-F.; Liu, G.-X.; Ren, J.-S.; Guo, L.-W.; Huang, W.-D.; Shi, L.-M.; Ma, Y.; Huang, H.-Y.; Bai, Y.-N.; et al. Medical expenditure for lung cancer in China: A multicenter, hospital-based retrospective survey. *Cost Eff. Resour. Alloc.* **2021**, *19*, 53. [[CrossRef](#)] [[PubMed](#)]
7. Ma, C.; Jiang, Y.; Li, Y.; Zhang, Y.; Wang, X.; Ma, S.; Wang, Y. Medical expenditure for middle-aged and elderly in Beijing. *BMC Health Serv. Res.* **2019**, *19*, 360. [[CrossRef](#)] [[PubMed](#)]

8. Gong, G.-w.; Chen, Y.-c.; Fang, P.-q.; Min, R. Medical expenditure for patients with hemophilia in urban China: Data from medical insurance information system from 2013 to 2015. *Orphanet J. Rare Dis.* **2020**, *15*, 137. [[CrossRef](#)]
9. Yang, L.; Min, Y.; Jia, Z.; Wang, Y.; Zhang, R.; Sun, B. Medical expenditure for strabismus: A hospital-based retrospective survey. *Cost Eff. Resour. Alloc.* **2022**, *20*, 27. [[CrossRef](#)] [[PubMed](#)]
10. Wang, L.; Tang, Y.; Roshanmehr, F.; Bai, X.; Taghizadeh-Hesary, F.; Taghizadeh-Hesary, F. The health status transition and medical expenditure evaluation of elderly population in China. *Int. J. Environ. Res. Public Health* **2021**, *18*, 6907. [[CrossRef](#)] [[PubMed](#)]
11. Han, J.; Zhang, X.; Meng, Y. The impact of internet medical information overflow on residents' medical expenditure based on China's observations. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3539. [[CrossRef](#)] [[PubMed](#)]
12. Morid, M.A.; Kawamoto, K.; Ault, T.; Dorius, J.; Abdelrahman, S. Supervised learning methods for predicting healthcare costs: Systematic literature review and empirical evaluation. *AMIA Annu. Symp. Proc.* **2018**, *2017*, 1312–1321. [[PubMed](#)]
13. Kaushik, S.; Choudhury, A.; Dasgupta, N.; Natarajan, S.; Pickett, L.A.; Dutt, V. Ensemble of multi-headed machine learning architectures for time-series forecasting of healthcare expenditures. In *Applications of Machine Learning*; Springer: Singapore, 2020; pp. 199–216.
14. Kaushik, S.; Choudhury, A.; Dasgupta, N.; Natarajan, S.; Pickett, L.A.; Bisht, D. Evaluating single-and multi-headed neural architectures for time-series forecasting of healthcare expenditures. In *Computational Intelligence Theoretical Advances and Advanced Applications*; De Gruyter Publisher: Berlin, Germany, 2020; pp. 159–176.
15. Kaushik, S.; Choudhury, A.; Sheron, P.K.; Dasgupta, N.; Natarajan, S.; Pickett, L.A.; Dutt, V. AI in healthcare: Time-series forecasting using statistical, neural, and ensemble architectures. *Front. Big Data* **2020**, *3*, 4. [[CrossRef](#)] [[PubMed](#)]
16. Yang, C.; Delcher, C.; Shenkman, E.; Ranka, S. Machine learning approaches for predicting high cost high need patient expenditures in health care. *Biomed. Eng. Online* **2018**, *17*, 131. [[CrossRef](#)] [[PubMed](#)]
17. Kuo, C.-Y.; Yu, L.-C.; Chen, H.-C.; Chan, C.-L. Comparison of models for the prediction of medical costs of spinal fusion in Taiwan diagnosis-related groups by machine learning algorithms. *Healthc. Inform. Res.* **2018**, *24*, 29–37. [[CrossRef](#)] [[PubMed](#)]
18. Zeng, X.; Feng, Y.; Moosavinasab, S.; Lin, D.; Lin, S.; Liu, C. Multilevel self-attention model and its use on medical risk prediction. In *Pacific Symposium on Biocomputing 2020*; World Scientific: Singapore, 2019; pp. 115–126.
19. Wang, K.-J.; Chen, J.-L.; Wang, K.-M. Medical expenditure estimation by Bayesian network for lung cancer patients at different severity stages. *Comput. Biol. Med.* **2019**, *106*, 97–105. [[CrossRef](#)] [[PubMed](#)]
20. Hu, J.; Wang, J.; Zhang, Z.; Liu, H. Continuous-discrete hybrid Bayesian network models for predicting earthquake-induced liquefaction based on the Vs database. *Comput. Geosci.* **2022**, *169*, 105231. [[CrossRef](#)]
21. Liu, J.; Kang, Y.; Liu, K.; Yang, X.; Sun, M.; Hu, J. Maize Carotenoid Gene Locus Mining Based on Conditional Gaussian Bayesian Network. *IEEE Access* **2020**, *8*, 15223–15231. [[CrossRef](#)]
22. Scutari, M.; Denis, J.-B. *Bayesian Networks: Examples R*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.