Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **11**, Number **1**, February **2026**

Accurate Recommendation of Hospital Online Diagnosis and Treatment Data Based on Big Data Logistic Regression Model

Tian-Feng Xu[1,*]

[1]Zhejiang Technical Institute of Economics, Hangzhou 310018, P. R. China
xutianfeng2024@163.com

Yi-Nuo Duan[2]

[2]Ulink College of Shanghai, Shanghai 201615, P. R. China
2334196088@qq.com

Lin Su[3,4]

[3]Minnan Science and Technology College, Quanzhou 362300, P. R. China
[4]School of Architecture and City, Royal Melbourne Institute of Technology, Melbourne VIC3000, Australia
3051700287@qq.com

[*]Corresponding author: Tian-Feng Xu

Received January 12, 2025, revised April 28, 2025, accepted July 27, 2025.

Abstract. *For the current hospital online diagnosis and treatment data precision recommendation model lack of big data training process, resulting in abnormal data recommendation results, this paper firstly based on the idea of non-convex penalty regression to improve the logistic regression model, to address the issue that the logistic regression model will crash. The mean-drift parameter vector is added to the logistic regression model to identify outliers, and the penalty function is used to generate sparsity, which is screened with its corresponding threshold law to obtain more robust estimates of the results. Then the neural network is applied to set the training model of electronic medical record big data and construct the key discriminant function of diagnosis and treatment information. On this basis, a modified logistic regression model was used to correlate different categories of consultation data, A matrix of latent similarity constraints was introduced to impact the user's eigenvectors, hyperparameter sampling was performed by Gibbs sampling method and Bayesian networks were used to calculate the maximum prior and posterior probabilities to obtain the user's ratings data on resources. Finally, through the PageRank approach to optimize the recommendation of medical treatment resources, to complete the accurate recommendation of online medical treatment data in hospitals. The experimental outcome indicates that the offered model exhibits a reduction of at least 0.1 in its Root Mean Square Error (RMSE) compared to other models, and more accurate recommendation outcome can be gained.*
Keywords: Big data; Logistic regression model; Online clinic data recommendation; Non-convex penalized regression; PageRank approach.

1. **Introduction.** Recent medical surveys have shown that the number of patients admitted to hospitals has increased geometrically, and the trend is increasing year by year and at a younger age [1]. The healing process of most diseases is a long-term and complex systematic project, which requires patients to improve their cooperation and actively participate in the treatment plan while following the doctor's advice. Due to the high cost and poor therapeutic effect of traditional medicine, it has gradually failed to meet the current needs of people's health. As a result of limited medical resources and untimely supply of

medicines, conflicts between doctors and patients have intensified, making it difficult to see a doctor and expensive to buy medicines a pain in the heart of the general public. With the increasing shortage of medical resources, the demand for intelligent online diagnosis and treatment is increasing [2, 3]. How to accurately recommend hospital online diagnosis and treatment data through big data technology is of great significance to the growth of pharmaceutical economy [4].

1.1. **Related work.** The research on accurate recommendation of hospital online diagnosis and treatment data belongs to the recommendation field. Shu et al. [5] proposed a Probability Matrix Factorization (PMF) algorithm, which complements the scoring matrix by filling in the missing values, and the recommendation based on the filled scoring matrix improves the recommendation accuracy to a certain extent. Liang et al. [6] introduced a constraint matrix in MF, which avoids the feature that users who have provided a limited number of ratings are near to the average of the predefined distribution and improves the accuracy of rating prediction. However, the overfitting problem may occur easily. Although some researchers have introduced Bayesian architecture on top of MF and used Monte Carlo sampling method to extract the hyper-parameters of the model, which can efficaciously address the model overfitting and improvement issues, there is still room for further optimization [7, 8].

Many scholars have studied recommender systems with the influence of time factor in mind, to improve overall recommended performance. Liu [9] used time decay function to measure the similarity of items, and clustered the items to compute the user's preference for the item category and the weight of the item within its own category, which improved the recommendation accuracy. Zhang et al. [10] added a time factor to add time weights to each user's predicted score, and concluded that the new algorithm is more accurate than the traditional collaborative filtering (CF), which effectively improves the traditional algorithm. To enhance the efficiency of recommendation, Przystupa et al. [11] class information is fused with item information so as to construct PMF model based on entity correlation, which shows strong advantages in rating prediction.

In recent years, deep learning (DL) models have brought new ideas for optimizing recommendation methods. Fang et al. [12] developed a multi-criteria collaborative filtering recommendation algorithm by combining with deep neural networks, and confirmed the effectiveness of the method on different datasets. Fu et al. [13] came up with recommendations derived from combining CF algorithms and DL models, optimizing the results for relevance, and then rearranging the optimized results, a method that allowed for increased diversity. Cai et al. [14] developed a genetic algorithm based recommender system. The system considers the similarity of items and rating satisfaction to assess the suitability of individual species, and the model has better performance in the calculation of similarity, but suffers from data sparsity.

Although DL application techniques are able to show better performance in learning about users and item representations than traditional personalized recommendation methods, such methods are still unable to change a number of common problems when making recommendations, such as data sparsity and cold starts. Logistic regression models are simple in structure, computationally inexpensive, and fast, and can be a good solution to the problems inherent in DL models. Lian et al. [15] used a social matrix decomposition recommendation algorithm with Logistic regression model to significantly alleviate the data sparsity problem. Zhang et al. [16] combined neural network and Logistic regression model to predict the final ratings of the items, and achieved better recommendation results.

1.2. **Contribution.** As can be seen from the analysis of the above research status, the current hospital online diagnosis and treatment data recommendation model has the problem of data sparsity, which leads to poor recommendation accuracy. For this reason, this paper proposes an accurate recommendation model for hospital online diagnosis and treatment data based on big data logistic regression model. Firstly, to address the issue that the logistic regression model breaks down when there are serious outliers in the sample data, this paper adds a sparse, case-specific mean-drift parameter vector to the logistic regression model to identify the outliers, and generates sparsity with a penalty function, and sifts the outliers with its corresponding threshold law. Subsequently, different information criteria are used to determine the tuning parameters and the choice of optimal parameter solutions for robust estimation of the results. Then the neural network is applied to set the training model of electronic medical record big data, and the Euclidean distance formula is used to construct the key discriminant function of diagnosis and treatment information. On this basis, a matrix incorporating latent similarity constraints was implemented to influence the user's eigenvectors by using an improved logistic regression model to correlate the different categories of consultation data, and the improved logistic regression model was adopted to represent the potential factors of the non-linear relationship. Through the Gibbs sampling method for hyper-parameter sampling and the use of Bayesian network to calculate the maximum a priori and a posteriori probability, to gain the user's evaluation data on the resources, and eventually select the

PageRank method to optimize the treatment resources with the top N ratings, to complete the accurate recommendation of the hospital's online treatment data. The experimental outcome indicates that the proposed model has high recommendation accuracy and can improve the recommendation quality to a certain extent.

2. **Theoretical analysis.**

2.1. **Logistic regression model.** Logistic regression model is a universal linear model which, despite the name 'regression', is mainly used for classification problems, especially binary classification problems [17]. It obtains the probability of an event occurring by constructing a linear model and mapping the output of the linear model between 0 and 1 using a Logistic function (usually a Sigmoid function) [18]. Logistic regression models are simple to implement, easy to understand and implement compared to DL models, and can better solve the data sparsity problem in DL.

When the model needs to establish a relationship among the dependent variable $p(X)$ and the explanatory variable X, it is generally assumed initially that they are linearly related, i.e., $P = X'\beta$. But when $p(X)$ expresses the probability of an event, as in $p(X) = \Pr(Y = 1|X)$, then $p(X)$ can only take values in the $[0, 1]$ interval, and in the neighborhood of $p(X) = 0$ and $p(X) = 1$, P is insensitive to changes in X. Faced with this situation, we can construct a connection function $\theta(p)$ so that the functional relationship reflects this property while ensuring that the fitted value of $p(X)$ is always meaningful, resulting in the following construction, where $\theta(p)$ is shown in Equation (2).

$$\frac{\partial\theta(p)}{\partial p} = \frac{1}{p} + \frac{1}{1-p} \tag{1}$$

$$\theta(p) = \ln\left(\frac{p}{1-p}\right) \tag{2}$$

Assuming a linear relationship between $\theta(p)$ and X, then having Equation (3).

$$\ln\frac{p}{1-p} = X'\beta \tag{3}$$

A common form of the dichotomous logistic model can then be obtained, as shown in Equation (4).

$$p(X) = \frac{e^{X'\beta}}{1+e^{X'\beta}} \tag{4}$$

Once the model is constructed, the parameters of the model can be estimated using the great likelihood method.

2.2. **Collaborative filtering and matrix factorization.** The CF algorithm is called user-based CF (CF) [19], which recommends items to the target user by discovering users with similar ratings to the target user and recommending items to the target user with their preferences. The item-based CF (Item Collaborative Filtering Recommendation Algorithm) calculates the items that are similar to the target user's previous preferences from the degree of similarity of the item scores and recommends them, as shown in Figure 1. Both of them belong to the category of memory-based CF [20]. The advantages of these algorithms are high accuracy and efficient execution, but the cost of training the model is high.

MF is a mathematical approach to CF, which essentially decomposes the user-item rating matrix into the product of two low-dimensional matrices. In this way, implicit feature vectors of users and items can be learnt for personalized recommendation. The representative models of MF are PMF [21], Bayesian probabilistic matrix factorization (BMF) [22], non-negative matrix factorization (NMF) [23], etc. MF is to find the space of coexistence of hidden factors of users and items by decomposing the user-item evaluation matrix. Assuming that the user-item evaluation matrix is $R_{m\times n}$, including M users and N items, two k-dimensional low-rank matrices $U_{m\times k}$ and $V_{k\times n}$ consisting of user and item latent factors are obtained by MF, such that $R \approx UV$. After fixing $U_{m\times k}$ and $V_{k\times n}$, the vector inner product can be used to predict the users' ratings of the items as shown below.

$$\arg\min_{U,V} L(R, UV^T) + \lambda(\|U\|_F^2 + \|V\|_F^2) \tag{5}$$

where L is a loss function to evaluate the gap between the two matrices before and after the decomposition, and λ is a regularization parameter, which is a regularized representation of the decomposed matrix to prevent training overfitting. The PMF model is shown in Figure 2.

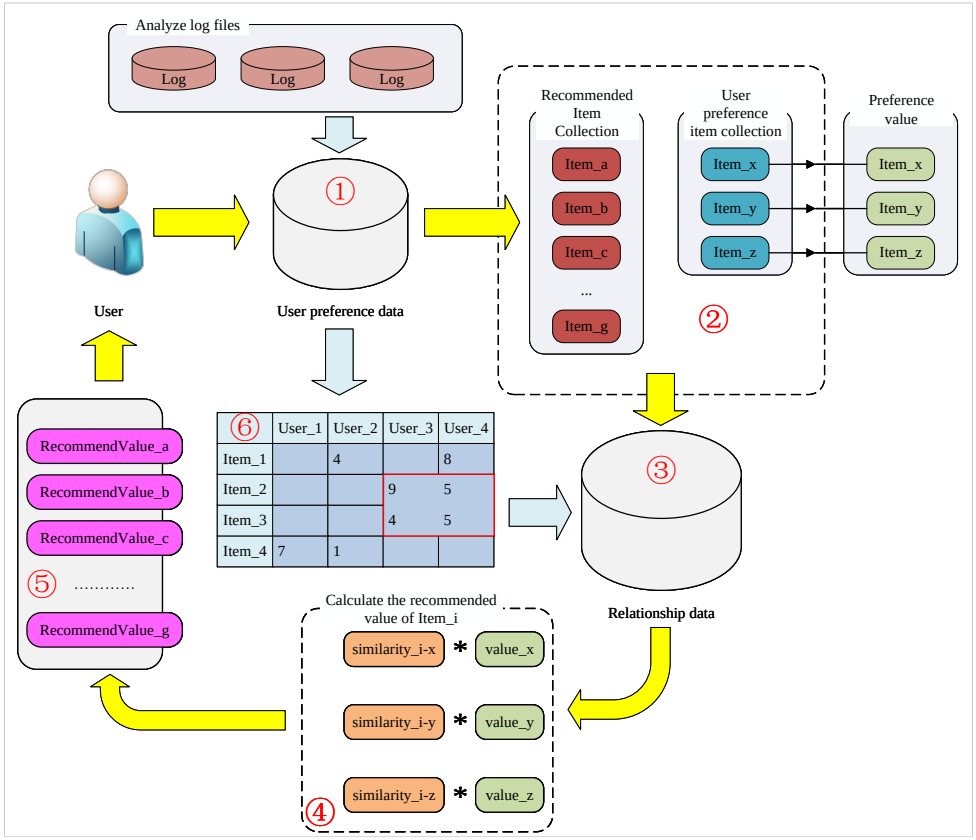

Figure 1. The structure of the item-based CF

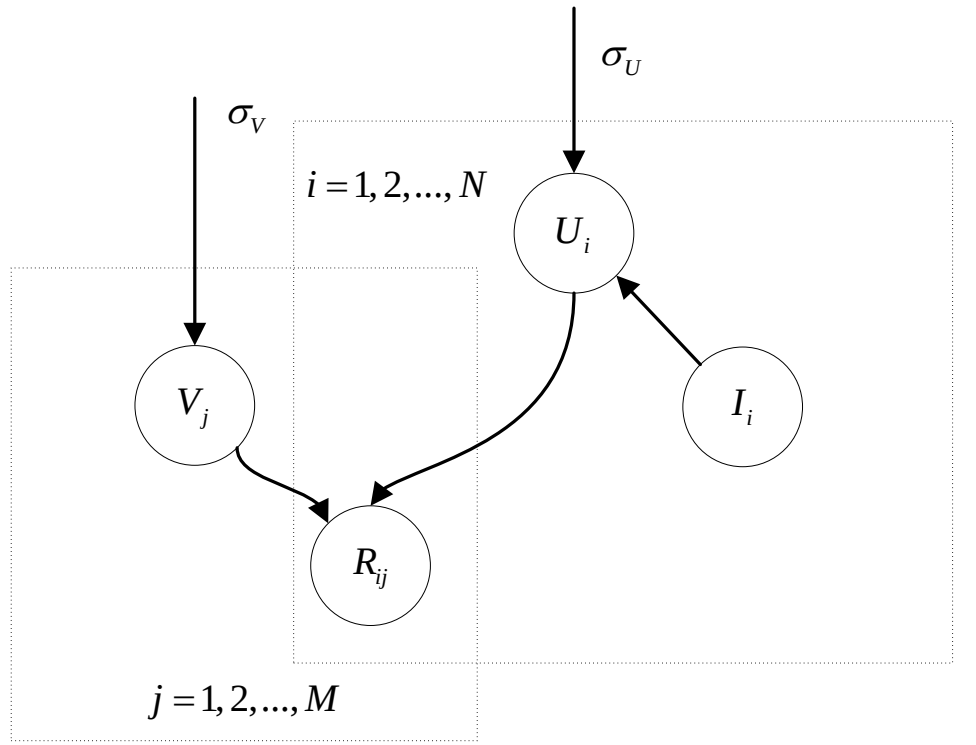

Figure 2. The PMF model

3. **Optimization of logistic regression models based on non-convex penalty regression.** Despite the simplicity and wide applicability of the classical Logistic model, the Logistic model breaks down when there are serious outliers in the data, for this reason, this paper introduces a mean drift parameter into the traditional Logistic model, thus proposing a robust Logistic regression method based on non-convex penalized regression. In addition, an iterative algorithm based on iterative threshold embedding is designed to ensure that the improved logistic regression method can simultaneously perform outlier detection and robust parameter estimation.

Drawing on the study of She and Owen [24], this paper develops the following robust logistic regression model.

$$\pi' = \ln\left(\frac{\pi}{1-\pi}\right) = X^T\beta + \gamma \tag{6}$$

where $X = (1, x_1, \ldots, x_{p-1})$, x_j are the j-th explanatory variables; $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$, β_j are the impact coefficients of the j-th explanatory variables; and $\gamma = (\gamma_1, \ldots, \gamma_n)^T$, γ_i is the mean drift parameter of the i-th observation. Obviously, the addition of a mean drift parameter without any constraints will over-parameterize the model. In this paper, assuming that γ is sparse and γ_i is non-zero when the observations are outliers, and conversely, γ_i takes the value of zero. Thus, the sparse estimation of γ provides a straightforward method for detecting outliers in a logistic regression model. The goal of this paper is to accurately estimate γ thus determining which data are outliers and to obtain a robust estimate of β. The following objective function is established and γ can be estimated by minimizing Equation (7).

$$f(\beta, \gamma; \lambda) = \frac{1}{2}\|\pi' - X\beta - \gamma\|_2^2 + \sum_{i=1}^{n} P_\lambda(|\gamma_i|) \tag{7}$$

where $P_\lambda(|\gamma_i|)$ is the penalty function of γ_i, and γ is the tuning parameter to control the degree of penalty. λ is an adaptive parameter. There are many choices of penalty functions: the commonly used l_1-paradigm penalty function $P_\lambda(|\gamma_i|) = \lambda|\gamma_i|$, l_0-paradigm penalty function $P_\lambda(|\gamma_i|) = \frac{\lambda^2}{2} I(\gamma_i \neq 0)$. By choosing a robust estimate of β as the initial value $\beta^{(0)}$, the new value of the dependent variable is obtained as follows.

$$Y_i(0) = x_i^\top \beta^{(0)} + \frac{Y_i - \frac{\exp\{x_i^\top \beta^{(0)}\}}{1+\exp\{x_i^\top \beta^{(0)}\}}}{w_i(0)} \tag{8}$$

where $i = 1, 2, \ldots, n$, $w_i(0)$ are weights with the following expression.

$$w_i(0) = \frac{\exp\{x_i^\top \beta^{(0)}\}}{1+\exp\{x_i^\top \beta^{(0)}\}} \times \left[1 - \frac{\exp\{x_i^\top \beta^{(0)}\}}{1+\exp\{x_i^\top \beta^{(0)}\}}\right] \tag{9}$$

Let $Y(0) = (Y_1(0), \ldots, Y_n(0))^T$, the initial value of the mean drift parameter γ can be expressed as $\gamma^{(0)} = Y(0) - X\beta^{(0)}$. This paper proposes an iteration-based threshold embedding algorithm to update γ, i.e., $\gamma^{(s+1)} \leftarrow \Theta(\xi; \lambda^*)$, where Θ is the threshold norm of l_1 norm and l_0 norm [25], $\xi = H\gamma^{(s)} + (I-H)Y(0)$, $\lambda^* = \lambda\sqrt{1-h_i}$, H is the hat matrices, and $H = X(X^TX)^{-1}X^T$, h_i is the i-th element of the diagonal of the H matrix.

In addition, to choose the optimal tuning parameter λ, the Bayesian Information Criterion (BIC) [26] is applied because of its computational efficiency and superiority in variable selection in the following form.

$$\begin{cases} BIC(\lambda) = \log(RSS) + (\log(n) + 1)k(\lambda) \\ RSS = \|(I-H)(Y(0) - \hat{\gamma})\|_2^2 \\ k(\lambda) = DF(\lambda) + 1 \end{cases} \tag{10}$$

For a specific value of λ, a set of estimates $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_n)^T$ of γ is obtained, and $DF(\lambda)$ denotes the number of non-zero elements in the estimated $\hat{\gamma}$-vector. By selecting n λ values equally spaced in the interval $[\lambda_{\min}, \lambda_{\max}]$ and bringing them into the threshold embedding algorithm described above, n different sets of γ estimates can be obtained, where the smallest BIC value corresponds to the optimal λ value and the optimal solution of γ.

4. **Neural network model-based training model design for big data training of hospital online diagnosis and treatment information.** To achieve accurate recommendation of hospital online diagnosis and treatment data, this paper firstly uses big data and artificial intelligence technology to collate hospital online diagnosis and treatment information data, and constructs the corresponding big

data training model. Literature research has found that diagnosis and treatment information is heterogeneous and complex. Therefore, this paper chooses to build a neural network to realize the training of information big data.

After obtaining the electronic medical records of a large number of patients, we extract the keywords of the medical records and record them as dimension 1, and then we construct a neural network on this dimension. The hidden layer of this network is set as A, the input layer is set as a, and the output layer is set as b. The following formula is used.

$$\begin{cases} A < b - 1 \\ A < \sqrt{a + b} + c \\ A = \lg b \end{cases} \tag{11}$$

where c represents the neurons to be added. The weight from the input level to the obscured level is set to E, the threshold is set to s, the weight from the obscured level to the input level is set to F, the threshold is set to x. The forward neural network is constructed, and the hidden layer, output layer and error of each information node are shown in Equation (12), Equation (13) and Equation (14), respectively.

$$G_j = \sum_{i=1}^{n} (E_{i,j} X_i) + s_j \tag{12}$$

$$G_{out} = \sum_{i=1}^{n} (F_{i,j} Z_i) + x_j \tag{13}$$

$$V(u) = \frac{\sum_{i=1}^{n} (R_i - Y_i)}{3} \tag{14}$$

where Y_i is the real data value, Z_i is the associated data value, X_i is the input raw data information, and x_j is the processed data value. Through the above processing, the raw data are integrated into three-dimensional form and trained to complete the data training process. The trained data are stored in the platform database to provide the data basis for the subsequent diagnosis and treatment.

After determining the matching nodes of each data, the neural network is used to construct the diagnosis and treatment information discriminant function as follows.

$$d_i(u) = \sum_{i=1}^{D} (x_i - Y_{ij})^2 \tag{15}$$

Applying this formula, the data in the database is activated and updated with the data weights.

$$T_{i,j(x)} = \exp(\frac{-Y_{ij}(x)}{2\eta^2}) d_i(u) \tag{16}$$

where η is the platform's preset data classification weight. Based on the above, the electronic medical records were classified into categories and the results were evaluated using the following formula.

$$DBI = \frac{\sum_{i=1}^{c} \max \left\{ \frac{\Delta(Y_i) + \Delta(Y_j)}{\alpha(Y_i, Y_j)} \right\}}{g T_{i,j(x)}} \tag{17}$$

where (Y_i, Y_j) is the distance between different categories of data; g is the weight of the number of categories, the smaller the DBI value is, the better the data training effect is. The above settings are used as the core data for the construction of the model to provide data support for the subsequent recommendation model.

5. Accurate recommendation model design for hospital online diagnosis and treatment data based on improved logistic regression model.

5.1. Data correlation based on improved logistic regression models.

After applying neural networks to set up a training model for electronic medical record big data, an improved logistic regression model was constructed to correlate different categories of medical data, and a similarity constraint matrix with latent was introduced to impact the user's eigenvectors, and the improved logistic regression model was adopted to represent the potential factors of the non-linear relationship. Through the Gibbs sampling method [27] for hyper-parameter sampling and the use of Bayesian networks to calculate the maximum a priori and a posteriori probabilities, integrated with the Markov chain Monte Carlo approach for training purposes, to obtain the user rating information on the resources, select the PageRank method [28] on the ratings ranked in the top N of the diagnosis and treatment resources recommended for optimisation, to

complete the hospital online diagnosis and treatment data accurate recommendation. The structure of the designed recommendation model is shown in Figure 3.

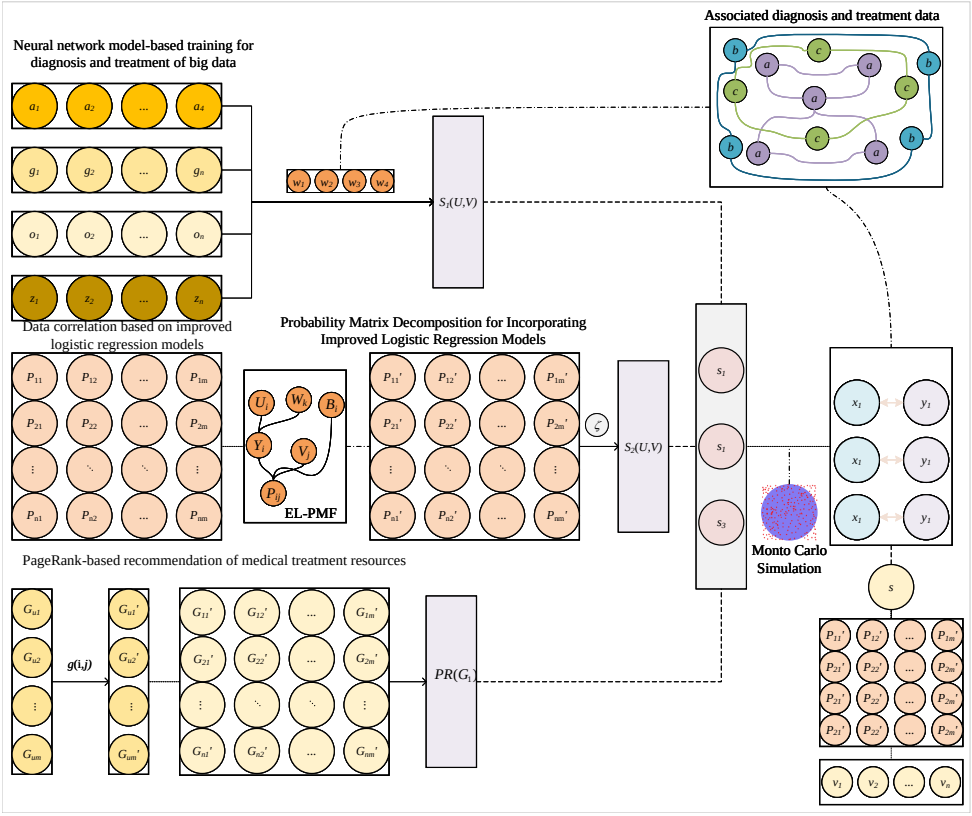

Figure 3. The structure of the designed recommendation model

To ensure that the algorithms are intelligently linked in the process of recommending medical data, the database in this study is set as a node type, so as to facilitate the analysis and organisation of medical data. The trained data are stored in different databases with different functions, and to improve the accuracy of extraction, an optimized logistic regression model is constructed in the database to ensure the intrinsic recommendation connectivity of the medical data, and the calculation formula of the improved logistic regression model is set as follows.

$$G = \frac{1}{1 + \exp(\gamma_0 + \gamma_0 k_0 + \gamma_1 k_1 + \cdots + \gamma_n k_n)} \tag{18}$$

where γ_0 is the distance of medical data categories between medical databases; k is the independent variable of the regression equation; and k_n is the amount of change in medical data categories.

The above equation can control the correctness of the information extraction results when the diagnosis and treatment information changes. The equation is implanted into the database, and the data storage mode is changed to binary mode to ensure its normal application in the recommendation algorithm.

5.2. **Probability matrix factorization for incorporating improved logistic regression models.** After associating the diagnosis and treatment data, this paper associates a constraint vector W for each diagnosis and treatment data that is independent of the basic feature vector, so that when calculating the user's eigenvalue, all the constraint vectors related to the user's participation in the evaluation data will have an impact on it, thus effectively preventing the user's eigenvalue from being too close to the mean of its a priori distribution. Meanwhile, to more accurately represent the nonlinear connection among the latent factors, this study conveys the nonlinear relationship of the potential significance factors through an improved logistic regression model, and the proposed EL-PMF structure is shown in Figure 4, where U_i is the original user feature vector, V_j is the healthcare resource feature vector, W_k is the constraint feature vector, and F_i is the new feature vector with U_i and W_k together.

In this paper, a constraint mechanism based on latent similarity is first introduced in order to avoid user features being too close to the centroid of the prior distribution. A constraint vector named W is

introduced, which is used to adjust and limit the representation of the original user characteristic vector, and the new user characteristic vector model is constructed as shown below.

$$F_i = U_i + \frac{\sum_{k=1}^{M} I_{ik} W_k}{\sum_{k=1}^{M} I_{ik}} \tag{19}$$

where I_{ik} is whether item k has been rated by user i. If the rating has taken place, $I_{ik} = 1$, otherwise $I_{ik} = 0$.

Assume that user i's rating R_{ij} of resource j follows a Gaussian distribution having a mean value of $B_i g(Y_i^T V_j)$ and variance of δ^{-1}. The user i's ratings of resource j are distributed as follows. To construct a new probability objective function for the EL-PMF, the conditional distribution of user ratings is defined as follows.

$$P(R|U, V, B, \alpha^{-1}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left[\mathcal{N}(R_{ij}|B_i G(Y_i^T Y_j), \delta^{-1})\right]^{I_{ij}} \tag{20}$$

where $G(x)$ is a modified logistic regression model and B_i is a parameter of user i's rating scale. Assuming that B obeys a Gaussian distribution with mean μ_B and variance Λ_B, the samples are extracted using the Markov Chain-Monte Carlo approach [29], and then the complex objective function is approximated using Equation (21).

$$P(R_{ij}|R, \Lambda_0) = \frac{1}{T} \sum_{t=1}^{T} P(R_{ij}|U_i^t, V_j^t, W_k^t, B_i^t) \tag{21}$$

where $\{U_i^t, V_j^t, W_k^t, B_i^t\}$ is produced through running a Markov chain, the Monte Carlo-based approach has the advantage of gradually producing accurate results. In this paper, a Markov chain with smooth distribution of expected joint probability is constructed, and the hyperparameters are sampled in T rounds with the help of Gibbs sampling method, in which the eigenvalues corresponding to the eigenvectors U, V, W and B are continuously traversed in a cyclic manner and updated to obtain the final MF results.

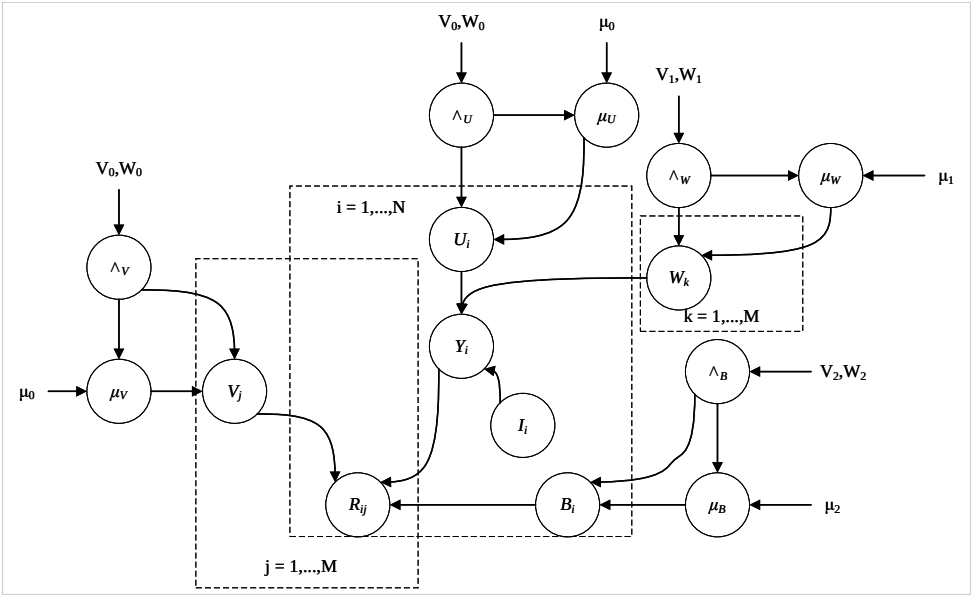

Figure 4. The proposed EL-PMF structure

5.3. **PageRank-based recommendation of medical treatment resources.** After obtaining the user rating data of the resources, this paper chooses the PageRank method to process the recommendation of diagnosis and treatment resources with the top N ratings. In the training process of massive medical big data, if a certain diagnosis and treatment information is linked by a lot of information, it means that this information is more important and the PageRank value is relatively high. Based on this assumption, the recommendation process is completed as follows.

The probability of visiting each medical resource is set, and each medical resource consists of $G_1, G_2, \ldots, G_n$ part of the degree of association, which is calculated by the formula (18) above. If all resources are linked

to G_1, then the PageRank of G_1 is the sum of the remaining $n-1$ parts, as shown below.

$$PR(G_1) = PR(G_2) + PR(G_3) + \cdots + PR(G_n) \tag{22}$$

In general, there will not be only one link per resource, but assuming that there are two or more links, the following equation applies.

$$PR(G_1) = \frac{PR(G_2)}{n'} + \frac{PR(G_3)}{n'} + \cdots + \frac{PR(G_n)}{n'} \tag{23}$$

where n' is the number of links. The final formula for recommending medical resources is as follows.

$$PR(G_1) = \frac{PR(G_2)}{L(G_2)} + \frac{PR(G_3)}{L(G_3)} + \cdots + \frac{PR(G_n)}{L(G_n)} \tag{24}$$

where $L(G_n)$ is the number of resource link nodes and databases. Apply this formula to complete the construction of the diagnosis and treatment resource recommendation module, and install the module into the original hospital online diagnosis and treatment platform.

6. **Performance testing and analysis.**

6.1. **Analyzing the results of the recommendation of online medical treatment data.**

In this paper, the online consultation data of a hospital collected from the literature [30] is selected as the dataset, which contains a total of 105,291 rating records of 397 healthcare resources by 6,039 users, and the dataset is divided in the ratio of 8:2, where 80% is used as the training set and the remaining 20% as the testing set. This experiment was conducted on a Lenovo laptop with Intel (R) Core (TM) i5-4258U CPU @ 2.40GHz processor and 12GB RAM. The experiments are programmed in Python, version 3.8, and the compilation environment is implemented in Anaconda's Spyder. In the experiments, the dimension of potential characteristics was established as 30, the amount of epochs was established as 50, the amount of instances was established as 100, the learning rate was set to 0.05, and the regularization factor was established as 0.02.

In this study, the information data to be recommended are set into two parts: "image" and "data," as shown in Table 1. The data in Table 1 are stored in the database, and the target data are output according to the preset query requirements to determine the information processing capability of each model.

Table 1. Diagnosis and treatment informationization platform test data

Test group number	Total information data/item	Data information/item	Image information/item
1	4287	1435	2852
2	6684	3265	3419
3	4677	3157	1520
4	7542	1315	6227
5	8463	6526	1937
Total	31653	15698	15955

The diagnosis and treatment images with features are selected as the test object, and this part of the images is retrieved using the models EL-PMF and TBCF [10], DLCF [13] and ANLO [16] in the paper, and the recommended results are analyzed, as shown in Table 2. The analysis of the data in the above table shows that the EL-PMF can recommend 99.0% of the online diagnosis and treatment data of hospitals after application, and the data are more complete. TBCF can only recommend 87.63% of the data, and the query results have a large number of missing items, while DLCF and ANLO can only recommend 94.12% and 97.13% of the data, respectively, and the data completeness is not high. Combining the above analyses, the optimization of the logistic regression model and the use of diagnosis and treatment data recommendation using the optimized model greatly improves the recommendation effect.

Table 2. Recommendation results of diagnosis and treatment data from different models

Test group number	Query result of EL-PMF/item	Query result of TBCF/item	Query result of DLCF/item	Query result of ANLO/item
1	4285	4200	4248	4269
2	6684	6154	6359	6492
3	4673	4032	4398	4527
4	7542	7065	7218	7456
5	8461	8143	8217	8391

6.2. **Performance analysis of online clinic data recommendation.** In addition to analyzing the diagnosis and treatment data recommendation results, this paper also uses the quantitative metrics Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Normalized Discounted Cumulative Gain NDCG@N, and Hit Rate HR@N, where N is the number of diagnosis and treatment resources. The recommendation performance of different models is analyzed. The RMSE and MAE for the different models are shown in Figure 5. From Figure 5(a) & Figure 5(b), it can be seen that the RMSE of EL-PMF decreases by about 0.13, 0.11 and 0.10 compared to TBCF, DLCF, and ANLO, respectively, as the number of iterations is incremented. Compared with TBCF, DLCF, and ANLO, the EL-PMF model experiences a decrease in MAE by approximately 0.04, 0.03, and 0.03 respectively. In comparison to the other three models, EL-PMF has smaller rating error values, faster convergence of the model, better extraction of potential eigenfactors in EL-PMF, and higher rating prediction accuracy. Therefore, EL-PMF is better than other models in solving the sparse rating problem. The EL-PMF in this paper can better solve the recommendation problem caused by sparse ratings and improve the recommendation quality to some extent.

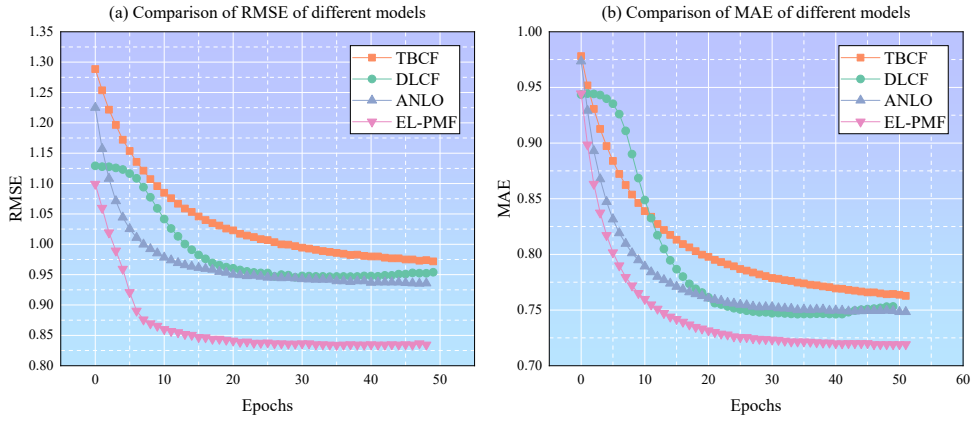

Figure 5. The RMSE and MAE for the different recommendation models

In order to fully verify the advantages of the EL-PMF model, the comparison of NDCG and HR of different models when N is taken as 5, 10 and 20 is shown in Figure 6. When N is taken as 5, the NDCG and HR of EL-PMF are 0.5637 and 0.6842, respectively, which are at least 12.33% and 6.69% higher compared to the other three models, respectively. When N is taken as 20, the NDCG and HR of EL-PMF are 0.7827 and 0.9635, respectively, which are at least 8.09% and 4.21% higher compared to the other three models, respectively. Although TBCF added time weights to the user rating matrix, it did not optimize MF, resulting in poor recommendations. DLCF combines CF with DL models for resource recommendation, but suffers from data sparsity. ANLO predicts the final score of a resource by combining a neural network with a logistic regression model, but does not optimize the logistic regression model, and therefore the recommendation is not as efficient as EL-PMF. Overall, the recommended performance of EL-PMF has some advantages.

7. **Conclusion.** As the Internet medical treatment rapidly growing, online diagnosis and treatment services in hospitals are becoming increasingly popular. How to mine valuable information from massive

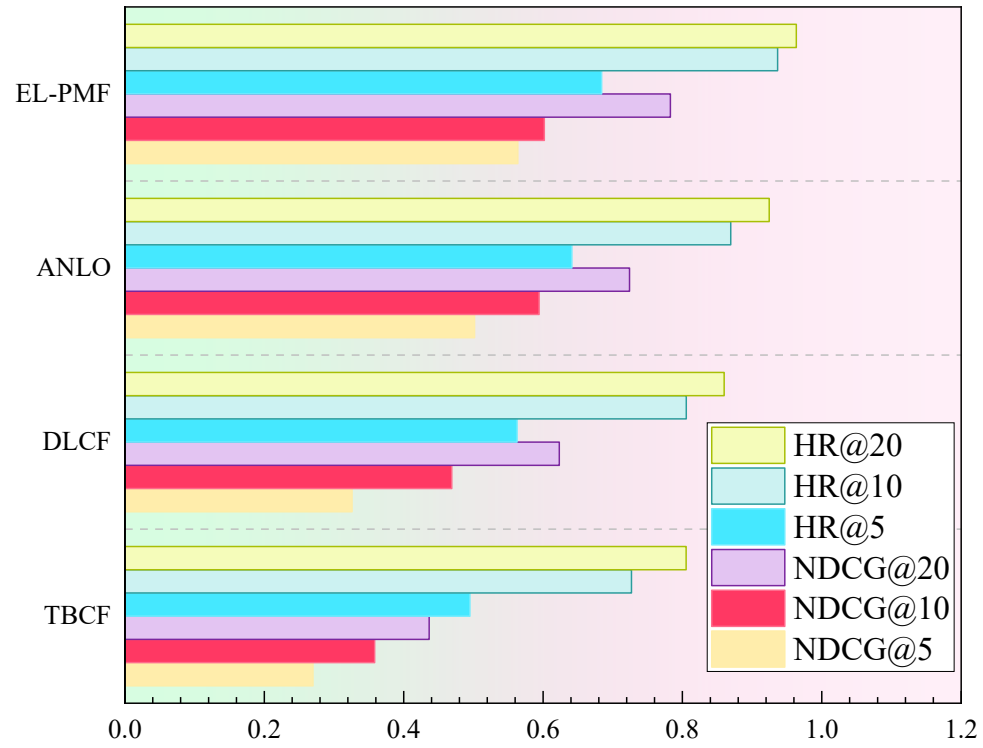

Figure 6. The comparison of NDCG and HR of different models

online diagnosis and treatment data and achieve accurate recommendation to improve diagnosis and treatment efficiency is imminent. To address the problem of unsatisfactory recommendation efficiency in the existing research, this paper firstly improves the logistic regression model based on the idea of non-convex penalty regression, add the sparse and situation-specific mean-drift parameter vectors into the logistic regression model to identify the outliers, and then generate sparsity by the penalty function, and then use the corresponding threshold law to screen the outliers. Subsequently, different information criteria are used to determine the tuning parameters and the choice of optimal parameter solutions for robust estimation of the results. Then the neural network is used to set up the big data training model of electronic medical record, and the key discriminant function of diagnosis and treatment information is built. A modified Logistic regression model is used to correlate different categories of diagnosis and treatment data to represent the non-linear relationship of potential features, which incorporates the Markov chain Monte Carlo method for the purpose of training by introducing a similarity constraints item matrix to influence the user's feature vector, and obtaining the user's rating data for the resource. The PageRank method is chosen to optimize the recommendation of treatment resources with the top N scores, and complete the accurate recommendation of hospital online treatment data. The experimental outcome implies that the HR of the proposed model is 0.6842, which improves at least 6.69% compared to the other three models respectively, and improves the recommendation performance to a larger extent.

In this paper, a modified logistic regression model is used to represent the nonlinear relationships of potential features. For further research, other linear and non-linear functions can also be considered as modules for conveying potential feature relationships. In addition, the hyperparameter sampling in this study is mainly done by Gibbs sampling. The subsequent research can improve the sampling effect by defining more hyperparameters and reduce the computational complexity to a certain extent.

Acknowledgment. This work was supported by the Key Project of the Annual Regular Program of the Zhejiang Provincial Philosophy and Social Sciences Planning, titled "Research on Digital Ethical Dilemmas of DeepSeek-based Generative AI from the Perspective of New-Quality Productive Forces" (No. 26NDJC064Z).

REFERENCES

[1] J. Chen, K. Li, H. Rong, K. Bilal, N. Yang, and K. Li, "A disease diagnosis and treatment recommendation system based on big data mining and cloud computing," *Information Sciences*, vol. 435, pp. 124-149, 2018.

[2] R. F. Mansour, A. El Amraoui, I. Nouaouri, V. G. Díaz, D. Gupta, and S. Kumar, "Artificial intelligence and internet of things enabled disease diagnosis model for smart healthcare systems," *IEEE Access*, vol. 9, pp. 45137-45146, 2021.

[3] S. Tian, W. Yang, J. M. Le Grange, P. Wang, W. Huang, and Z. Ye, "Smart healthcare: making medical care more intelligent," *Global Health Journal*, vol. 3, no. 3, pp. 62-65, 2019.

[4] U. A. Bhatti, M. Huang, D. Wu, Y. Zhang, A. Mehmood, and H. Han, "Recommendation system using feature extraction and pattern recognition in clinical care systems," *Enterprise Information Systems*, vol. 13, no. 3, pp. 329-351, 2019.

[5] J. Shu, X. Shen, H. Liu, B. Yi, and Z. Zhang, "A content-based recommendation algorithm for learning resources," *Multimedia Systems*, vol. 24, no. 2, pp. 163-173, 2018.

[6] N. Liang, Z. Yang, Z. Li, W. Sun, and S. Xie, "Multi-view clustering by non-negative matrix factorization with co-orthogonal constraints," *Knowledge-Based Systems*, vol. 194, 105582, 2020.

[7] F. Colace, D. Conte, M. De Santo, M. Lombardi, D. Santaniello, and C. Valentino, "A content-based recommendation approach based on singular value decomposition," *Connection Science*, vol. 34, no. 1, pp. 2158-2176, 2022.

[8] P. Symeonidis, and D. Malakoudis, "Multi-modal matrix factorization with side information for recommending massive open online courses," *Expert Systems with Applications*, vol. 118, pp. 261-271, 2019.

[9] X Liu, "An improved clustering-based collaborative filtering recommendation algorithm," *Cluster Computing*, vol. 20, pp. 1281-1288, 2017.

[10] C. Zhang, M. Yang, J. Lv, and W. Yang, "An improved hybrid collaborative filtering algorithm based on tags and time factor," *Big Data Mining and Analytics*, vol. 1, no. 2, pp. 128-136, 2018.

[11] K. Przystupa, M. Beshley, O. Hordiichuk-Bublivska, M. Kyryk, H. Beshley, J. Pyrih, and J. Selech, "Distributed singular value decomposition method for fast data processing in recommendation systems," *Energies*, vol. 14, no. 8, 2284, 2021.

[12] J. Fang, B. Li, and M. Gao, "Collaborative filtering recommendation algorithm based on deep neural network fusion," *International Journal of Sensor Networks*, vol. 34, no. 2, pp. 71-80, 2020.

[13] M. Fu, H. Qu, Z. Yi, L. Lu, and Y. Liu, "A novel deep learning-based collaborative filtering model for recommendation system," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1084-1096, 2018.

[14] X. Cai, Z. Hu, and J. Chen, "A many-objective optimization recommendation algorithm based on knowledge mining," *Information Sciences*, vol. 537, pp. 148-161, 2020.

[15] D. Lian, X. Xie, and E. Chen, "Discrete matrix factorization and extension for fast item recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 5, pp. 1919-1933, 2019.

[16] J. Zhang, Y. Lv, J. Hou, C. Zhang, X. Yua, Y. Wang, T. Yang, X. Su, Z. Ye, and L. Li, "Machine learning for post-acute pancreatitis diabetes mellitus prediction and personalized treatment recommendations," *Scientific Reports*, vol. 13, no. 1, 4857, 2023.

[17] S. Domínguez-Almendros, N. Benítez-Parejo, and A. R. Gonzalez-Ramirez, "Logistic regression models," *Allergologia Et Immunopathologia*, vol. 39, no. 5, pp. 295-305, 2011.

[18] J. Kuha, and C. Mills, "On group comparisons with logistic regression models," *Sociological Methods & Research*, vol. 49, no. 2, pp. 498-525, 2020.

[19] F. Wang, H. Zhu, G. Srivastava, S. Li, M. R. Khosravi, and L. Qi, "Robust collaborative filtering recommendation with user-item-trust records," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 4, pp. 986-996, 2021.

[20] P. B. Thorat, R. M. Goudar, and S. Barve, "Survey on collaborative filtering, content-based filtering and hybrid recommendation system," *International Journal of Computer Applications*, vol. 110, no. 4, pp. 31-36, 2015.

[21] J. Liu, C. Wu, Y. Xiong, and W. Liu, "List-wise probabilistic matrix factorization for recommendation," *Information Sciences*, vol. 278, pp. 434-447, 2014.

[22] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 3964-3977, 2012.

[23] S. A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1364-1377, 2010.

[24] Y. She, and A. B. Owen, "Outlier detection using nonconvex penalized regression," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 626-639, 2011.
[25] P. A. Loring, "Threshold concepts and sustainability: Features of a contested paradigm," *Facets*, vol. 5, no. 1, pp. 182-199, 2020.
[26] M. Drton, and M. Plummer, "A Bayesian information criterion for singular models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 79, no. 2, pp. 323-380, 2017.
[27] D. Marcotte, and D. Allard, "Gibbs sampling on large lattice with GMRF," *Computers & Geosciences*, vol. 111, pp. 190-199, 2018.
[28] S. S. Shaffi, and I. Muthulakshmi, "Weighted PageRank algorithm search engine ranking model for web pages," *Intelligent Automation & Soft Computing*, vol. 36, pp. 183-92, 2023.
[29] C. Andrieu, A. Doucet, and R. Holenstein, "Particle markov chain monte carlo methods," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 72, no. 3, pp. 269-342, 2010.
[30] J. Jiang, A.-F. Cameron, and M. Yang, "Analysis of massive online medical consultation service data to understand physicians' economic return: Observational data mining study," *JMIR Medical Informatics*, vol. 8, no. 2, e16765, 2020.