ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Assessment of the Impact of Big Data Auditing on Hospital Internal Controls in Smart Healthcare Construction

Siyu Chen[1,*]

[1] Department of Research Administration, Shandong Second Provincial General Hospital, Jinan, Shandong, 250023, China

SUMMARY: *This study assesses the mechanism of influence and methods of improvement of big data auditing on the quality of internal control in hospitals against the background of smart healthcare construction. First, we build the framework of smart audit platform covering audit management, audit operation and audit collaboration system. Then, with the help of Lasso idea, the penalty parameters of the model are set and the variables that are not significantly related to the model are eliminated. After variable screening, the key variables of internal control quality of hospitals are retained by fusing the nonlinear model Logistic, and whether the internal control quality of hospitals meets the standard is assessed. The SHAP method was then utilized for feature attribution analysis to enhance the predictive efficacy and interpretability of the model. The results showed that the Lasso-Logistic model had over 90% accuracy in discriminating the quality of internal control in hospitals, revealing the non-linear association between the hospital's large equipment utilization rate and prescription compliance rate with the quality of internal control. This study not only quantifies the impact of big data auditing on the quality of internal control in hospitals, but also provides a method for optimizing management strategies and improving the quality of internal control in hospitals.*

KEYWORDS: *Lasso-Logistic model; SHAP method; internal control quality; feature attribution; big data audit*

1 Introduction

In the era of big data, with the application of Internet of Things, big data and other technologies in the medical field, a new type of intelligent medical service model has been formed. The construction and development of smart healthcare realizes telemedicine and self-help healthcare through information technology, which is conducive to alleviating the pressure of shortage of medical resources; to the sharing and exchange of medical information and resources, thus significantly improving the rationalization of medical resources allocation; and to the modernization of healthcare services and the level of services [1-5]. Big data auditing, as a key part of smart medical construction, optimizes the auditing process.

Under the new situation, the audit work is especially inseparable from the support of big data, and the use of big data for auditing will become an important means for audit institutions to deal with the complex social and economic management situation and improve the quality of audit work. Literature [6] points out that big data technology has been deeply integrated into the audit process, covering data collection, storage, screening and analysis, significantly improving the breadth and depth of the audit; and emphasizes that in medical insurance auditing,

*Chenxiaoxiao940917@163.com
https://doi.org/10.65102/is2026664 Published: 30 April, 2026

big data auditing helps to ensure the safety of the fund, improve the efficiency and reduce the cost, and its application is of great significance in safeguarding the well-being of people's lives and promoting the healthy development of the social insurance system. Literature [7] focuses on big data auditing, emphasizing that big data analytics improves audit quality by integrating structured and unstructured data, and reduces sampling dependency and risk; it provides real-time monitoring and in-depth financial insights, and helps strategic planning and compliance supervision. Literature [8] optimizes the financial audit method based on big data by using the powerful multi-source feature abstraction and fusion capability of convolutional neural network, which can effectively improve the accuracy of audit judgment. Literature [9] illustrates that the application of big data analytics in financial auditing has great potential, and its effective implementation relies on information technology knowledge, data categorization, and multivariate skills, but also faces challenges such as data size and lack of standards.

On the impact of big data auditing on internal control. Literature [10] showed that in the era of big data, the independence of the internal audit department in reporting to the audit committee, the soft skills of the chief audit executive and his involvement in risk management, fraud and IT audits are positively associated with the use of data analytics. Literature [11] found that big data technology can enhance audit efficiency, especially in internal control, risk management and fraud detection by identifying anomalies and causality advantages. Literature [12] identifies key risk factors affecting internal control governance in a big data environment, reveals their complex associations, and proposes a hybrid architecture that combines AI rule generation and multi-rule base decision-making, with the control environment and IT control construction as the core dimensions, which provides a key path to optimize the internal control of the enterprise and improve the success rate of auditing. Literature [13] based on data analysis of 102 companies found that big data applications significantly improved the internal control environment and communication channels, and emphasized that combining business intelligence and blockchain technology plays a key role in supporting accounting practices and enhancing auditor functions. Most of the current research on the impact of big data auditing on internal control targets enterprises and lacks research on healthcare organizations. And with the construction of smart healthcare, it has an impact on the internal control of hospitals, and the impact of big data auditing on it should be assessed to continue to optimize hospital auditing.

This study integrates predictive modeling algorithms and interpretable analysis methods to construct a method for assessing the impact of big data auditing on hospital internal control. First, a Lasso-Logistic regression model was used to screen out key influencing factors from the potential influencing factors, avoiding the overfitting problem that is prone to occur during data analysis, and constructing an assessment model of internal control quality. To further reveal the mechanism of each factor, this study introduces the SHAP attribution method, which quantifies the global contribution of each feature to the prediction results and enhances the interpretability and transparency of the model. Multidimensional data of 360 comprehensive tertiary hospitals in China from 2019-2023 were collected, and their internal control quality was predicted using the Lasso-Logistic model, and then the SHAP method was applied to interpret the specific degree of influence of the factors to validate the usability of this paper's method.

2 Logical framework construction of hospital intelligent audit platform

2.1 Supply and Demand for Smart Auditing in Hospitals

This paper draws on scholars about the audit quality supply and demand framework model, on the basis of its proposed static model, to make an analysis of the hospital smart audit supply and demand, the hospital smart audit supply and demand structural framework shown in Figure 1.

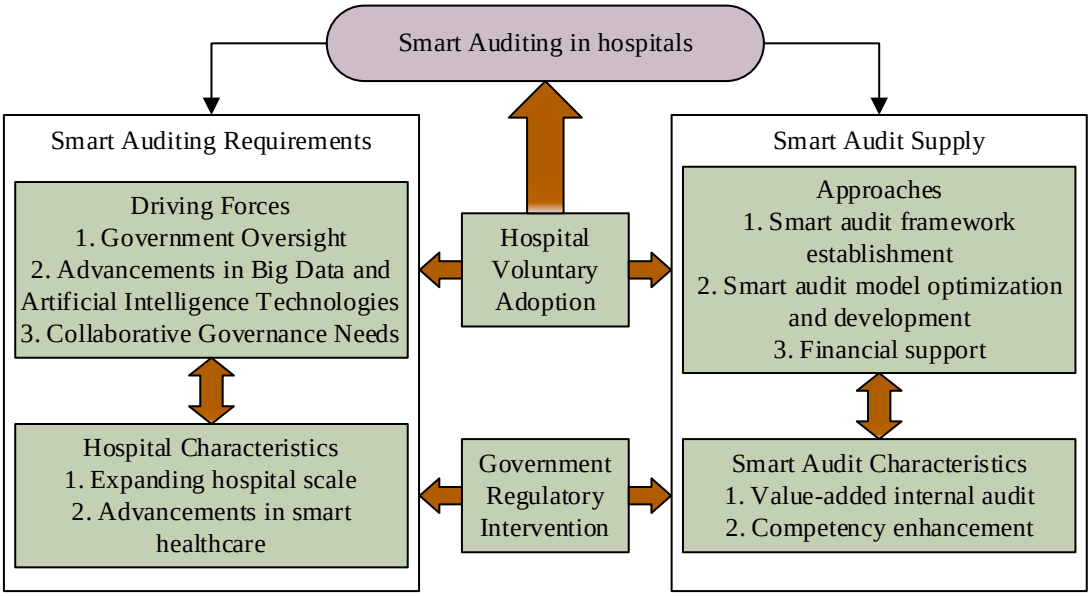

Figure 1: Hospital intelligent audit supply and demand structure framework

2.2 Framework Construction of Hospital Smart Audit Platform

Based on the theory of reference information system, this paper builds hospital intelligent audit platform with the help of cloud computing, big data, Internet of Things, blockchain, artificial intelligence and other technologies. Through data processing, risk early warning, audit management, audit collaboration and other ways to sort out the audit operation from audit preparation, audit implementation, audit report, audit rectification and ultimately to the audit archive of the whole process of closed-loop management. The architecture of the hospital intelligent audit platform constructed in this paper is shown in Figure 2.

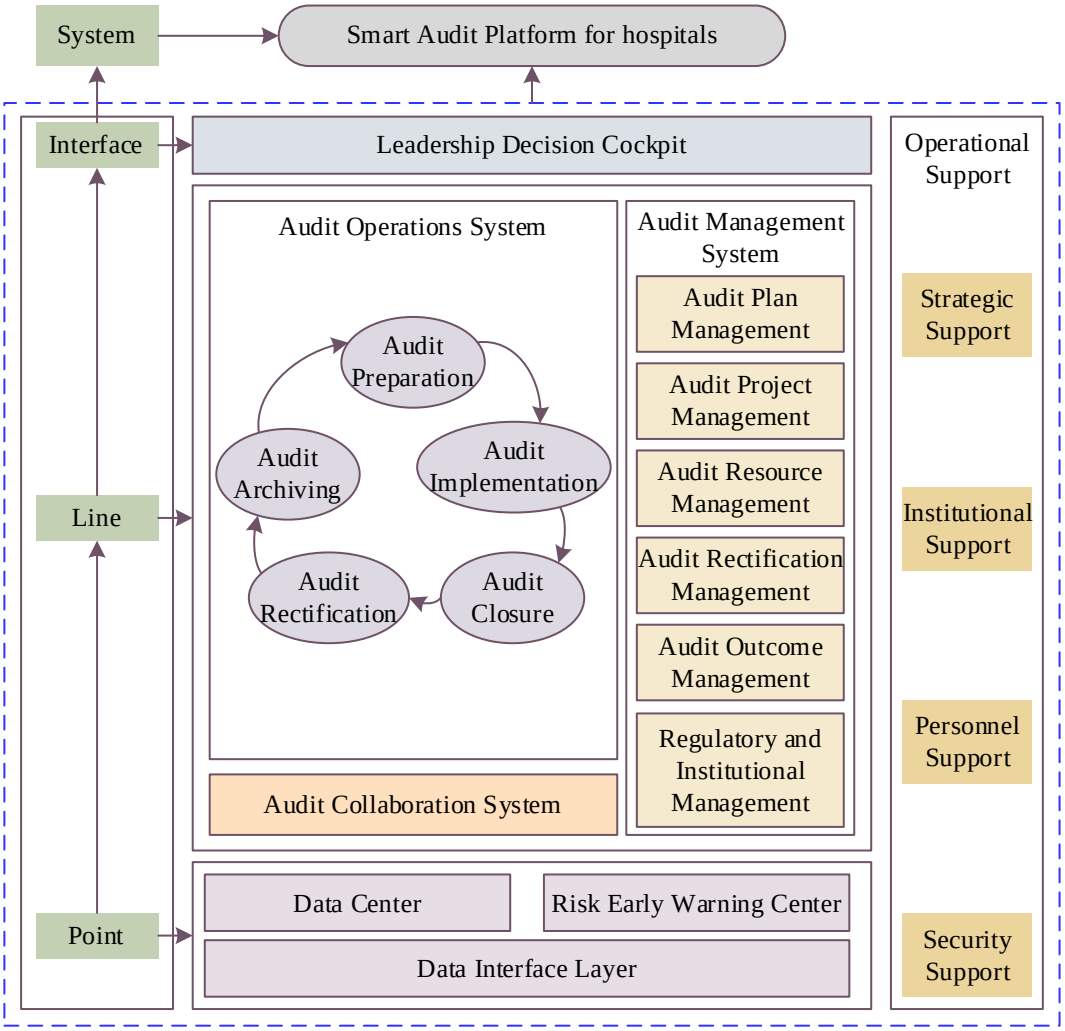

Figure 2: Hospital intelligent audit platform architecture

2.2.1 Audit management system

The audit management system is the control center of internal audit management, which integrates the management of audit knowledge, audit resources and other information, realizes the interactive sharing of information, and thus enables the auditors to grasp the progress and effectiveness of their work in real time. The audit management system includes audit plan management, audit project management, audit resource management, audit result management, audit correction management and audit system management.

2.2.2 Audit operating system

The audit operation system is based on the requirements and guidelines of the Audit Commission, IAIS and regulatory bodies, and realizes the closed-loop management of the whole process from audit preparation, audit implementation, audit report, audit rectification and audit archiving. Figure 3 shows an overview of the audit operation process.

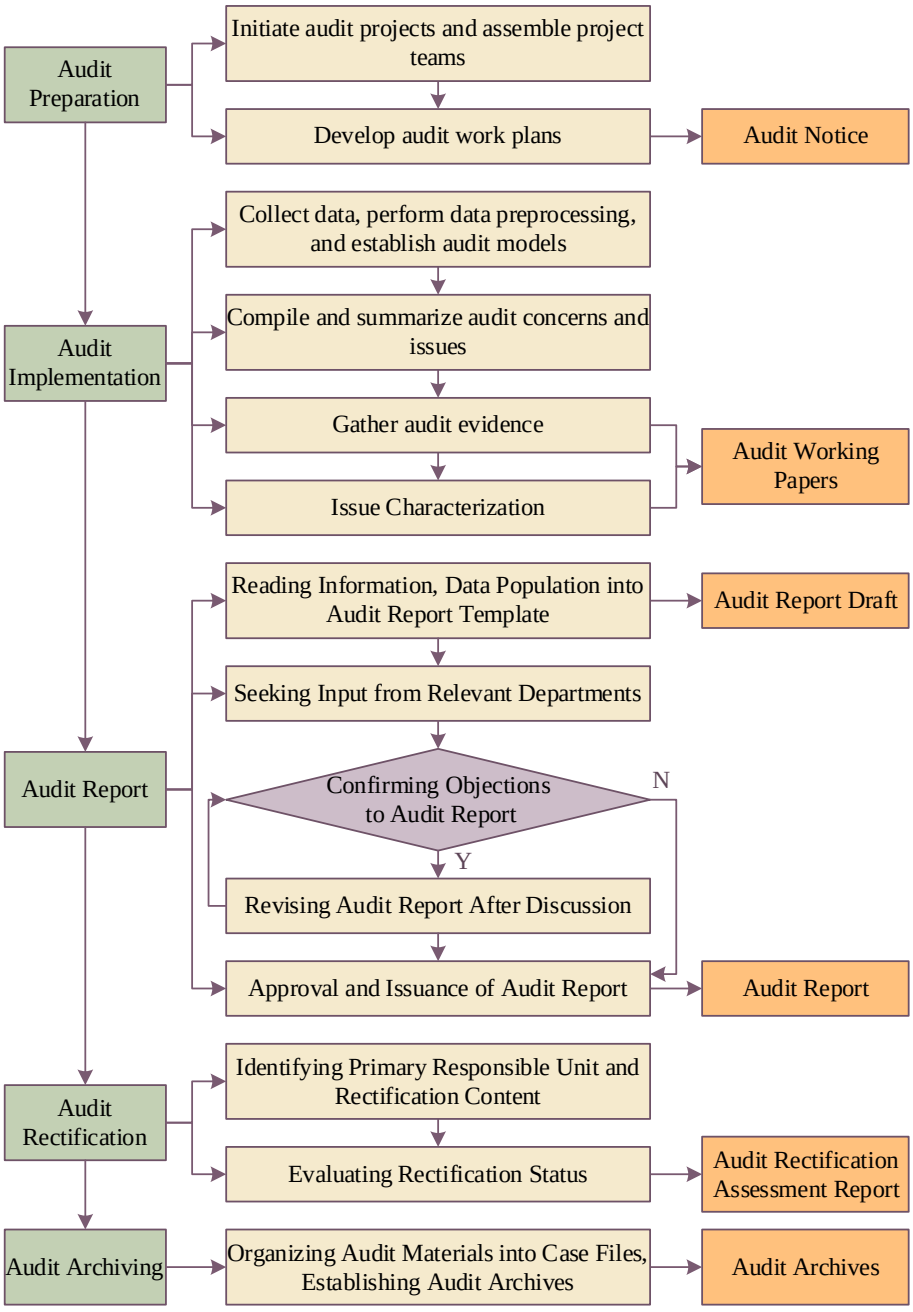

Figure 3: Intelligent audit operation process

2.2.3 Audit synergy system

Audit collaboration system supports data interoperability of OA, WeChat, Nail, online meeting platform and other systems to enhance the collaboration efficiency of various types of collaborative work scenarios. The audit collaboration system includes internal multi-departmental collaboration within the hospital as well as multi-governance body collaboration in the hospital. Through the audit collaborative governance, strengthen the effectiveness of audit governance, improve the operation and management of the hospital, can further improve its risk management and control capabilities, and better promote the implementation of the new health care reform to do a good job of system security.

2.3 Practical Path of Smart Audit Platform for Hospitals

In order to realize the concept and working methods of smart auditing, it is indispensable to build a smart audit platform. The accumulated experience of internal auditing and the strategic insights of the leadership are fully integrated into the top-level design, the development direction of smart auditing is clarified from the top down, and the feasibility of the digitization and smart support of internal auditing is fully assessed.

According to the needs of data model construction, analyze the data sources based on business process sorting and audit data requirements, put forward the audit database fetching requirements, extract data from the source system layer such as the database, internal system, and manually entered data, and through data cleaning, data analysis, and data integration, establish a unified data model layer, so as to gradually form the data aggregation layer, realize data structuring, and build the hospital's audit Database.

Internal audit extends from the basic functions of supervision and evaluation to the expanded functions of governance and counseling. The science and timeliness of intelligent auditing are fully utilized in order to achieve effective control of auditing costs, strengthen the role of auditing as a pre-warning and preventive measure, and provide reference for improving the quality of hospital medical services.

The audit team is an important part of realizing the transformation of digital and intelligent auditing, and the implementation of intelligent auditing requires the strong support of a professional audit team. However, the demand for smart auditing is often mismatched with the ability of auditors. Smart auditing not only requires auditors to have basic logical thinking, structured thinking, communication skills, management skills and other general capabilities, but also additional capabilities such as data mining skills, risk prediction skills and business skills.

3 Methodology for assessing the impact of big data auditing on internal control in hospitals

3.1 Lasso-Logistic prediction modeling

3.1.1 Main Ideas of Lasso-Logistic Modeling

Lasso is a new method of compression estimation proposed in the field of regression, this method is in the least squares method, by setting the L1 regular term, so that the sum of the coefficients of the variables must be less than a certain value, so as to compress the coefficients of some variables to 0. The variables that are retained are the key variables screened out by the model, and the variables that are excluded are those that the model considers to be unimportant. When the number of variables involved in the model is too large, the correlation between the variables is strong, and it is easy to produce the problem of overfitting, as can be seen from the characteristics of Lasso, the method is more suitable for dealing with high-dimensional data, so that it can achieve the effect of screening variables. The final model established is more refined and interpretable than it would be. The model can handle both discrete and continuous variables. Before using LASSO, it is better to process the original data, otherwise, because of the large difference in the range of values of the variables, the results of variable screening will be inaccurate.

$$\arg\min\left\{\sum_{i=1}^{n}\left(y_i-\alpha-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \tag{1}$$

y_i denotes the ith dependent variable, β_j is the coefficient of the independent variable, λ is the reconciliation parameter, used to adjust the complexity of the model, the reconciliation parameter is too large λ, the penalty is increased, which may make some effective variables are excluded; the reconciliation parameter λ is too small, the penalty is weakened, and all the variables are retained. information (similar to least squares), the model becomes more complex and the possibility of overfitting is very high. Thus, it can be found that the key of Lasso model lies in the selection of the reconciliation parameter.

3.1.2 Logistic model

Logistic model is a non-linear model that is more flexible compared to traditional linear models. Trained by given samples, the output of the model is according to the threshold rule, and the final outputs are 0 and 1, thus it is commonly used in two-class problems. In this paper, we study the internal control quality of hospitals under big data auditing, and categorize the dependent variable into hospital internal control quality compliance and hospital internal control quality noncompliance. $Y=1$ denotes hospital internal control quality compliance and $Y=0$ denotes hospital internal control quality noncompliance. P denotes the probability of $Y=1$ under the action of N independent variables, then the logistic regression model is established as:

$$\ln\left(\frac{P}{1-P}\right)=Z=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots\cdots+\beta_nX_n \tag{2}$$

By parameter estimation, the parameter estimates of $\beta_j\left(j=0,1,2,3,4,\cdots,n\right)$ can be obtained, and for the variable observations that are $X_1,X_2,X_3,X_4\cdots\cdots X_n$ customers, one can predict the probability that $Y=1$, i.e:

$$\begin{aligned}P\left(Y=1\right)&=\frac{\exp\left(Z\right)}{1+\exp\left(Z\right)}\\&=\frac{\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_nX_n\right)}{1+\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_nX_n\right)}\end{aligned} \tag{3}$$

In addition, the probability of non-compliance with the quality of the hospital's internal controls is:

$$\begin{aligned}1-P\left(Y=1\right)&=\frac{1}{1+\exp\left(Z\right)}\\&=\frac{1}{1+\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots\cdots+\beta_nX_n\right)}\end{aligned} \tag{4}$$

The point of note for this model is that it is better not to use least squares for the loss function

of the Logistic regression model, because the non-convex function obtained in this way makes it difficult to find a globally optimal solution. For the estimation of the parameters of the Logistic model, the loss function obtained by maximum likelihood can be used, which can quickly estimate the coefficients of the variables.

3.1.3 Lasso-Logistic Models

Lasso-Logistic regression model is with the help of Lasso's idea to achieve the effect of eliminating model variables by setting penalty parameters. The Lasso-Logistic model established after variable screening retains the most critical variable information and meets the requirements of this paper.

The parameter estimates in the Lasso-Logistic regression model can be expressed as:

$$Min\left\{\sum_{i=1}^{n}\left[y_i\left(\beta_0+\beta^T x_i\right)-\ln\left(1+e^{\beta_0+\beta^T x_i}\right)\right]-\lambda\sum_{i=1}^{n}\left|\beta_i\right|\right\} \tag{5}$$

The value of the reconciliation parameter λ directly affects the model's filtering of variables. If the value of the parameter is too small, the model will contain unimportant information, and if the value of the parameter is too large, the model will exclude too much critical information. In this paper, the ten-fold cross-validation method is used to determine the reconciliation parameter λ. The idea of the ten-fold worse validation method is to divide the sample into 10 parts, take turns to use the data of 9 of them to build a model, and use the built model to predict the data of the remaining 1 part of the test.

Assume that the model with combinations and reconciliation parameters is represented as $f^k\left(x,\lambda\right)$, defined:

$$CV\left(f,v\right)=\frac{1}{N}\sum_{i=1}^{n}L\left(y_i,f^{k(i)}\left(x_i,\lambda\right)\right) \tag{6}$$

Then $CV\left(f,v\right)$ is a test error curve associated with λ, and finding the λ that minimizes the test error enables the construction of a Lasso-Logistic model.

Reconciliation parameters of the regression model:

$$\lambda=\arg\min CV\left(f,v\right) \tag{7}$$

The final variables retained after the above process, i.e., the key variables for model screening, are the finalized Lasso-Logistic regression equations:

$$\log\left(\frac{P}{1-P}\right)=\beta_0+\sum_{i=1}^{n}\left[\beta_i X_i\right]=X^T\beta^{Lasso} \tag{8}$$

3.2 SHAP-based feature volume screening approach

3.2.1 Basic overview of SHAP

SHAP is a solution to the cooperative game, and the SHAP values are applicable to a wide range of interpretable models for categorical and regression models. All characteristic variables of the model inputs to be interpreted can be regarded as “contributors” to the model outputs. The principles are described in the following sections.

SHAP is currently used in automation, economy, finance, medical care, transportation and other fields, the method can effectively analyze the relationship between variables and results between systems, but also allows the relevant personnel to intuitively understand the judgment of the relevant information, and provides an effective reference for the subsequent judgment of the improvement of the basis.

Some algorithms have been applied to provide partial dependency graphs to explain some of the results; however, their analyses are still not explanatory enough and cannot be used to explain all machine learning models. Therefore, SHAP is used to complement the machine learning models and to make the analysis of results more comprehensive and effective.

3.2.2 Basic Principles of SHAP

The traditional method can only provide information about the order of importance of the features, but not the way the features affect the target features.The LGBM model consists of N trees and is a “black-box” model, so for a specific sample, it is unknown how its data affect the L index. In specific applications, we still hope to clarify the relationship between the eigenvalues and the results, and we also hope to analyze the results from two perspectives: the eigenvalues and the final results.

SHAP is a method based on game theory to explain the output of machine learning models, the core of which is: Shapley value, is an additive interpretation model. In this paper, the SHAP algorithm is used to explain and analyze the factors affecting the L indicator in the model to increase the interpretability of the model.

The role of the SHAP value is to quantify the influence (contribution) of each feature in the sample to the final output of the prediction model. Its simple common understanding is: for example, a job is done by more than one person, and how much each person does is not the same, in the distribution of wages, trying to be fair and reasonable distribution, at this time, based on the SHAP value to calculate the amount of money allocated to each person, i.e.: the degree of each person's contribution to the work.

In practice, the output values are attributed to the SHAP value of each feature value and used to measure its impact.SHAP can perform the following tasks: debugging the model, doing characterization work, guiding the direction of data collection, and guiding decision making.

Assuming the i th sample x_i and its k th feature value is x_{ik}, the marginal contribution of the feature is m_{ij}, the weights of the edges are w_i, and the SHAP value of the feature is $f\left(x_{ik}\right)$, then the contribution to the predicted value y_i from the k th feature of the i th sample is value, i.e., the SHAP value corresponding to that feature is:

$$f\left(x_{ik}\right) = m_{ik}w_1 + m_{ik}w_2 + \cdots + m_{ik}w_n \tag{9}$$

The baseline for the entire model, typically the mean of all sample target variables:

$$y_i = y_{base} + f\left(x_{i1}\right) + \cdots + f\left(x_{ik}\right) \tag{10}$$

And the SHAP value also satisfies equation (10). When the formula $f(x_{ik}) > 0$, it indicates that the feature plays a positive role in the prediction value, also a boosting role; on the contrary, it is the opposite, a lowering role. The biggest feature is: not only can reflect the influence of the feature input in each sample, but also can indicate the positive and negative influence.

4 Results of the assessment of the impact of big data audits on the quality of internal controls in hospitals

4.1 Results of the Lasso-Logistic Model Assessment

4.1.1 Construction of the model

In this paper, the relevant data of 60 hospitals from January 1, 2019 to December 31, 2023 were randomly selected from Chinese comprehensive tertiary-level hospitals as the study sample, and then 300 hospitals with internal control quality standards were selected as the control group in accordance with the same conditions in a ratio of 1:5. The sample of 360 hospitals used was approximately the same in terms of size and industry. All data and hospital information are obtained from the financial information disclosed by the hospitals to ensure that all data are true and reliable.

The selected model indicators need to be able to objectively respond to the changes in the quality of internal control in hospitals, and if the selected indicators are non-financial, their authenticity and validity cannot be guaranteed. Therefore, the indicators selected in this paper try to reflect the financial indicators of the quality of internal control in hospitals. After many aspects of reviewing and researching, this paper adopts four aspects of financial compliance, operational efficiency, sustainable development ability and medical quality and safety, and 17 indicators are used to react to the changes in the quality of internal control of hospitals under the big data audit. At the same time, these indicators can be obtained from real and reliable ways, which are used to ensure the scientificity and accuracy of this experiment. The specific indicators included in the variables are shown in Table 1.

Table 1: Variable meaning

Index reaction	Concrete index	Referencing symbol
Financial compliance	Asset ratio	A1
	Mobility ratio	A2
	Health receivable turnover	A3
	Budget execution deviation rate	A4
	The conversion rate of the drug and materials warehouse	A5
	Government procurement compliance rate	A6
Operating efficiency	Average residence	A7
	Bed rate	A8
	The outpatient rate is compared to the doctor	A9
	Large equipment utilization	A10
Sustainable development ability	The funding ratio of scientific research projects	A11
	Medical staff continued the education participation rate	A12
	Patient satisfaction score	A13
Medical quality and safety	Incidence of hospital infection	A14
	Prescription rate	A15
	Incidence of medical disputes	A16
	Full rate of electrical record	A17

4.1.2 Lasso-Logistic Regression Analysis

Lasso-Logistic regression analysis mainly uses the Glmnet package of the R program to construct

the model. This method solves the shortcoming of Logistic regression that cannot screen the factors, and adds an $L1$ penalty number to Logistic regression to eliminate redundant variables.Lasso-Logistic regression analysis needs to include all the dependent and 17 independent variables in the database, and get the model parameters and reconciliation parameters through the sample set. Finally, the predictions were obtained through SPSS.

Figure 4 shows the change in model AUC for different values of λ. As the reconciliation parameters change, the top of the figure indicates the number of variables retained in the corresponding model. It can be seen from the figure that the number of independent variables selected by the model changes as the value of the reconciliation parameter λ is taken, the larger λ is, the greater the degree of compression the smaller the number of independent variables finally selected, and the independent variables that are not selected to enter the model are eliminated. It can be seen that the AUC is maximum when compressed to 14 variables.

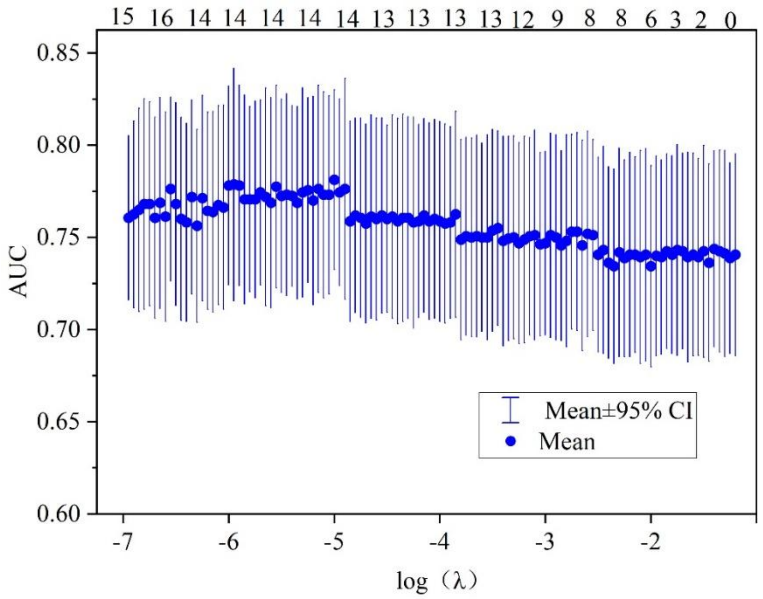

Figure 4: AUC changes in different values

Meanwhile, through the program calculation, the compression degree of each variable under different penalty parameter lambda can be obtained, so as to further discriminate the effective influencing factors with the practical background. For the given lambda values, cross-validation is carried out, and the lambda value with the smallest cross-validation error is selected, and the results of parameter estimation of the Lasso-Logistic model are shown in Table 2. It can be seen that the model with 14 independent variables is obtained.

Table 2: The Lasso-Logistic model parameter estimation results

Var_names	Coef	Expcoef
Intercept	-5.056	15.675
A12	-1.855	0.154
A16	-8.605	0.421
A10	-6.362	0.936
A11	-5.257	0.945
A2	-1.692	0.986
A4	-4.103	0.997
A3	1.103	1.000
A15	9.335	1.000
A6	1.726	1.000
A14	2.905	1.002
A9	6.711	1.005
A1	2.645	1.025
A17	8.977	1.094
A13	2.534	1.265

Using the fourteen indicators mentioned above as independent variables, logistic regression analysis was performed using SPSS 21.0 software on 360 training samples of credit risk related indicators, and finally the backward wald method was used for the dependent variable, i.e., whether there is a substandard quality of internal control in the hospital. In constructing the model, the results were homogeneous in and out, and the final risk factor was taken as P-value < 0.05.Nagelkerke is the result of further adjustment of Cox&Snell and -2$\hat{LL}s$, which is more intuitive and quasi-Nagelkerke takes the value interval of (0, 1). Its closer to 1 the better the goodness of fit. The value of Nagelkerke for the logistic regression model in this paper is 0.554, which is greater than 0.5 and has a better fitting effect.The Sig. value obtained from the H-L test is 0.859, with a significance greater than 0.05, which indicates that there is no significant difference between the fitted values of the grouping test and the observed values, which means that the model results fit the sample data more accurately.

By using the stepwise backward wald method and eliminating insignificant variables, the variables in the equation are shown in Table 3, and only six variables and one constant were finally entered into the model, and the variables include the percentage of research project funding, the incidence of medical disputes, the ratio of outpatient visits to doctors, the utilization rate of large equipment, the turnover rate, and the rate of qualified prescriptions, and with P<0.05, the variables are significantly related to the results.

Table 3: Variable in equation

	B	S.E	Wals	df	Sig.	Exp(B)	EXP(B)95%C.I	
							Lower limit	Upper limit
A11	-0.056	0.017	10.905	1	0.022	0.945	0.911	0.975
A16	-0.073	0.011	36.258	1	0.043	0.927	0.905	0.952
A9	-2.435	0.546	19.694	1	0.048	0.085	0.032	0.256
A10	3.412	0.855	15.944	1	0.006	30.304	5.684	161.521
A2	-2.056	0.378	29.403	1	0.033	0.126	0.062	0.265
A15	1.011	0.402	6.406	1	0.035	2.755	1.255	6.032
Constant	7.258	1.182	37.724	1	0.000	0.103		

4.1.3 Testing of model results

The relevant results predicted by the Lasso-Logistic model are shown in Table 4, which shows that the overall prediction for the 360 hospitals selected in the model sample is better, basically predicting most of the hospitals. The highest discrimination rate of 94% was found for 300 hospitals with internal control quality standards. Meanwhile, the discrimination of 60 hospitals whose internal control quality does not meet the standard is 90% correct, which can be seen that the model is more effective in identifying the problem of whether the quality of the internal control of hospitals is compliant. The statistical results of the Lasso-Logistic regression empirical model in this paper have high accuracy and robustness, and can explain the relevant coefficients of the hospitals accordingly, and this model imports the relevant coefficients into it, and carries out calculations and analyses, so as to derive the probability of the internal control quality of the hospitals failing to meet the standard.

Table 4: Model prediction

Iteration number	Observed		Predictive value		
			The internal control quality is standard		Correct percentage/%
			NO	YES	
1	The internal control quality is standard	NO	282	18	94
		YES	6	54	90
2	The internal control quality is standard	NO	282	18	94
		YES	6	54	90
3	The internal control quality is standard	NO	281	19	93.7
		YES	7	53	88.3
4	The internal control quality is standard	NO	282	18	94
		YES	6	54	90
5	The internal control quality is standard	NO	281	19	93.7
		YES	7	53	88.3
6	The internal control quality is standard	NO	282	18	94
		YES	6	54	90
7	The internal control quality is standard	NO	281	19	93.7
		YES	7	53	88.3
8	The internal control quality is standard	NO	282	18	94
		YES	6	54	90
9	The internal control quality is standard	NO	281	19	93.7
		YES	7	53	88.3
10	The internal control quality is standard	NO	282	18	94
		YES	6	54	90

4.2 SHAP-based feature volume screening analysis

4.2.1 Characteristic importance analysis results

The feature attribution methods Gain, Cover and Weight of the Lasso-Logistic model all belong to the global attribution method, reflecting the changes in the expected accuracy of the model when a set of features is deleted, i.e., the effect of the features on the global accuracy. Among them, Gain reflects the effect of feature variables on the accuracy, measured by the degree of improvement of model accuracy after the features are brought into the model; Cover reflects the degree of coverage of the feature variables on the observation objects, measured by the

number of observation objects related to the features; Weight reflects the frequency of the feature variables being used in the model. , Weight and Cover values are shown in Table 5. Among them, A15 has the largest Weight value of 0.083, which indicates the high frequency of prescription compliance rate being used in the model.

Table 5: Gain, Weight and Cover

Characteristic variable	Gain	Weight	Cover
A10	0.206	0.075	0.032
A11	0.083	0.002	0.061
A2	0.076	0.031	0.051
A3	0.072	0.017	0.053
A16	0.053	0.039	0.059
A9	0.054	0.054	0.023
A17	0.044	0.069	0.039
A15	0.04	0.083	0.033
A13	0.049	0.024	0.055
A1	0.039	0.080	0.034
A12	0.043	0.036	0.035
A4	0.034	0.082	0.028
A14	0.023	0.034	0.045
A6	0.029	0.038	0.033

Figure 5 shows the SHAP impact of all feature values of all samples in the training set, each row in the figure represents a feature, the horizontal axis is the SHAP value, where each point represents a sample, the darker the color indicates that the feature itself is larger in value, the lighter the color indicates that the feature itself is smaller in value. Figure 6 shows the mean of the absolute values of the SHAP values of each feature as the importance of the feature. We found that the top three features with the largest SHAP values for the contribution of the Lasso-Logistic model to the prediction results are A10, A15, and A1, i.e., large equipment utilization rate, prescription compliance rate, and gearing ratio.

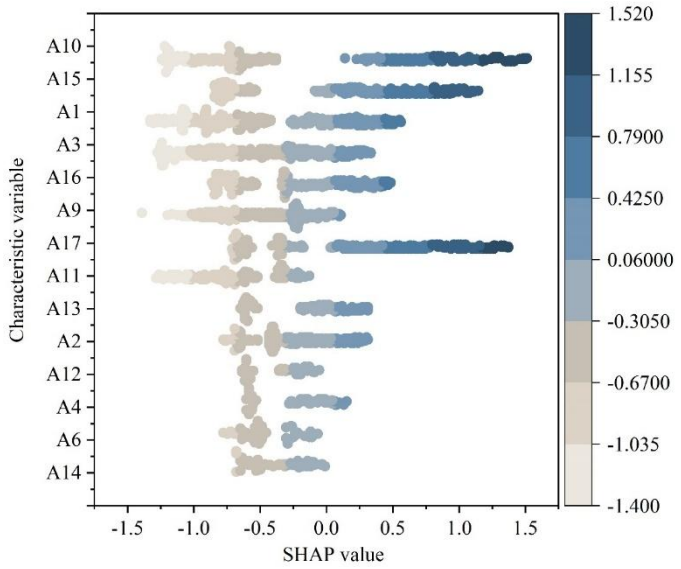

Figure 5: The SHAP impact of all eigenvalues

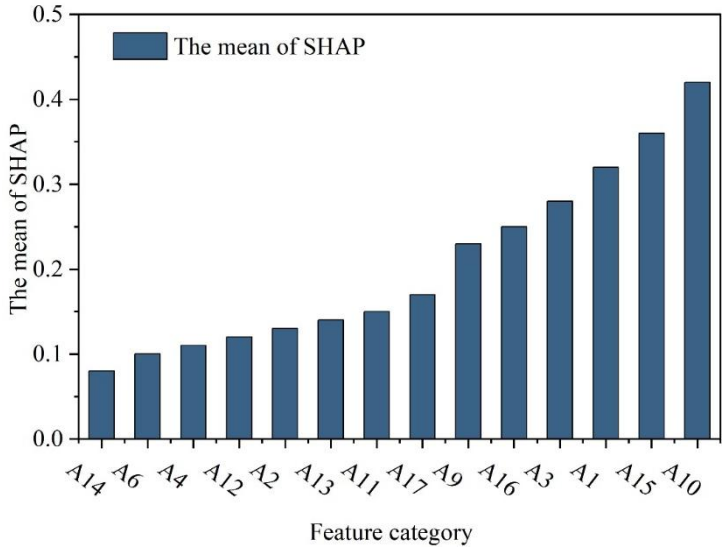

Figure 6: Model characteristics SHAP global mean

4.2.2 Prediction of the quality of internal control by significant characteristics

In an analysis of the relative importance of internal control quality in hospitals under big data auditing, this paper finds that large equipment utilization and prescription compliance rate are the most important predictors of internal control quality characteristics in hospitals under big data auditing. So what exactly is the association between these two characteristics and hospital internal control quality? To answer this question, partial dependency graphs are used in this paper to examine the predictive patterns of large equipment utilization and prescription compliance rate for hospital internal control quality. Figure 7 shows the prediction of large equipment utilization on the quality of hospital internal control under big data auditing, with the horizontal axis representing large equipment utilization and the vertical axis representing the quality of hospital internal control. Overall, the quality of internal control in hospitals shows an upward trend as the utilization rate of large equipment increases. When the utilization rate of large equipment in hospitals under big data auditing is around 30%, the quality of internal control in hospitals will be significantly improved, and when the utilization rate of large equipment reaches 83%, the quality of internal control in hospitals will not be significantly improved even if the utilization rate of large equipment is increased again.

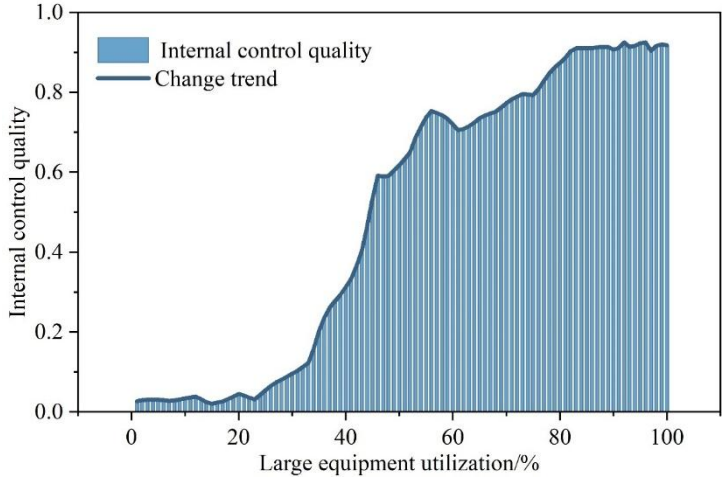

Figure 7: The prediction of the quality of internal control of hospitals

Figure 8 shows the relationship curve between the prescription compliance rate of hospitals under big data auditing and the quality of internal controls in hospitals. The prescription pass rate of hospitals under big data audit also has a significant nonlinear predictive effect on the quality of internal control of hospitals, in general, as the prescription pass rate of hospitals under big data audit grows, the quality of internal control of hospitals becomes higher, and when the prescription pass rate of the quality of internal control of hospitals under big data audit is around 60%, the quality of internal control of hospitals improves significantly, and the When the per capita age of hospitals' internal control quality under big data audits exceeds 78%, the quality of hospitals' internal controls does not become significantly higher, even if the general prescription pass rate of hospitals' internal control quality under big data audits increases further.

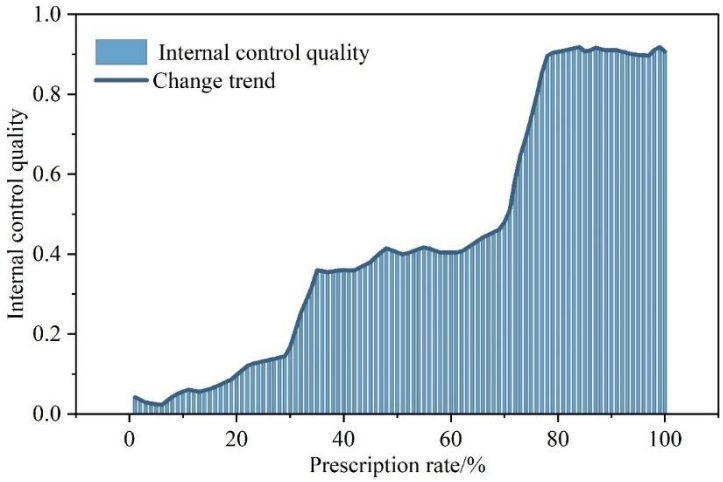

Figure 8: Hospital prescription rate and internal control quality

5 Conclusion

This paper adopts an integrated approach of machine learning to mine the association between multidimensional features of hospital internal control under big data auditing and the quality of hospital internal control using a Lasso-Logistic integrated learning model. The article takes comprehensive Tier 3A hospitals in China from 2019-2023 as a research sample, and uses the Lasso-Logistic model to assess the predictive ability of the features of hospital internal control quality under big data auditing on the quality of hospital internal control, based on which, the feature importance and partial dependency graphs are utilized to analyze the feature prediction patterns of which have a greater predictive ability on the quality of hospital internal control.

From the results of Lasso-Logistic regression, the variables that enter the metric model are the percentage of research project funding, the incidence of medical disputes, the ratio of outpatient visits to doctors, the utilization rate of large equipment, the turnover rate, and the prescription pass rate. In these data, the current ratio reflects the financial compliance of the hospital, the ratio of outpatient visits to doctors and the utilization rate of large equipment maps the operational efficiency of the hospital, the sustainability of the hospital is presented through the ratio of funding for scientific research projects, and the quality and safety of medical care is shown through the prescription compliance rate. Inputting the relevant data indicators into the model reveals that the highest discrimination rate for both hospitals that meet and do not meet the internal control quality standards reaches over 90%, with high accuracy.

In order to further enhance the interpretability of the model, this study visualizes the contribution of each characteristic variable in the Lasso-Logistic model based on the SHAP method, which realizes the global attribution explanation and enhances the interpretability and

prediction ability of the Lasso-Logistic model.

In the context of big data auditing, there is a nonlinear correlation between the hospital's large equipment utilization rate and prescription compliance rate with the quality of internal control. This finding not only advances the research related to the relationship between hospital operational characteristics and internal control quality in the context of big data auditing at the academic level, but also provides practical references for hospitals to design in terms of audit strategies.

Funding

This research was supported by the Shandong Provincial Medical and Health Science and Technology Project: Design and Practice Path of Intelligent Internal Control Mechanism for Research Funds under the Background of Digital Transformation (No.: 202515020140). Source: Shandong Provincial Health Commission.

References

[1] Yin, H., Akmandor, A. O., Mosenia, A., & Niraj K, J. (2018). Smart healthcare. Foundatiosn and Trends in Electronic Design Automation, 12(4), 401-466.

[2] Tian, S., Yang, W., Le Grange, J. M., Wang, P., Huang, W., & Ye, Z. (2019). Smart healthcare: making medical care more intelligent. Global Health Journal, 3(3), 62-65.

[3] Mao, Y., & Zhang, L. (2021). Optimization of the medical service consultation system based on the artificial intelligence of the internet of things. IEEE Access, 9, 98261-98274.

[4] Zhang, G. W., Gong, M., Li, H. J., Wang, S., & Gong, D. X. (2023). The “Trinity” smart hospital construction policy promotes the development of hospitals and health management in China. Frontiers in public health, 11, 1219407.

[5] Guo, J., & Sun, S. (2024, July). Research on intelligent healthcare scenario construction of" one hospital for the whole city". In Third International Conference on Biomedical and Intelligent Systems (IC-BIS 2024) (Vol. 13208, pp. 665-676). SPIE.

[6] Xu, K. (2025). Applications and Challenges of Big Data Auditing in the Health Insurance Industry. Academic Journal of Business & Management, 7(4), 173-181.

[7] Dako, O. F., Onalaja, T. A., Nwachukwu, P. S., Bankole, F. A., & Lateefat, T. (2020). Big data analytics improving audit quality, providing deeper financial insights, and strengthening compliance reliability. Journal of Frontiers in Multidisciplinary Research, 1(2), 64-80.

[8] Zhao, H., & Wang, Y. (2023). A big data-driven financial auditing method using convolution neural network. Ieee Access, 11, 41492-41502.

[9] Mohammed Ismail, I. H., & Abdul Hamid, F. Z. (2024). A systematic literature review of the role of big data analysis in financial auditing. Management & Accounting Review (MAR), 23(2), 321-350.

[10] Rakipi, R., De Santis, F., & D'Onza, G. (2021). Correlates of the internal audit function's use of data analytics in the big data era: Global evidence. Journal of International Accounting, Auditing and Taxation, 42, 100357.

[11] Kaya, I., Akbulut, D. H., & Ozoner, K. (2018). Big data analytics in internal audit. PressAcademia Procedia, 7(1), 260-262.

[12] Chen, F. H., Hsu, M. F., & Hu, K. H. (2022). Enterprise's internal control for knowledge discovery in a big data environment by an integrated hybrid model. Information Technology and Management, 23(3), 213-231.

[13] Elmashtawy, A., & Salaheldeen, M. (2022, September). Big data techniques and internal control: evidence from Egypt. In International Conference on Emerging Technologies and Intelligent Systems (pp. 14-23). Cham: Springer International Publishing.