BMC Infectious Diseases

RESEARCH

Open Access

Predicting hospitalization costs for pulmonary tuberculosis patients based on machine learning

Shiyu Fan[1†], Abudoukeyoumujiang Abulizi[2,3†], Yi You[4†], Chencui Huang[4], Yasen Yimit[2,3], Qiange Li[1], Xiaoguang Zou[3,5*] and Mayidili Nijiati[3,6*]

Abstract

Background Pulmonary tuberculosis (PTB) is a prevalent chronic disease associated with a significant economic burden on patients. Using machine learning to predict hospitalization costs can allocate medical resources effectively and optimize the cost structure rationally, so as to control the hospitalization costs of patients better.

Methods This research analyzed data (2020–2022) from a Kashgar pulmonary hospital's information system, involving 9570 eligible PTB patients. SPSS 26.0 was used for multiple regression analysis, while Python 3.7 was used for random forest regression (RFR) and MLP. The training set included data from 2020 and 2021, while the test set included data from 2022. The models predicted seven various costs related to PTB patients, including diagnostic cost, medical service cost, material cost, treatment cost, drug cost, other cost, and total hospitalization cost. The model's predictive performance was evaluated using R-square (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) metrics.

Results Among the 9570 PTB patients included in the study, the median and quartile of total hospitalization cost were 13,150.45 (9891.34, 19,648.48) yuan. Nine factors, including age, marital status, admission condition, length of hospital stay, initial treatment, presence of other diseases, transfer, drug resistance, and admission department, significantly influenced hospitalization costs for PTB patients. Overall, MLP demonstrated superior performance in most cost predictions, outperforming RFR and multiple regression; The performance of RFR is between MLP and multiple regression; The predictive performance of multiple regression is the lowest, but it shows the best results for *Other costs*.

Conclusion The MLP can effectively leverage patient information and accurately predict various hospitalization costs, achieving a rationalized structure of hospitalization costs by adjusting higher-cost inpatient items and balancing different cost categories. The insights of this predictive model also hold relevance for research in other medical conditions.

Keywords Pulmonary tuberculosis, Multilayer perceptron, Cost prediction, Influencing factors, Machine learning

†Shiyu Fan, Abudoukeyoumujiang Abulizi and Yi You contributed equally to this work.

*Correspondence:
Xiaoguang Zou
zxgks@163.com
Mayidili Nijiati
mydl0911@163.com
Full list of author information is available at the end of the article

Background

Pulmonary tuberculosis (PTB) is a highly infectious disease caused by Mycobacterium tuberculosis (MTB) infection that poses a significant threat to human life and health. It has been referred to as "phthisis" in ancient Chinese folk medicine. This disease has been historically associated with large-scale epidemics and high mortality rates, resulting in the name "white plague" [1]. Global data from the World Health Organization (WHO) reveal that approximately 1.7 billion people worldwide have latent PTB infection, accounting for 23% of the global population. China is among the 30 countries with a high prevalence of PTB, with 780,000 new cases reported in 2021, representing 7.4% of the total global new cases and ranking third worldwide [2]. In Kashgar, epidemiological surveys have shown a high incidence of PTB, reaching 315.50 per 100,000 people, which is 5.21 times the national average [3].

PTB is characterized by a long course and long treatment duration, often lasting approximately six months or longer. This imposes a significant economic burden on patients and their families, impacting employment opportunities and worsening financial hardships. According to WHO data, tuberculosis is the seventh leading cause of death in low- and middle-income countries and ranks seventh in terms of the overall economic burden of disease worldwide [4]. Among more than thirty countries with a high burden of tuberculosis worldwide, China ranks second in terms of new tuberculosis cases and cases of multidrug-resistant tuberculosis. While China has made progress in tuberculosis control through the implementation of the DOTS strategy, the country still faces challenges in tuberculosis prevention and control, with the incidence of PTB decreasing by approximately 61% from 2000 to 2010. However, this reduction has not significantly changed the situation in high-burden tuberculosis countries, highlighting the ongoing severity of the disease. The healthcare costs associated with tuberculosis in China have been continuously increasing. From 2006 to 2018, the total healthcare costs and average cost per patient for tuberculosis increased, with a budget of 7.19 million US dollars allocated for tuberculosis healthcare costs in 2019. Despite China's implementation of a free service policy for tuberculosis, patients still need to allocate a significant portion of their annual income to treatment costs. This has earned tuberculosis, the reputation of being a "disease of poverty," pushing many families into or back into financial hardship.

With the development of machine learning algorithms, predictive models have been widely applied in fields such as education, finance, and healthcare. Recently, there has been a growing interest in predictive models targeting high-cost patients to reduce healthcare expenditures compared to interventions targeting the entire population [5]. Existing "medical cost prediction models" can be broadly classified into two categories [6]: rule-based prediction models and supervised machine learning models. Rule-based models utilize traditional algorithms based on predefined cost rules. However, these models require substantial domain knowledge and may struggle to adapt to complex data in practical scenarios. On the other hand, supervised machine learning models, such as random forests and support vector machines, can capture relationships between data. However, these models require extensive feature engineering for different characteristics and may not adapt well to high-dimensional data. Despite these drawbacks, both types of models have shown good prediction performance in various domains, including house price prediction and the prediction of diagnosis and treatment costs for chronic diseases such as hepatitis B and coronary heart disease. For example, Taloba et al. [7] utilized linear regression analysis, the naive bayes classifier, and random forest models to predict medical costs related to spinal fusion, highlighting the good prediction performance of linear regression models based on the MAPE and R^2. Similarly, Gowd et al. [8] employed logistic regression, K nearest neighbor, random forest, naive Bayes, decision tree, gradient elevation tree, and other models to predict total medical costs after total shoulder arthroplasty, revealing that the random forest gradient elevation tree performed the best by comparing accuracy and the area under the receiver operating characteristic curve.

However, the abovementioned models are often considered incapable of adequately detecting complex patterns in large-scale population data [9]. Fortunately, rapidly developing deep learning techniques, including the latest deep neural network structures and quantitative methods, have shown promise in overcoming these challenges. Recent studies have demonstrated the success of deep learning in various medical applications, such as helping dermatologists examine skin cancer [10], predicting patient outcomes using medical text data [11], and as a clinical diagnostic to streamline the triaging of patients and facilitate the clinical decision-making process [12]. The early perceptron model was limited to simple binary classification problems and failed to solve complex nonlinear problems. However, the introduction of hidden layers and activation functions allowed the multilayer perceptron (MLP), also known as multilayer neural network, to be widely used in data mining, pattern recognition, machine learning, and other fields [13–15]. Introducing skip-layer connections in feedforward neural networks allows for the full utilization of input information to supplement the parts of the original feature information lost

during the network training compression process. The predictive performance has been validated and proven effective, MA Morid et al. [16] attempted to improve the performance of healthcare cost prediction methods by leveraging the feature learning power of convolutional neural networks for temporal pattern detection. P Drewe-Boss et al. [17] have shown that neural networks compare favorably to several baseline methods and that tools such as integrated gradients can be used to explain predictions of population health costs. A Al Bataineh et al. [18] proposed an MLP neural network trained with PSO for heart disease detection. The findings demonstrated that the MLP-PSO model can assist healthcare providers in more accurately diagnosing patients and recommending better treatments. Therefore, given the limited complexity of multiple regression and random forest regression (RFR), we selected the MLP model in our manuscript to validate the results and compare the three models.

In China, there have been few studies on healthcare cost prediction due to the public welfare and nonprofit nature of healthcare services. Therefore, the innovative analysis of the introduction of the MLP model provides insights into the future application of artificial intelligence in the medical field. By utilizing the MLP model to predict hospitalization costs, we can potentially provide more intelligent economic management tools for medical institutions and offer more personalized and economically reasonable medical services for patients. This has significant social and economic implications for promoting the coordinated development of the "artificial intelligence + medicine" field and improving the quality and efficiency of medical services.

Method

Data

This study aimed to investigate patients with PTB (coded as A15-A16 according to the ICD-10) who were admitted to Kashgar Pulmonary Hospital between 2020 and 2022 and subsequently discharged. The data are derived from the hospital information system of a pulmonary hospital in Kashgar. Medical professionals record and organize patient medical records within 24 h and promptly upload them to the hospital information system. Additionally, data within the system are independently collected and organized by two individuals, followed by cross-checking to ensure the accuracy and completeness of the date within the system. Patients with missing data and inability to verify cost were excluded from the study, a total of 9,570 eligible patients with PTB were included as subjects for this study (Fig. 1).

Total hospitalization cost

The costs were categorized according to the cost items listed on the first page of the latest version of the medical records in 2017. Similar cost items were combined, resulting in six categories, namely, diagnosis cost (including pathological diagnosis, laboratory diagnosis, imaging diagnosis, clinical diagnosis items, etc.), medical service cost (including general medical services, general treatment operations, nursing cost, etc.), material cost (including examination, treatment, surgical disposable medical materials, etc.), treatment cost (including non-surgical treatment items, surgical treatment cost, etc.), drug cost (including western medicine, Chinese patent medicine, Chinese herbal medicine, etc.), and other cost

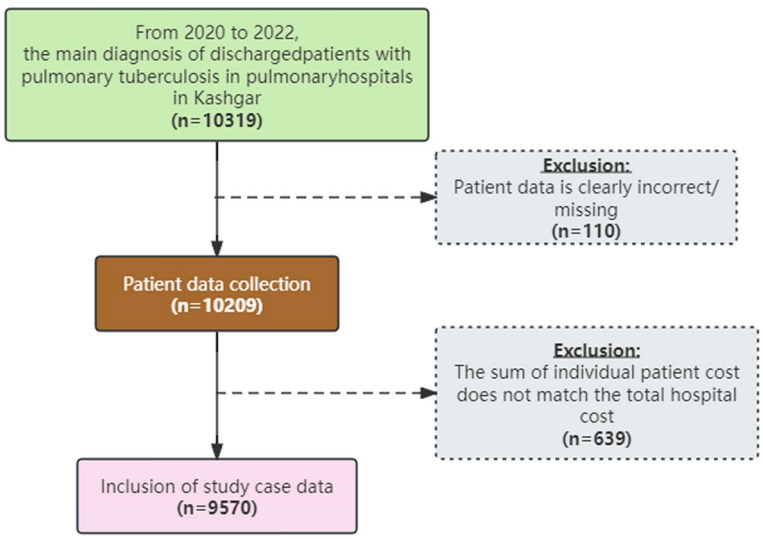


Fig. 1 This figure shows the patient inclusion and exclusion process

(including rehabilitation cost, traditional Chinese medicine cost, blood and blood product cost, etc.).

Sociological data

Sociological data such as gender, age, marital status, and payment method were collected. Disease Characteristics Data: Information related to the admission condition, length of hospital stay, initial treatment, presence of other diseases, transfer, allergy, drug resistance, and admission department were also collected.

It is important to note that this study obtained approval from the hospital ethics committee, and all patient information used in the research was completely anonymous.

Statistical methods

A comprehensive database was created using SPSS 26.0, and the data were collated and cleaned. Descriptive analysis was conducted on the hospitalization cost data, sociological data, and disease characteristic data of all patients with PTB. Categorical data was described using frequencies and constituent ratios, while medians and interquartile ranges were used to describe the central tendency and dispersion tendency of continuous data with a skewed distribution.

In univariate analysis, Mann–Whitney U test was used for pairwise comparisons, including 6 factors such as gender, initial treatment, presence of other diseases, transfer, allergy, and drug resistance; Kruskal–Wallis H test was used for multiple-group comparisons, including 6 factors such as age, marital status, payment method, admission condition, length of hospital stay, and admission department. After the results of the univariate analysis, multiple stepwise regression analysis was performed to analyze the main factors influencing the total hospitalization cost. $P < 0.05$ indicated a statistically significant difference.

Based on the collected data, multiple regression, RFR and MLP prediction models were established to predict the diagnosis cost, medical service cost, material cost, treatment cost, drug cost, other cost, and total hospitalization cost of patients with PTB. The prediction efficacy of the three models was compared using R-square (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). R^2 reflects the proportion of variation explained by the model and is usually used to evaluate the predictive ability of the model. A higher R^2 value indicates a better fit of the model to the data. The RMSE is considered to measure the average error that a model produces when predicting an observation. The lower the RMSE is, the better the predictive ability of the model. Like the RMSE, the MAE measures prediction errors but is less sensitive to outliers; the lower the MAE is, the better the model. An essential use of predictive models is identifying key data features and their impact on predictions. We leveraged Permutation Importance to assess the best model's feature importance.

Predictive model build

All the data were divided into a training set (2020–2021) and a test set (2022) according to the year. In terms of feature processing, multicategory variables were transformed into dummy variables; for binary variables, they were converted into 0 and 1; and the predictive costs were transformed using the logarithm. The numerical variables such as age and length of stay, were divided into ordinal categories by interval. The variable assignments were detailed in Appendix Table.

Seven multiple regression models were separately established for diagnosis cost, medical service cost, material cost, treatment cost, drug cost, other cost and total hospitalization cost. The forward method was employed to select the optimal variables.

With sklearn package in Python 3.7, seven RFR models also were separately established. In order to select the optimal hyper-parameters, a fivefold cross-validation method with grid search was used to tune below parameters, n_estimators (5 ~ 100), max_features (['auto', 'sqrt', 'log2']), max_depth (2 ~ 12), min_samples_split (5 ~ 150), and min_samples_leaf (5 ~ 50). By systematically exploring the parameter space through this comprehensive tuning strategy, we aimed to identify the optimal hyperparameter configuration for each RFR model, ensuring optimal performance and generalization ability.

The model architecture adopted in this study was different from the traditional multilayer perceptron. MLP encompasses an input layer, one or multiple hidden layers, and an output layer. Neurons in a layer are fully connected to those in the subsequent layer, enabling information transmission through weighted connections and biases. In our model's design,the original input was connected to the hidden vector in the last hidden layer and fed to the output layer. This setup permitted the network to harness both simple input–output relationships and residuals from complex deep learning architectures. Then, the final seven target costs including diagnosis cost, medical service cost, material cost, treatment cost, drug cost, other cost, and total hospitalization cost were predicted. The model was built and optimized by the Keras package in Python 3.7. To select the optimal hyperparameters, a fivefold cross-validation method with grid search was used to tune the number of hidden layers (3 ~ 10), the number of neurons per layer (10 ~ 50), the dropout (0.25, 0.5) and the learning rate (0.001, 0.01). Besides, Rectified Linear Units (ReLU) was applied as the activation function, the Adam optimizer for parameter optimization.

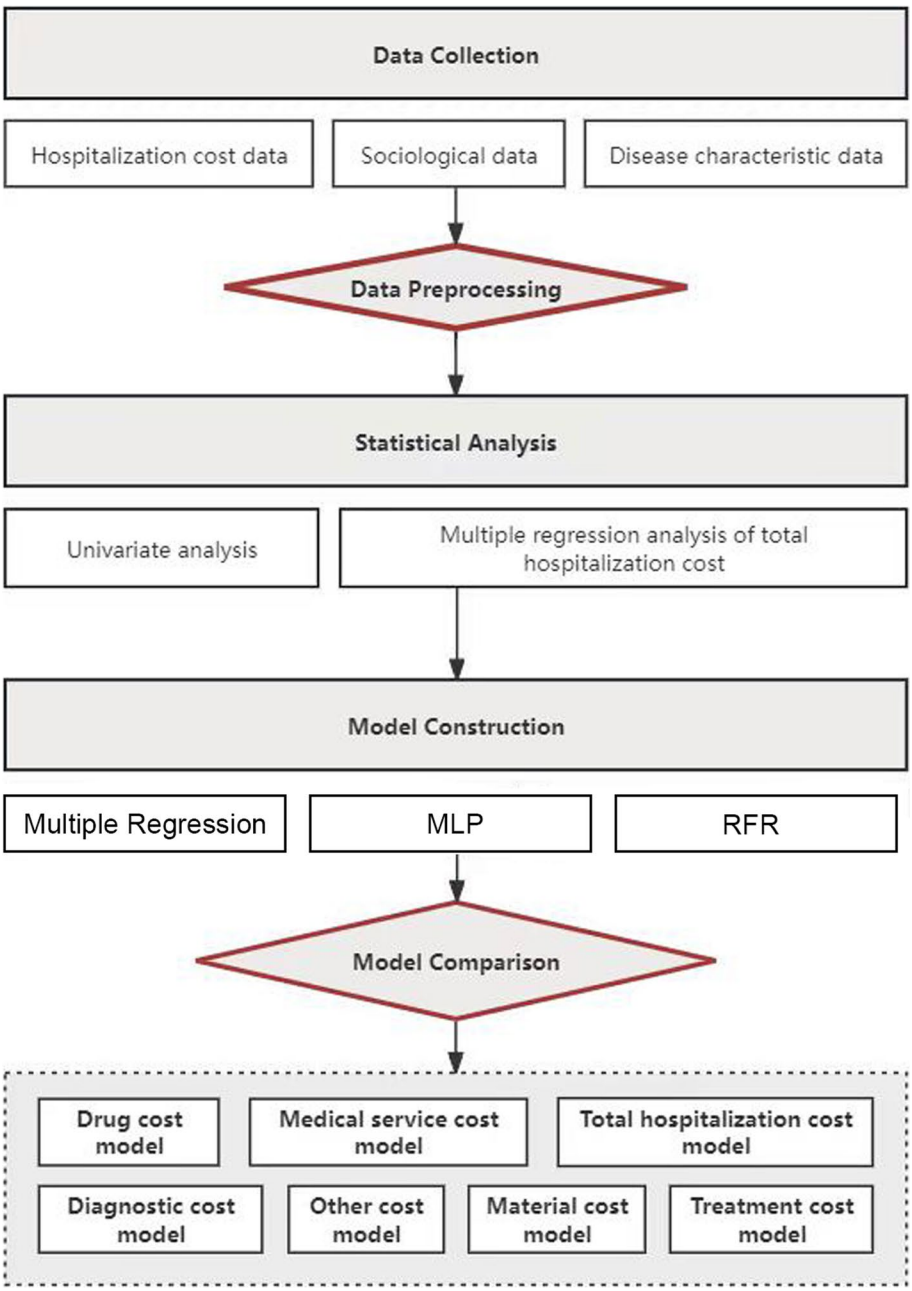


Fig. 2 This is an overview of Inference and Analysis Framework

In the fivefold cross-validation, 300 epochs were used for each training, and to prevent overfitting, the early stopping mechanism was used, indicating that training would be halted if the loss in the validation set did not decrease for 10 consecutive epochs. We used the Adam optimizer and the mean squared error as the loss function. The optimal model was evaluated with an average of 5 loss results. An overview of inference and analysis framework was illustrated in Fig. 2.

Results

Description of general feature

Among the 9,570 patients with PTB, the proportion of female patients (53.77%) was higher. The median and quartile of patient age were 67.00 (55.00, 74.00) years. The majority of patients were married (91.41%). Most patients (81.04%) used urban and rural residents' medical insurance as the payment method. The majority of patients were admitted as general patients (53.11%). The

median and quartile length of hospital stay for patients were 14.00 (11.00, 21.00) days, and 79.51% of patients were newly diagnosed with PTB. A total of 61.78% of patients had comorbidities. A total of 95.17% of patients were not transferred to another department. A total of 98.51% of patients had no allergies. A total of 97.77% of patients did not have resistance to medications. The proportion of PTB patients admitted to the respiratory department was the largest (17.46%).

Univariate analysis revealed significant differences in the total hospitalization cost among the 11 factors except allergy ($P=0.358$), including gender, age, marital status, payment method, admission condition, length of hospital stay, initial treatment, presence of other diseases, transfer, drug resistance, and admission department ($P<0.05$) (Table 1).

The hospitalization costs of various categories of PTB patients exhibited a skewed distribution, the median total hospitalization cost was 13,150.45 yuan, with an interquartile range between 9,891.34 yuan and 19,648.48 yuan. According to the analysis of hospitalization costs, material costs accounted for the greatest proportion of patients with PTB, followed by drug costs (Table 2).

Multiple regression analysis revealed that the F value was 1377.916, indicating a highly significant relationship ($P<0.001$). The R^2 was 0.760, suggesting that the regression model explained 76.0% of the total variation. The RMSE was 0.130, indicating a relatively small average difference between the predicted and actual values. Additionally, the coefficient of variance expansion (VIF) for all independent variables was less than 5, indicating the absence of multicollinearity among the variables included in the analysis.

The analysis revealed nine statistically significant factors (age, marital status, admission condition, length of hospital stay, initial treatment, presence of other diseases, transfer, drug resistance and admission department) that were the main factors influencing the hospitalization costs of patients with PTB ($P<0.05$) (Table 3).

Modeling Results

The seven multiple regression models were developed using the forward method, with varying characteristics included for cost prediction. Age and length of hospital stay were used in all models; Admission condition, initial treatment, transfer and admission department were effective variables for most models; Less used in the model were marital status, presence of other diseases and drug resistance (Table 4).

As to RFR, the best parameters were different in 7 cost prediction models. In total hospitalization cost prediction model, the optimal parameter were 'max_depth' = 8, 'max_features' = 'auto', 'min_samples_leaf' = 10, 'min_samples_split' = 10, 'n_estimators' = 10.

Figure 3 showed the relationship between the number of trees and RMSE when using the RFR algorithm to predict total hospitalization cost, with the RMSE being minimal near 10 trees. While figure (b) showed that the optimal max_depth was 8.

Based on the average loss results from fivefold cross-validation, the optimal parameter combination for the MLP model was determined(number of hidden layers = 3, number of neurons per layer = 50, dropout rate = 0.25, and learning rate = 0.001). With the optimal parameter combination, the validation set loss was minimized at the 18th epoch and did not decrease further after 10 epochs (Fig. 4). Figure 5 shows the true and predicted values of the MLP in the training and test sets, the scatter plot highlights the great predictive performance of the MLP. Figure 6 showed the ranking of feature importance for the model, with admission department and length of hospital stay ranking highly.

Table 5 showed the evaluation results of the three modeling methods. Overall, MLP demonstrated the superior performance in most cost predictions, outperforming RFR and multiple regression. For instance, as to the prediction of total hospitalization cost, both in the training set and the test set, MLP achieved the highest R^2 values (0.817 and 0.832, respectively), surpassing RFR's R^2 (0.758 and 0.809, respectively) and multiple regression's R^2 (0.707 and 0.777, respectively); Additionally, MLP had the lowest RMSE (0.103 and 0.114, respectively) and MAE (0.078 and 0.086, respectively), lower than the RMSE (0.118 and 0.122, respectively) and MAE (0.086 and 0.091, respectively) of RFR and the RMSE (0.130 and 0.132, respectively) and MAE (0.098 and 0.099, respectively) of multiple regression. However, in the prediction of diagnostic cost, RFR exhibited the best performance in the test set, with an R^2 of 0.629, RMSE of 0.231, and MAE of 0.167, better than MLP (0.609, 0.237 and 0.163, respectively) and multiple regression (0.518, 0.264 and 0.175, respectively). Despite multiple regression showed low performance across all models, in the prediction of other cost, multiple regression revealed the best results in the test set, with an R^2 of 0.352, RMSE of 0.792, and MAE of 0.627, better than MLP (0.313, 0.816 and 0.611, respectively) and RFR (0.342, 0.799 and 0.619, respectively).

Discuss

The combined results of univariate analysis and multivariate linear regression analysis revealed that the main factors affecting the hospitalization costs of PTB patients were, age, marital status, admission condition, length of hospital stay, initial treatment, presence of other diseases, transfer, drug resistance and admission department.

Table 1 Basic information of patients with PTB

Variable	Inpatient		Hospitalization cost (Yuan)		
	Number of subjects (person-time)	Composition ratio (%)	*M* (P_{25},P_{75})	*Z*/*H* value	*P* value
Gender				-2.213	0.027
Male	4425	46.24	13012.71 (9668.82,19609.62)		
Female	5145	53.76	13284.21 (10066.49,19733.87)		
Age(years)				340.37	<0.001
≤15	50	0.52	10194.28 (7125.30,15014.94)		
16-30	577	6.03	9981.99 (6803.84,15703.38)		
31-45	782	8.17	10975.33 (7891.00,16115.45)		
46-60	1758	18.37	12548.56 (9528.83,17919.80)		
61-75	4586	47.92	13990.43 (10627.84,21314.23)		
≥76	1817	18.99	13463.17 (10411.41,19538.85)		
Marital status				107.152	<0.001
Unmarried	498	5.20	10306.86 (6776.75,15785.45)		
Married	8748	91.41	13261.20 (10037.95,19735.77)		
Divorce	228	2.38	14083.54 (10633.55,21107.92)		
Widowed	96	1.00	14453.72 (10523.80,23956.15)		
Payment method				151.295	<0.001
Corps Medical Insurance	27	0.28	14506.26 (9521.99,22024.80)		
Urbanand rural residents	7756	81.04	13448.13 (10151.61,20408.62)		
Urban Workers	1219	12.74	12260.32 (9205.85,17218.68)		
Off-site health insurance	348	3.64	12127.39 (8830.54,16792.04)		
Self-pay	220	2.30	9676.71 (6819.64,13290.25)		
Admission condition				313.071	<0.001
General	5083	53.11	12737.53 (9727.38,18372.96)		
Emergency	2448	25.58	13894.13 (10376.44,21635.34)		
Sick	1182	12.35	16056.53 (11402.74,26732.78)		
Critical	857	8.96	11227.71 (8288.83,14816.97)		
Length of hospital stay (days)				5496.206	<0.001
≤10	2316	24.20	8534.42 (6581.87,10867.26)		
11-20	4787	50.02	12737.45 (10620.33,15273.81)		
21-30	1283	13.41	21806.43 (17403.86,26554.09)		
31-40	504	5.27	30719.36 (25863.00,36136.04)		
≥41	680	7.11	48813.73 (37608.93,65533.00)		
Initial treatment				-60.977	<0.001
Yes	7609	79.51	11851.97 (9173.40,14936.82)		
No	1961	20.49	30531.42 (23572.72,43639.10)		
Presence of other diseases				-12.435	<0.001
Yes	5912	61.78	12447.76 (9492.16,18281.68)		
No	3658	38.22	14503.09 (10751.23,21035.02)		
Transfer				-11.472	<0.001
Yes	462	4.83	17735.98 (13010.11,28358.99)		
No	9108	95.17	12978.66 (9778.39,19239.64)		
Allergy				-0.920	0.358
Yes	143	1.49	13647.12 (9661.46,22477.21)		
No	9427	98.51	13146.25 (9892.78,19618.11)		
Drug resistance				-17.598	<0.001
Yes	213	2.23	47359.01 (24594.49,67134.57)		
No	9357	97.77	13023.42 (9822.66,19081.58)		
Admission department				1221.911	<0.001

Table 1 (continued)

Variable	Inpatient		Hospitalization cost (Yuan)		
	Number of subjects (person-time)	Composition ratio (%)	*M* (P_{25},P_{75})	*Z*/*H* value	*P* value
ICU	442	4.62	22044.65 (14815.62,35472.80)		
Respiratory	1671	17.46	10405.34 (7723.47,13916.47)		
Tuberculosis I	1332	13.92	16126.22 (12032.04,22693.79)		
Tuberculosis II	825	8.62	18176.32 (11875.56,34979.83)		
Outpatient emergency department	589	6.15	9889.17 (7853.47,12614.11)		
Internal Division I	1538	16.07	12975.58 (10015.79,18010.73)		
Internal Division II	1639	17.13	13000.31 (10434.01,17895.37)		
Cardiology	1534	16.03	13301.91 (10587.58,18674.75)		

Table 2 Composition of hospitalization costs for patients with PTB

Cost Category	Total cost (ten thousand RMB)	Median cost [yuan, *M* (Q_1, Q_3)]	Composition ratio (%)
Diagnostic cost	3261.81	2550.30 (1647.15, 4300.03)	19.34
Medical service cost	1291.87	800.00 (562.00, 1397.00)	7.66
Material cost	5101.30	4209.74 (3444.53, 5728.59)	30.25
Treatment cost	3057.79	2405.50 (1542.00, 3742.25)	18.13
Drug cost	4047.92	2885.02 (1665.10, 4880.01)	24.01
Other cost	101.59	7.00 (0.00, 56.00)	0.60
Total hospitalization cost	16,862.29	13,150.45 (9891.34, 19,648.48)	100.00

Table 3 Multiple linear regression analysis results of influencing factors of hospitalization costs in patients with PTB

Influencing factors	Regression Coefficient β	Standard Error	Standard regression coefficient	*t* value	*P* value	VIF
Constant	3.816	0.027		141.634	< 0.001	
AgeX_2	0.026	0.001	0.107	18.034	< 0.001	1.411
Marital status X_3(Ref: Unmarried)						
Married	0.063	0.007	0.067	9.158	< 0.001	2.127
Divorce	0.088	0.011	0.051	8.037	< 0.001	1.590
Widowed	0.050	0.015	0.019	3.231	0.001	1.332
Admission condition X_5	-0.005	0.001	-0.018	-3.419	0.001	1.046
Length of stay X_6	0.160	0.002	0.657	86.069	< 0.001	2.320
Initial treatment X_7	0.131	0.005	0.200	27.741	< 0.001	2.062
Presence of other diseases X_8	0.010	0.003	0.019	2.971	0.003	1.584
Transfer X_9	0.036	0.006	0.029	5.598	< 0.001	1.099
Drug resistance X_{10}	0.047	0.010	0.026	4.782	< 0.001	1.178
Admission department X_{11} (Ref: ICU)						
Respiratory	-0.331	0.007	-0.474	-46.665	< 0.001	4.115
Tuberculosis I	-0.264	0.008	-0.345	-33.847	< 0.001	4.150
Tuberculosis II	-0.225	0.008	-0.239	-26.814	< 0.001	3.166
Outpatient emergency department	-0.323	0.008	-0.293	-38.857	< 0.001	2.266
Internal Division I	-0.289	0.007	-0.401	-39.845	< 0.001	4.029
Internal Division II	-0.272	0.007	-0.386	-37.695	< 0.001	4.190
Cardiology	-0.250	0.007	-0.346	-34.878	< 0.001	3.392

Table 4 Feature selection in multiple regression models for seven hospitalization costs

Cost Category	Age	Marital status	Admission condition	Length of hospital stay	Initial treatment	Presence of other diseases	Transfer	Drug resistance	Admission department
Diagnostic cost	√		√	√	√	√	√		
Medical service cost	√		√	√	√				√
Material cost	√		√	√	√		√		√
Treatment cost	√	√		√		√	√	√	√
Drug cost	√			√	√	√	√		√
Other cost	√	√	√	√	√		√		√
Total hospitalization cost	√		√	√	√		√	√	√

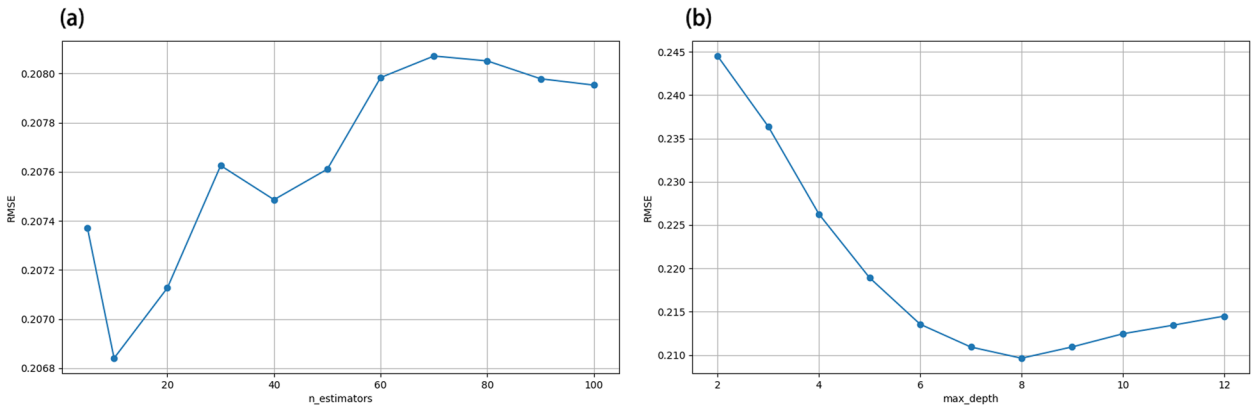


Fig. 3 This diagram shows training process of the RFR

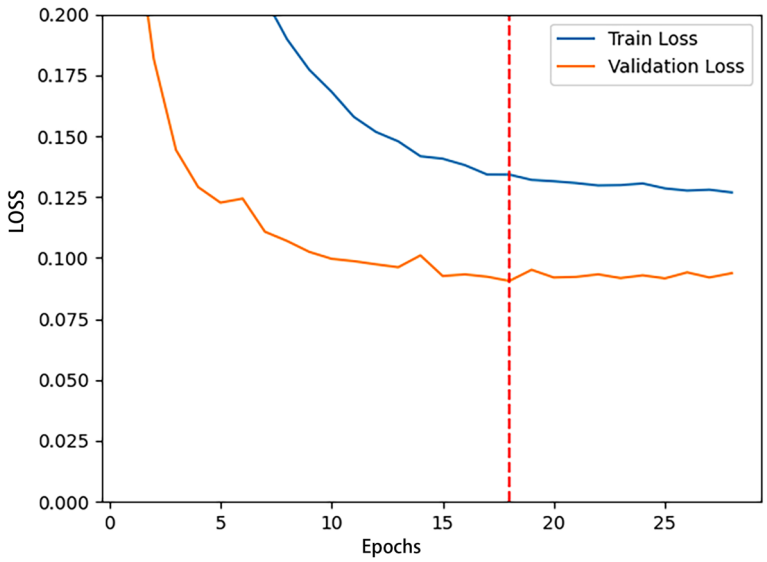
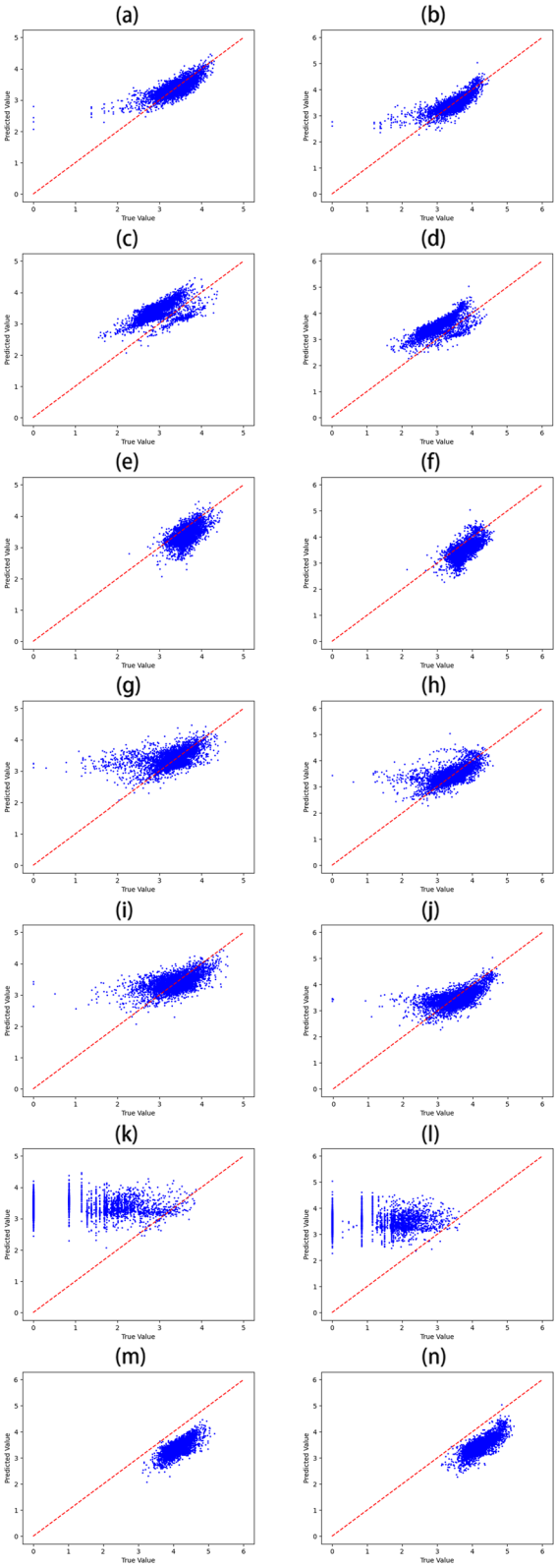


Fig. 4 This graph shows the loss curve of train and validation set of the MLP under the optimal parameter combination

Length of hospital stay, initial treatment and age were important factors affecting the hospitalization costs of PTB patients, with standard regression coefficients of 0.657, 0.200 and 0.107, respectively. Notably, younger patients tend to demonstrate better physical fitness and fewer underlying health conditions, which contributes to



◀ **Fig. 5** This picture shows the scatter plot of the true and predicted values of seven hospitalization costs in MLP model. Shown in (**a**, **c**, **e**, **g**, **i**, **k**, **m**) are the training while shown in (**b**, **d**, **f**, **h**, **j**, **l**, **n**) are testing sets

more favorable treatment outcomes and, consequently, lower hospitalization costs. With increasing age, the patient's own immunity will decrease, which will have a certain impact on the therapeutic effect, resulting in increased hospitalization costs; widowed patients need to bear the economic productivity of the entire family alone, and various external pressures may also cause psychological and physical double pressures, so the patient's physical health will be adversely affected, indirectly leading to an increase in the total hospitalization cost; critically ill patients face a large number of treatment costs and examination costs, which may require shortening the length of hospital stay as much as possible during the diagnosis and treatment period, strengthening the intensity of treatment, and avoiding unnecessary treatment items, examination items and drugs as much as possible, leading to lower per capita hospitalization costs; the longer the length of hospital stay is, the greater the number of medical and health resources, such as examination costs, medical costs and bed costs, and thus, it is bound to increase the total hospitalization costs [19–21]. Multiple treatment, combined with other diseases, transfer, drug resistance, and PTB patients in ICU departments may be more complex and correspondingly more difficult to treat, usually accompanied by more health resource consumption, which invisibly leads to an increase in the total hospitalization cost [22–24]. The above influencing factors suggest that relevant medical insurance policies should be improved, standardized diagnosis, treatment and management models should be improved, and popular science awareness of PTB prevention and control should be promoted to optimize the structure of hospitalization costs and reasonably control the increase in medical costs.

The results of our study showed that the MLP model had better predictive performance for hospitalization costs than did the multiple regression and RFR, which was similar to the findings of previous studies. [25–27]. Traditional models such as multiple regression and RFR might not be able to identify complex relationships among variables, leading to suboptimal fitting performance. However, the MLP model has the characteristics of large-scale parallel processing, high fault tolerance, self-organization, self-adaptive ability and an association function [28]. This approach necessitated a less restrictive specification of independent variables and could effectively solve highly complex classification problems.

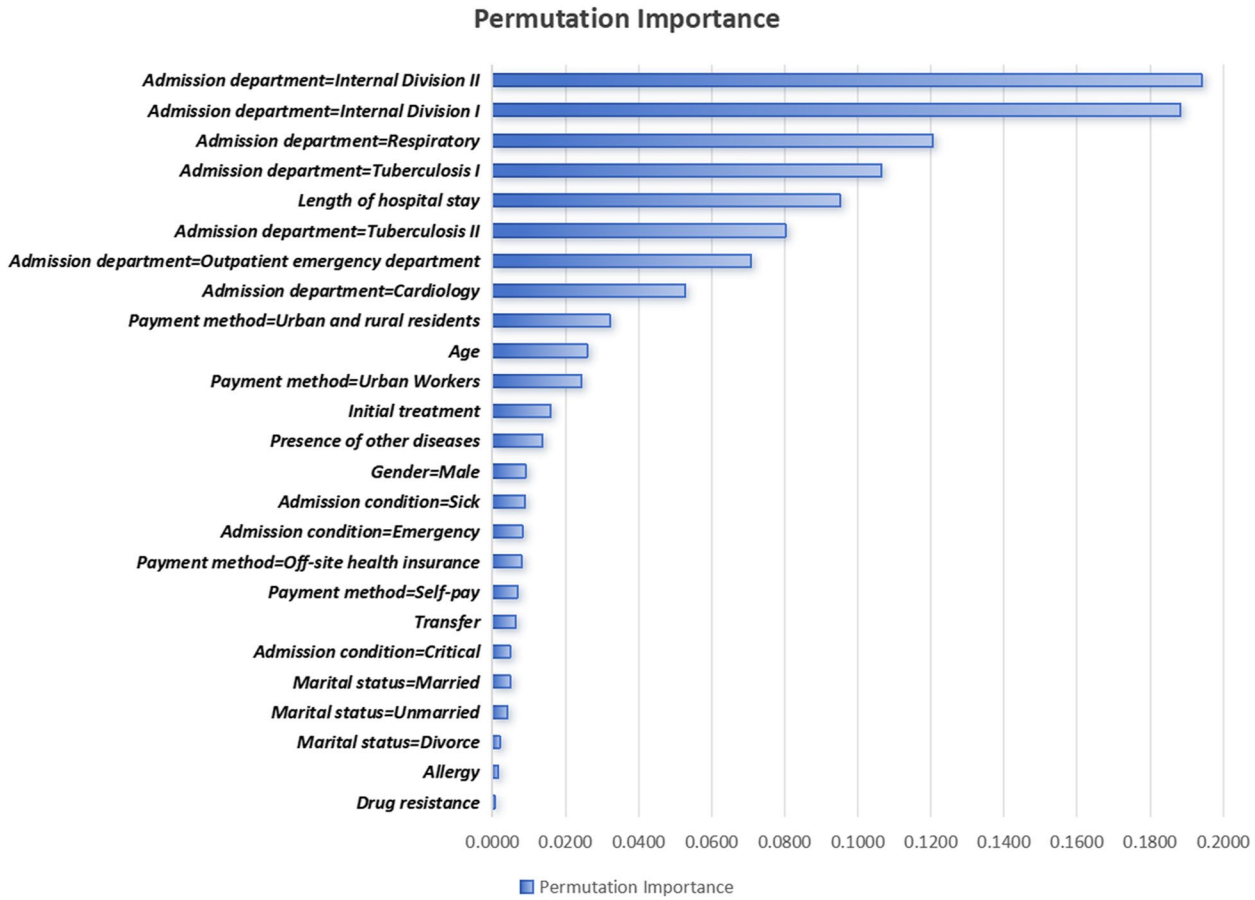


Fig. 6 This picture shows the ranking of feature importance for the model

It was suitable for processing information that involves considering numerous factors and conditions simultaneously, especially in situations that are imprecise and fuzzy [29]; high nonlinear global action, automatic extraction of "reasonable" solution rules, and certain promotion and generalization ability were also the advantages of the MLP model. Chen Y et al. [30] proposed a new LPR-MLP hybrid model, which uses LBP, PCA, and Relief F to process image data and meteorological mechanics data, respectively, and then uses MLP to predict its health level, thus solving the challenge of predicting the health status of transmission lines under high-dimension, multimode, nonlinear, and heterogeneous data. The experimental results have shown that the LPR-MLP model has high prediction accuracy and performance.

Recently, the success of deep learning has led to a resurgence of interest in MLP. MLP, which is regarded as a standard supervised learning algorithm in the field of pattern recognition and continues to become a subject of research in the field of computational neurology and parallel distributed processing, is often used as a back propagation algorithm for learning [31]. At present, the MLP has been proven to be a general functional approximation method that can be used to fit complex functions or solve classification problems. Compared with other prediction models, MLP models have shown unique advantages, and their popularity and application are wider [32]. The utilization of the MLP model for predicting hospitalization costs positively impacts medical cost management, aiding in the rational allocation of medical resources by diverse healthcare institutions and balancing various hospitalization costs through adjustments to high-cost items. This ultimately achieves the goal of controlling patients' medical costs and alleviating their economic burden.

At present, there is limited research on predicting medical costs. The introduction of the MLP model for hospitalization cost prediction holds significant social and economic importance in enhancing the quality and efficiency of medical services. The MLP model is currently capable of effectively addressing more complex problems, but it also exhibits notable shortcomings, such as the challenging task of determining the appropriate number of hidden nodes in the network, potential inadequacies in learning, and lengthy training times. Consequently,

Table 5 The comparison results of multiple regression, RFR and MLP

Cost Name	Multiple regression						MLP						RandomForesRegression					
	2020–2021 (Training Set)			2022 (Test Set)			2020–2021 (Training Set)			2022 (Test Set)			2020–2021 (Training Set)			2022 (Test Set)		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Diagnostic cost	0.445	0.264	0.181	0.518	0.264	0.175	0.596	0.224	0.161	0.609	0.237	0.163	0.57	0.232	0.173	0.629	0.231	0.167
Medical service cost	0.716	0.176	0.121	0.753	0.181	0.125	0.826	0.138	0.093	0.816	0.156	0.107	0.686	0.185	0.115	0.687	0.203	0.135
Material cost	0.657	0.117	0.085	0.668	0.133	0.098	0.708	0.107	0.079	0.712	0.124	0.092	0.668	0.115	0.084	0.703	0.126	0.092
Treatment cost	0.527	0.302	0.201	0.564	0.27	0.187	0.698	0.241	0.158	0.629	0.249	0.162	0.578	0.285	0.178	0.618	0.253	0.161
Drug cost	0.463	0.3	0.218	0.54	0.296	0.22	0.585	0.263	0.197	0.581	0.282	0.212	0.494	0.29	0.211	0.54	0.296	0.216
Other cost	0.456	0.76	0.591	0.352	0.792	0.627	0.6	0.65	0.486	0.313	0.816	0.611	0.471	0.748	0.57	0.342	0.799	0.619
Total hospitalization cost	0.707	0.13	0.098	0.777	0.132	0.099	0.817	0.103	0.078	0.832	0.114	0.086	0.758	0.118	0.086	0.809	0.122	0.091

future research on MLP models should further expand the research sample and include more influencing factors and comparative models.

This study was constrained by the limitations of its data source, as patient data ware derived from a single center, and its generalizability may be restricted. Furthermore, as a retrospective study, the cases were sourced from the hospital information systems, limiting access to additional variables that may impact hospitalization costs. Therefore, a multicenter study with a greater number of variables is required.

Conclusion

In general, the MLP model has demonstrated significant advantages over the traditional multiple regression models in terms of prediction efficacy. It enabled the utilization of comprehensive patient information and effectively predicted hospitalization costs, thereby facilitating the rationalization of cost structures and reducing the economic burden on patients. Furthermore, the insights gained from the MLP model hold considerable reference value for research on other diseases, highlighting its broader applicability in the field of healthcare economics.

Abbreviations

PTB	Pulmonary Tuberculosis
WHO	World Health Organization
MLP	Multilayer Perceptron
R^2	R-square
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
RFR	Random Forest Regression
ReLU	Rectified Linear Units

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-024-09771-6.

Additional file 1.

Acknowledgements

We thank the patients and their family for their consent to participate in this study. We also appreciate all medical staff's cooperation and care of the patient.

Authors' contribution

(I) Conception and design: Shiyu Fan, Abudoukeyoumujiang Abulizi, Yasen Yimit. (II) Administrative support: Xiaoguang Zou, Mayidili Nijiati. (III) Provision of study materials or patients: Abudoukeyoumujiang Abulizi, Mayidili Nijiati. (IV) Collection and assembly of data: Shiyu Fan, Yasen Yimit, Qiange Li. (V) Data analysis and interpretation:.Yi You, Chencui Huang. (VI) Manuscript writing: All authors. (VII) Final approval of manuscript: All authors.

Funding

The project was funded from two sources. One is the Tianshan Innovation Team Program of Autonomous Region (Grant number 2022D14007). The other is the Health Kashgar National Regional Medical Center Talent Cultivation Demonstration Base (KSRC-2022001).

Availability of data and materials

The data underlying this manuscript cannot be shared publicly due to the privacy of individuals in the study. These data are stored in a protected information system at a pulmonary hospital in Kashgar, Xinjiang, China. The datasets manipulated or generated in our research are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

The 1964 Declaration of Helsinki and its later amendments or equivalent ethical standards were followed in all procedures carried out in studies involving human subjects. This retrospective study was approved by the Ethics Committee of The First People' s Hospital of Kashi (Kashgar) Prefecture 2023–60, which waived the need for written informed consent from the patients.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Present Address: Department of Preventive Healthcare, Shihezi University, Shihezi 832000, China. [2]Department of Radiology, The First People's Hospital of Kashi (Kashgar) Prefecture, Kashgar 844000, China. [3]Present Address: Xinjiang Key Laboratory of Artificial Intelligence Assisted Imaging Diagnosis, Kashgar 844000, China. [4]Present Address: Department of Research Collaboration, Hangzhou Deepwise & League of PHD Technology Co., Ltd, R&D Center, Hangzhou 311101, China. [5]Xinnjiang Health Commission, Urumqi 830000, China. [6]The Fourth Affiliated Hospital of Xinjiang Medical University, Urumqi 830000, China.

Received: 29 March 2024 Accepted: 20 August 2024
Published online: 28 August 2024

References

1. Chakaya J, Khan M, Ntoumi F, Aklillu E, Fatima R, Mwaba P, et al. Global Tuberculosis Report 2020–Reflections on the Global TB burden, treatment and prevention efforts. Int J Infect Dis. 2021;113:S7–12.
2. Philippe G, Katherine F, Mario R. Iconography: global epidemiology of tuberculosis. Seminars Respir Crit Care Med. 2018;39(03):271–85.
3. Chen W, Zhang H, Du X, Li T, Zhao Y. Characteristics and Morbidity of the Tuberculosis Epidemic-China, 2019. China CDC Weekly. 2020;2(12):181–4.
4. World Health Organization. World health statistics 2021: monitoring health for the SDGs, sustainable development goals. World Health Organ. 2021;2021:1–136.
5. Osawa I, Goto T, Yamamoto Y, Tsugawa Y. Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. NPJ Digit Med. 2020;3(1):148.
6. Desai RJ, Wang SV, Vaduganathan M. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA Netw Open. 2020;3(1):e1918962.
7. Taloba AI, El-Aziz A, Rasha M, El-Bagoury AAH. Estimation and prediction of hospitalization and medical care costs using regression in machine learning. J Healthc Eng. 2022;2022:7969220.
8. Gowd AK, Agarwalla A, Beck EC, Rosas S, Waterman BR, Romeo AA, et al. Prediction of total healthcare cost following total shoulder arthroplasty utilizing machine learning. J Shoulder Elbow Surg. 2022;31(12):2449–56.
9. Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, et al. Canadian association of radiolo-gists white paper on artificial intelligence in radiology. Can Assoc Radiol J. 2018;69(2):120–35.
10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Der-matologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115.

11. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Dig Med. 2018;1(1):18.
12. Zhou HY, Yu Y, Wang C, Zhang S, Gao Y, Pan J, et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nat Biomed Eng. 2023;7(6):743–55.
13. Desai M, Shah M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). Clin eHealth. 2021;4:1–11.
14. Mishra S, Tripathy HK, Mallick PK, Bhoi AK, Barsocchi P. EAGA-MLP—an enhanced and adaptive hybrid classification model for diabetes diagnosis. Sensors. 2020;20(14):4036.
15. Poongodi M, Malviya M, Kumar C, Hamdi M, Vijayakumar V, Nebhen J, et al. New York City taxi trip duration prediction using MLP and XGBoost. Int J Syst Assurance Eng Manage. 2022;1:1–12.
16. Morid MA, Sheng ORL, Kawamoto K, Abdelrahman S. Learning hidden patterns from patient multivariate time series data using convolutional neural networks: A case study of healthcare cost prediction. J Biomed Inform. 2020;111:103565.
17. Drewe-Boss P, Enders D, Walker J, Ohler U. Deep learning for prediction of population health costs. BMC Med Inform Decis Mak. 2022;22(1):32.
18. Al Bataineh A, Manacek S. MLP-PSO hybrid algorithm for heart disease prediction. J Personal Med. 2022;12(8):1208.
19. Moreira ASR, Kritski AL, Carvalho ACC. Social determinants of health and catastrophic costs associated with the diagnosis and treatment of tuberculosis. J Bras Pneumol. 2020;46:e20200015.
20. Samuel R, Natesan S, Bangera MK. Quality of life and associating factors in pulmonary tuberculosis patients. Indian Journal of Tuberculosis. 2023;70(2):214–21.
21. Assebe LF, Negussie EK, Jbaily A, Tolla MTT. Financial burden of HIV and TB among patients in Ethiopia: a cross-sectional survey. BMJ Open. 2020;10(6):e036892.
22. Li XZ, Jin F, Zhang JG, Deng YF, Shu W, Qin JM, et al. Treatment of coronavirus disease 2019 in Shandong, China: a cost and affordability analysis. Infect Dis Poverty. 2020;9(03):31–8.
23. Oga-Omenka C, Tseja-Akinrin A, Sen P. Factors influencing diagnosis and treatment initiation for multidrug-resistant/rifampicin-resistant tuberculosis in six sub-Saharan African countries: a mixed-methods systematic review. BMJ Global Health. 2020;5(7):e002280.
24. Wang Y, McNeil EB, Huang Z, Chen L, Lu X, Wang C. Household financial burden among multidrug-resistant tuberculosis patients in Guizhou province, China: a cross-sectional study. Medicine. 2020;99(28):e21023.
25. Gopukumar D, Ghoshal A, Zhao H. Predicting readmission charges billed by hospitals: machine learning approach. JMIR Med Inform. 2022;10(8):e37578.
26. Theerthagiri P. Forecasting hyponatremia in hospitalized patients using multilayer perceptron and multivariate linear regression techniques. Concurr Comput: Pract Exp. 2021;33(16):e6248.
27. Chen M, Wu X, Zhang J, Dong E. Prediction of total hospital expenses of patients undergoing breast cancer surgery in Shanghai, China by comparing three models. BMC Health Serv Res. 2021;21(1):1–9.
28. Miranda AC, Santana JCC, Yamamura CLK, Rosa JM, Tambourgi EB, Ho LL, et al. Application of neural network to simulate the behavior of hospitalizations and their costs under the effects of various polluting gases in the city of São Paulo. Air Qual Atmos Health. 2021;1:1–9.
29. Karnuta JM, Navarro SM, Haeberle HS, Helm JM, Kamath AF, Schaffer JL, et al. Predicting inpatient payments prior to lower extremity arthroplasty using deep learning: which model architecture is best? J Arthroplasty. 2019;34(10):2235–41.
30. Chen Y, Chen S, Zhang N, Liu H, Jing H, Min G. LPR-MLP: A novel health prediction model for transmission lines in grid sensor networks. Complexity. 2021;2021(1):8867190.
31. Car-Pusic D, Petruseva S, Zileska Pancovska V, Zafirovski Z. Neural network-based model for predicting preliminary construction cost as part of cost predicting system. Adv Civil Eng. 2020;2020:1–13.
32. Al-Taie RRK, Saleh BJ, Saedi AYF, Salman LA. Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: a case study in Iraq. Int J Electr Comput Eng. 2021;11(6):5229.

Publisher's Note