Key Factor Analysis of Digital Health Ability in the Elderly Using SVM and Shapley Algorithm

Yi Ren #, Yihan Zhao *, #, Hanbo Min #, Siyu Luo #, Muyang Liu #

College of Science, Minzu University of China, Beijing, China, 100081

* Corresponding Author Email: Zhaoyihan4012@163.com

#These authors contributed equally.

Abstract. This research centers on the analysis of critical factors influencing the digital health ability of older adults in a city during the post-pandemic period, addressing the current situation of low digital health proficiency and a pronounced digital divide in a specific country by developing a comprehensive evaluation framework. This study employed a SVM model with K-fold cross-validation for optimization and utilized the Shapley algorithm to quantitatively assess the contribution of critical factors to digital health ability. The findings indicate that technology anxiety, community support, and mental health are the primary factors affecting the digital health ability of older adults, with technology anxiety exerting a pronounced negative influence. The optimized SVM model showed substantial improvements over traditional stepwise regression models regarding predictive accuracy and robustness, highlighting its strengths in capturing nonlinear relationships. This research offers a scientific foundation for policymaking, stressing the importance of alleviating technology anxiety, strengthening community support, and enhancing mental health to narrow the digital divide among older adults and elevate their digital health literacy comprehensively. Furthermore, the approach and findings of this study are applicable to other areas of health management and digital literacy research, offering robust support for the advancement of healthy aging and the realization of social equity.

Keywords: Digital Healthcare, Elderly, Key Factors, SVM Shapley Value.

1. Introduction

With the accelerated process of global aging and the rapid development of digital technologies, improving the digital health literacy of the elderly has become a focal point of societal concern. Digital health literacy refers to the ability of older adults to access, understand, and apply health information in a digital environment. This capability significantly influences their efficiency in health management and overall quality of life. However, due to generally low digital literacy, there exists a significant disparity between urban and rural elderly populations, with rural older adults facing a more severe digital divide [1] [2] [3]. In the digital era, many elderly individuals, lacking technological knowledge and adaptability, have gradually become "disadvantaged groups in the digital society" [4]. Consequently, empowering the elderly through digital technologies to enhance their health management capabilities has become an urgent social issue.

In recent years, scholars have conducted extensive research on the factors influencing digital health literacy among older adults and strategies for its improvement, focusing on aspects such as digital literacy, technology anxiety, social support, and the digital divide. Studies have shown that digital literacy has a significant impact on improving the health of the elderly, especially in rural areas, among individuals aged 65 and above, and within female groups [1]. Social support has been found to alleviate technology anxiety and enhance digital literacy [5], while age-friendly designs can effectively lower technological barriers [5]. Furthermore, unequal digital resource distribution and insufficient social support have been identified as major obstacles contributing to the digital divide [4], with the issue being particularly pronounced among rural left-behind elderly populations, where the widening digital divide exacerbates health inequalities [3]. While these studies provide theoretical support, they often focus on single-dimensional analyses and lack in-depth exploration of the complex interactions and nonlinear relationships among multiple factors.

Existing studies have made progress in defining digital health literacy for the elderly and offering policy recommendations, but they are largely constrained by traditional linear models, which fail to fully capture the complex interactions between multiple factors. This limitation is particularly evident in the study of dynamic psychological factors, such as technology anxiety and privacy concerns, and their regional differences [3] [6] [7]. Moreover, there has been limited research on the quantitative analysis and modeling of digital health literacy, and the lack of advanced machine learning methods has hindered the precision and scientific validity of policy recommendations [8].

To fill the research gap, this paper innovatively combines the SVM model optimized by K-fold cross-validation with the Shapley value algorithm, and for the first time applies it to quantify the key factors affecting the digital health capabilities of the elderly, providing a new method for quantitative analysis. Based on the research results, specific policy recommendations are put forward to alleviate technological anxiety, enhance community support, and improve mental health, providing a solid scientific basis for practical application.

The structure of the paper is organized as follows: Chapter 1 serves as the introduction, outlining the research background, current status, research gaps, and contributions of the study; Chapter 2 provides a literature review and theoretical framework, synthesizing relevant theories and advancements in digital health capabilities; Chapter 3 elaborates on the research design, covering data sources, model development, and optimization approaches; Chapter 4 focuses on empirical analysis, highlighting model results and key factor exploration; Chapter 5 presents conclusions and policy recommendations, summarizing the findings and envisioning future research avenues.

2. Relevant Theories

2.1. One-Hot Encoding

One-Hot Encoding is a widely used technique to transform categorical variables into numerical variables, commonly employed during the data preprocessing phase in machine learning models. The core principle involves converting each category into a distinct binary vector, with 0 and 1 denoting the presence or absence of a specific category. This removes ordinal relationships among categories, preventing algorithms from inferring magnitude or distance between them.

2.2. SVM

Support Vector Machine (SVM) is a commonly applied machine learning algorithm for classification tasks, effective in processing categorical data. The fundamental concept of SVM lies in constructing one or more hyperplanes within the data space to maximize the margin between classification boundaries and data points [8]. Table 1 contains explanations of the relevant concepts:

Table 1. Parameter Descriptions for the SVM Regression Model

Parameter Name	Detailed Explanation
Hyperplane	In three-dimensional space, a hyperplane acts as a plane that divides data points of distinct categories, positioning one category on one side and the other on the opposite side.
Margin	SVM classifies by identifying the hyperplane that maximizes the margin between data points of different categories and the hyperplane.
Support Vectors	The points nearest to the hyperplane are known as support vectors, which SVM utilizes to define the classification boundary.
Kernel Function	SVM leverages kernel functions to transform low-dimensional data into a higher-dimensional space to construct a linear classifier. Popular kernel functions include the linear kernel, RBF kernel, and polynomial kernel.

2.3. K-Fold Cross Validation

The core concept of K-Fold Cross-Validation is splitting the dataset into k equal-sized subsets, using k-1 subsets for training the model and the remaining subset for testing. This process is repeated k times, with each subset taking turns as the test set, while the others are used as the training set. Finally, the weights from the k validation rounds are aggregated to assess their consistency and stability across various training sets. If the projection directions and weights show minimal variation across validations, it suggests that the model is stable and the weights are reasonable.

2.4. Shapley Value

The Shapley algorithm, derived from game theory, focuses on fairly distributing the contribution of each participant to quantify their importance in cooperative scenarios. The algorithm is extensively applied in analyzing feature importance and explaining the decision-making processes in machine learning models. The Shapley algorithm excels in fully accounting for the marginal contributions of features in various combinations, offering theoretically robust quantifications of variable influence [9]. In the context of machine learning models, Shapley values assess feature importance by computing the average marginal contribution of each feature across various combinations. This approach addresses the constraints of single-feature importance assessments, enabling more precise analysis of feature interactions and their effects on model predictions [10]. This research employs the Shapley algorithm to measure the contribution of various influencing factors (secondary indicators) to older adults' digital health capabilities, enabling more accurate identification and explanation of each factor's impact.

3. Empirical Analysis

The objective of this study is to identify the critical factors affecting the digital health capabilities of older adults in Beijing and quantify their influence using an SVM model. The main objective of the experiment is to pinpoint the key variables impacting the digital health ability of older adults and assess their contributions to the model outcomes via the Shapley algorithm, offering a theoretical foundation for policy-making. The detailed research process is illustrated in Figure 1:

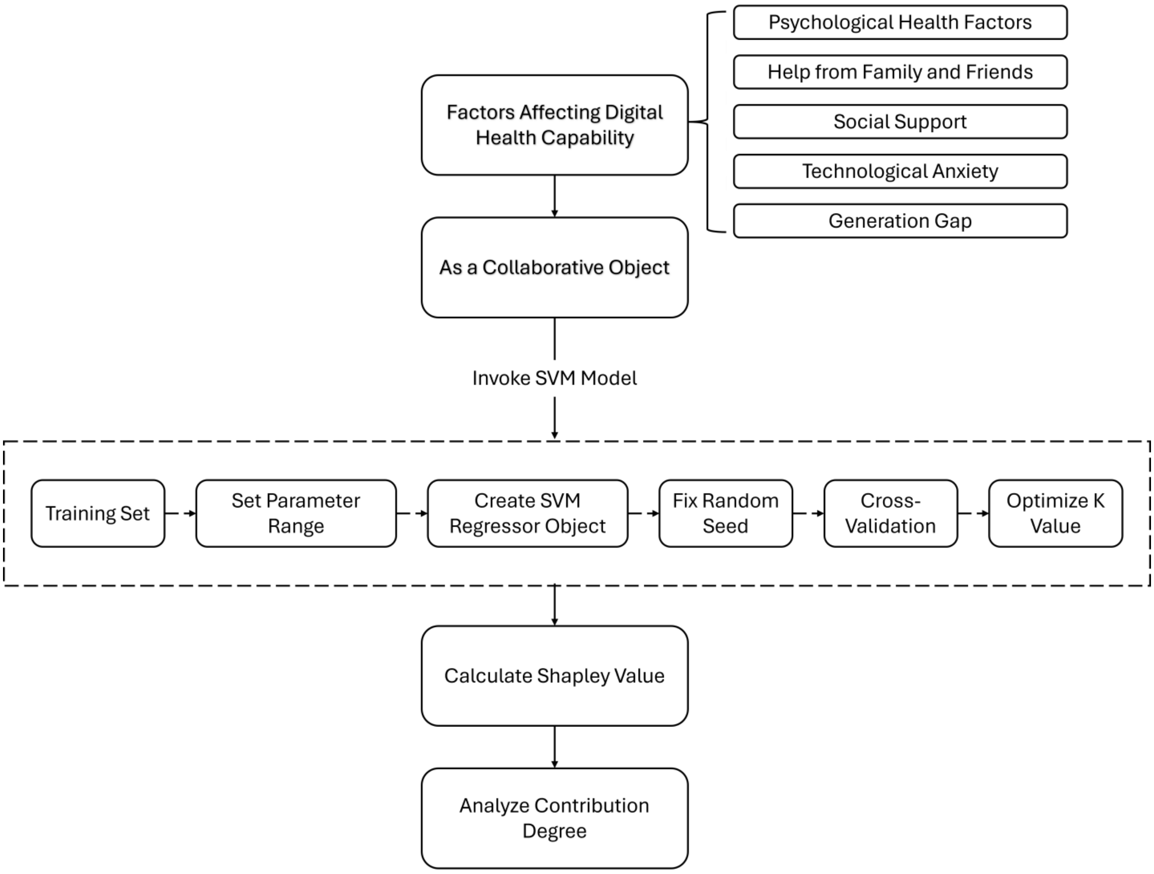

Figure 1. Research Workflow

3.1. Data Preprocessing

The data utilized in this study were obtained via questionnaires. The eHealth Literacy Scale was employed to assess the digital health capabilities of older adults in a specific region. The scale encompasses three dimensions: online health information and service application ability (F1~F5), valuation ability (F6), and decision-making ability (F7~F8), comprising 8 items scored on a 5-point Likert scale from "strongly disagree" (1) to "strongly agree" (5). The total score of the scale is between 8 and 40, where a higher score reflects greater digital health capability. Following data cleaning, the study examined the correlations among six highly dispersed influencing factors, with results presented in Figure 2.

After data cleaning, the correlations among six highly dispersed influencing factors were analyzed, and the results are shown in Figure 2.

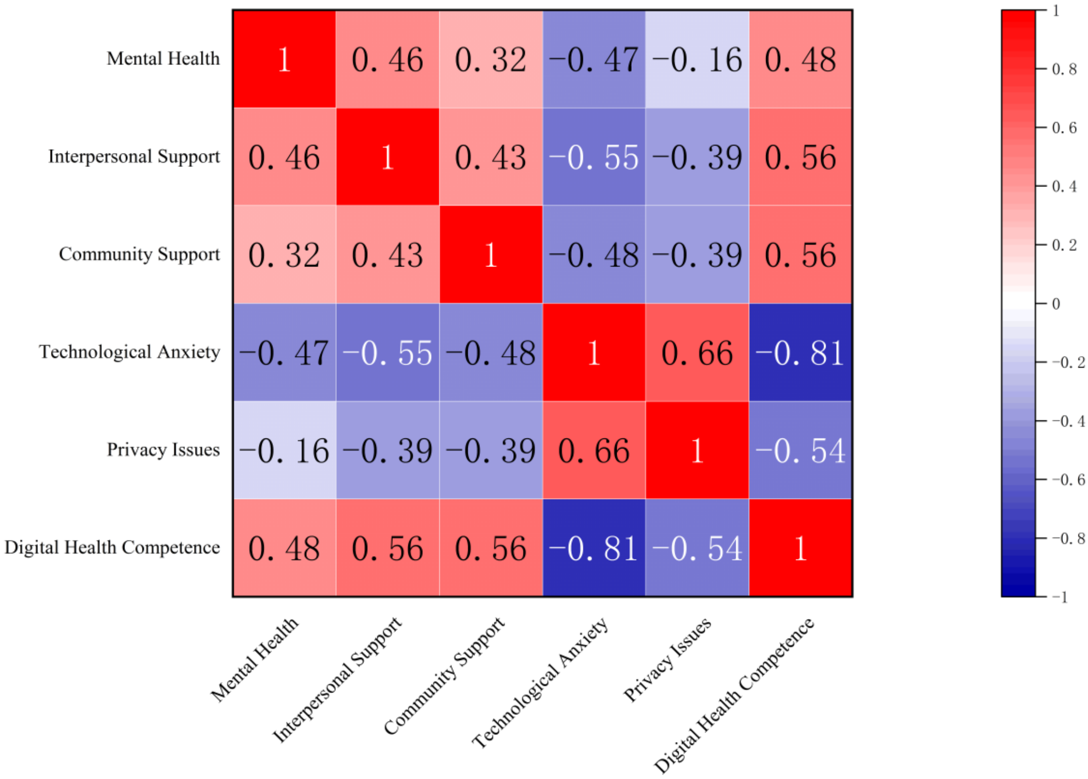

Figure 2. Heatmap of Key Factors

This research builds an evaluation framework for the digital health ability of older adults based on the social ecosystem theory. The framework establishes three primary dimensions: individual health, social support, and temporal environment, and seven secondary dimensions: demographics, physical health, mental health, interpersonal relationships, community support, technology anxiety, and privacy concerns.

This research utilized standardization and one-hot encoding techniques for data processing. For general numerical variables, standardization was applied to the raw data to remove the impact of dimensional inconsistencies and enable horizontal comparison among secondary indicators. For unordered categorical variables, one-hot encoding transformed them into numerical variables to prevent introducing false ordinal relationships to the algorithm, thereby improving model accuracy and retaining category independence. For ordered categorical variables, label encoding was employed to transform them into numerical variables, simplifying data representation and streamlining subsequent data processing and analysis.

3.2. Model Construction

In conventional K-fold cross-validation, this study aimed to optimize the choice of K and enhance model accuracy and stability by fixing a random seed, exploring the K value range (2 to 20), and applying grid search for parameter optimization. The detailed process involves performing 5-fold cross-validation for each K value, computing the mean squared error (MSE) and its standard deviation to evaluate model performance comprehensively. A fixed random seed was used throughout the experiment to ensure result reproducibility. The objective function assigns 70% weight to the mean MSE for evaluating the model's overall predictive accuracy and 30% weight to the MSE standard deviation to assess stability across data splits. The K value minimizing this function is chosen as the optimal K, balancing accuracy and stability.

In conclusion, the objective function for this study is defined as:

$$\hat{K} = min(0.7 \times MSE_{avg} + 0.3 \times MSE_{std}) \tag{1}$$

$$MSE_{avg} = \frac{1}{K}\sum_{i=1}^{K} MSE_i \tag{2}$$

$$MSE_{std} = \sqrt{\frac{1}{K}(MSE_K - MSE_{avg})^2} \tag{3}$$

In the aforementioned formula, $\widehat{K}$ denotes the optimal K value, MSE_{avg} refers to the average MSE, MSE_{std} indicates the standard deviation of the MSE, and MSE_K is the MSE for the Kth value.

This study focuses on selecting parameters for the SVM algorithm based on a Gaussian kernel function, using grid search to optimize kernel parameters and penalty coefficients, with a parameter range of 30 values logarithmically distributed between10^{-2} and 10^2. MATLAB code is used to implement the method, identify the optimal K value and model parameters, and validate the results, which are presented in Figure 3.

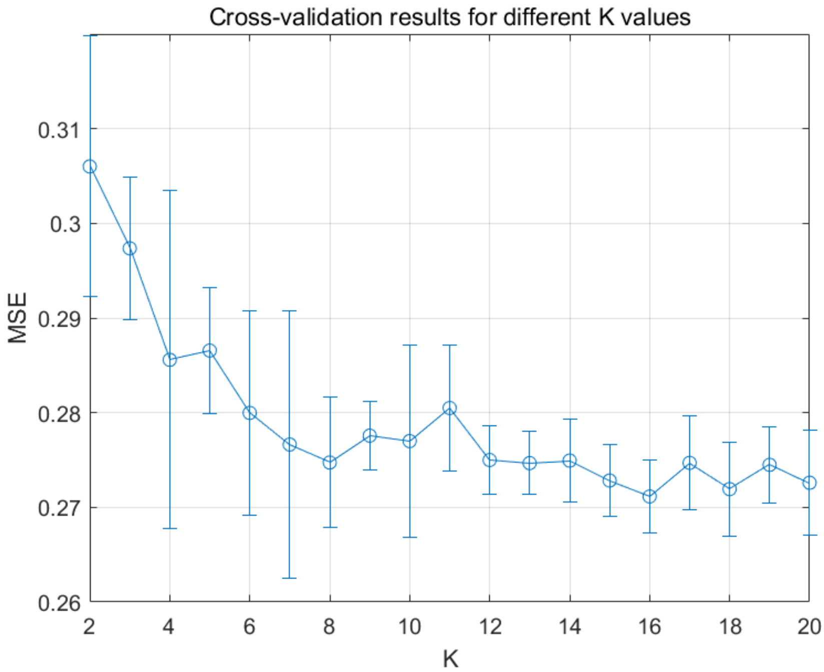

Figure 3. Results of Cross-Validation for Various K Values

From Figure 3, it is evident that when K=16, the average MSE is 0.2711, and the standard deviation of MSE is 0.0038. Compared with conventional fixed K values (e.g., 5 or 10), the optimized average MSE is significantly reduced, and the model shows enhanced robustness under different data splits, confirming that K=16 achieves the best accuracy and stability. Thus, this study adopts K=16 as the model parameter, splitting the dataset into 16 subsets, where 15 subsets are used for training, and 1 subset is allocated for testing.

Using the constructed optimal K-value SVM regression model, this study applies the Shapley value method to analyze feature contributions to digital health ability. The detailed steps are as follows: First, the various features of digital health ability (e.g., mental health condition, support from family and friends, technology anxiety) are defined as cooperative "players" in the Shapley algorithm. Next, the optimal K-value SVM regression model is applied, with the eHealth Literacy Scale score as the dependent variable and influencing factor scores (secondary indicators) as independent variables. Subsequently, the Shapley algorithm is used to compute the contributions of individual features to the digital health ability of older adults, yielding Shapley values. Finally, the results are analyzed to determine the influence of each feature on digital health ability, clarifying key factors affecting user willingness, such as whether perceived ease of use is a critical factor.

For the selection of the K value, this study evaluated candidate values within the range of 2 to 20, conducting 5-fold cross-validation and selecting the optimal K value based on a weighted objective function (70% weight for average MSE and 30% for MSE standard deviation). Furthermore, the grid search method was used to identify the optimal regularization parameter and kernel function scale, ensuring the model's optimal performance. The final outcomes are presented in Table 2.

Table 2. Results of Cross-Validation for Various K Values

Relevant Outcomes	Quantitative Data
Best K Value	16
Mean MSE	0.2711
Standard Deviation of MSE	0.0038
Best Regularization Parameter	0.6210
Best Kernel Function Scale	2.2122
Best Cross-Validation MSE	0.2696
Model Stability (Mean MSE)	0.1975
Model Stability (Standard Deviation of MSE)	0.0908
Feature Significance (Ranked by SHAP Values)	G,E,C,H,D

In the experiment, the best K value was 16, with an optimal regularization parameter of 0.6210 and a kernel function scale of 2.2122. With these settings, the average MSE was 0.2711, and the standard deviation was 0.0038, demonstrating satisfactory model fitting performance.

4. Results

Based on the Shapley calculation of the Shapley game model using the optimal K-value SVM, Figure 4 reveals that technology anxiety (G, feature 4) has the most significant impact on older adults' digital health ability, with a SHAP value of about 0.45, indicating that it substantially obstructs their adoption and use of digital technology. Community support (E, feature 3) ranks second, with a SHAP value around 0.25, positively influencing older adults by providing education and resources. Mental health (C, feature 1) ranks third with a SHAP value of 0.20, indicating that adaptability and a positive outlook support the improvement of digital health literacy. Privacy concerns (H, feature 5) and interpersonal support (D, feature 2) have smaller effects, with SHAP values of 0.15 and 0.10, respectively, suggesting that older adults prioritize technology usability and training resources over privacy and interpersonal support. Consequently, interventions should focus on mitigating technology anxiety, enhancing community support, and implementing mental health interventions to holistically improve older adults' digital health capabilities.

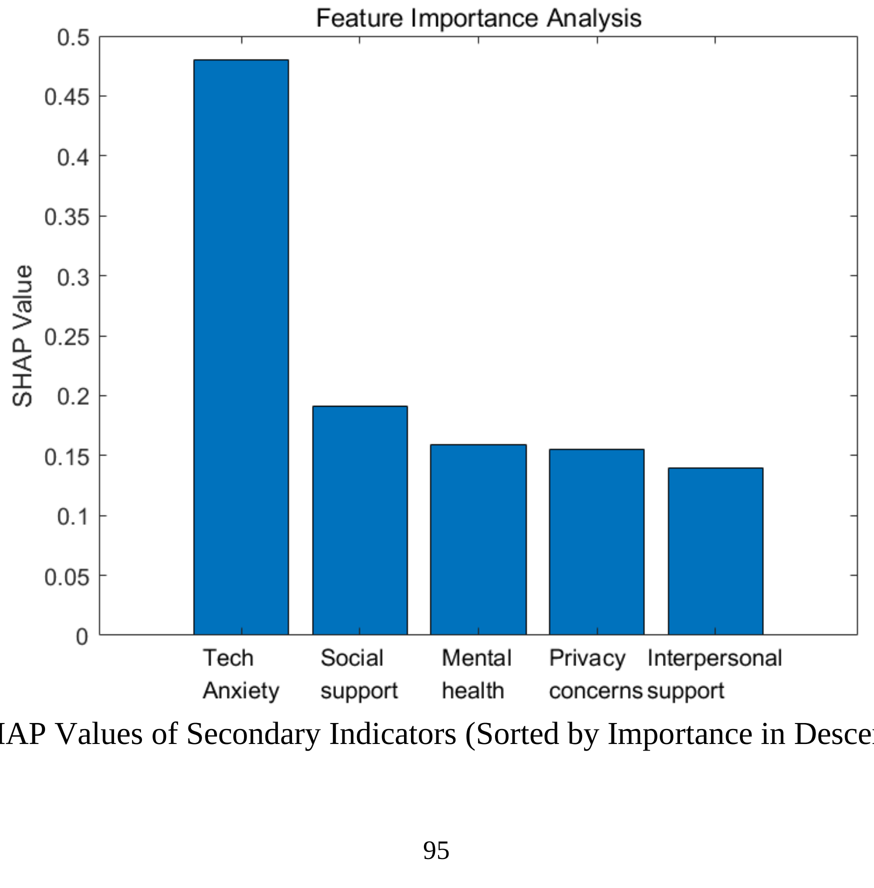

Figure 4. SHAP Values of Secondary Indicators (Sorted by Importance in Descending Order)

5. Conclusion

This study addresses the improvement of digital health capabilities among older adults in Beijing during the post-pandemic period, focusing on identifying critical influencing factors. The study employs an SVM model optimized through K-fold cross-validation, paired with the Shapley algorithm, to quantitatively assess the contributions of various factors to digital health capabilities. Findings indicate that technology anxiety is the primary barrier to improving digital health capabilities among older adults, whereas community support and mental health significantly contribute to enhancement. Additionally, the study validates the robustness of the SVM model in capturing intricate nonlinear relationships and multi-factor interactions, demonstrating superior predictive accuracy and stability compared to traditional stepwise regression models. The study's findings and analytical approach present novel insights into the quantification of influencing factors while providing clear policy recommendations, such as reducing technology anxiety, bolstering community support, and prioritizing mental health interventions. Future studies could extend to older populations across diverse regions and cultural settings, investigating additional dynamic factors (e.g., technological progress and habitual long-term use) affecting digital health capabilities. Moreover, the study's models and methodology can be adapted to other domains of health management and digital literacy research, offering holistic support for advancing health management and social equity among older adults.

References

[1] Yao Yao, Zhao Lindu. The Influence of Digital Literacy on Elderly Health (English) [J]. Journal of Southeast University (English Edition), 2024, 40 (2): 1 - 10.

[2] Ran Xiaoxing, Hu Hongwei. The Urban-Rural Divide, Digital Divide, and Inequalities inElderly Health [J]. Population Journal, 2022, 44 (03): 46 - 58.

[3] Jiang Yunjie, Shao Yuzhao, Zhang Huijuan, Zhou Chenyu. Digital Challenges and Breakthroughs for Rural Left-behind Elderly in Nanjing under Digital Health [J]. Journal of Nanjing Medical University (Social Science Edition), 2024, 24 (04): 343 - 351.

[4] Huang Jiahao, Ma Li. Bridging the Digital Divide for Older Adults: Logic, Challenges, and Optimized Solutions [J]. Journal of Panzhihua University, 2024, 41 (03): 11 - 17.

[5] Xie Yongfei, Liu Yifeng. Exploring the Causes and Management of the Elderly Digital Divide from a Healthy Aging Perspective [J]. Journal of Minzu University of China (Philosophy and Social Sciences Edition), 2024, 51 (05): 125 - 134.

[6] Xu Hao. Enhancing the Digital Literacy of Older Adults: Definitions, Challenges, and Pathways [J]. Journal of Xinjiang Open University, 2024, 28 (03): 13 - 17.

[7] Shan Na, Zhang Xianling. Study on the Digital Disabilities of Older Adults and Their Measurement Indicators within the Digital Divide [J]. Gerontology Research, 2023, 11 (08): 50 - 62.

[8] Hu Yangming, Luo Jun, Gong Rengui. The Role of Socioeconomic Status, Digital Literacy, and Intergenerational Support in Paying for Smart Community Elderly Care: A Moderated Mediation Model [J]. China Health Service Management, 2023, 40 (07): 494 - 499+535.

[9] Zhang Xia. Data Analysis and Research on Diabetes Using SVM and Logistic Regression [J]. Journal of Cangzhou Normal University, 2023, 39 (01): 19 - 23+84.

[10] Wu Mengda, Mao Ziyang, Wang Dan. Shapley Value and Applications [J]. Mathematical Modeling and Its Applications, 2024, 13 (01): 110 - 119.