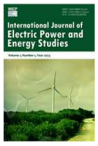
WEP Warwick Evans Publishing

International Journal of Electric Power and Energy Studies

ISSN: 3006-2004 (Print), ISSN: 3006-0826 (Online) | Volume 4, Number 1, Year 2025
DOI: https://doi.org/10.62051/ijepes.v4n1.05
Journal homepage: https://wepub.org/index.php/IJEPES/index



A Deep Learning-Based Approach for Relative Poverty Identification and Classification Prediction

Tianqi Ding

School of Computer Science and Technology, Taiyuan Normal University, Jinzhong, Shanxi, China

ABSTRACT

By predicting and classifying relative poverty, we can spot and tell the difference between potentially impoverished groups early on. This allows for early intervention and efficient resource allocation, aiding long - term poverty governance. Given the lack of algorithmic research in relative poverty identification using multi - year data, this paper proposes the RP - DCSA model. It blends deep learning (DNN) with the interpretable SHapley Additive exPlanation (SHAP) model. The 2020 China Family Panel Studies (CFPS) survey data form the research base. Spearman correlation coefficients are applied for feature selection to eliminate redundant ones. Next, the DNN - based RP - DCSA model is built and compared experimentally with LR, RF, etc. Finally, SHAP is used for interpretable analysis to identify key features affecting relative poverty classification and assess their impact on results. The RP-DCSA model achieves an 89.55% classification accuracy on the CFPS2020 dataset, outperforming other algorithms in various indicators.

KEYWORDS

DNN Model; SHAP Model; Relative Poverty Classification and Prediction; Feature Selection.

1. INTRODUCTION

In 2020, China achieved the historical goal of eliminating absolute poverty, signifying a major breakthrough in poverty reduction [1]. China is now shifting focus to relative poverty. Accurate identification and early warning of relative poverty are crucial for timely intervention [2]. Thus, applying deep learning models to relative poverty prediction becomes more important. They can leverage big data and advanced algorithms to precisely and rapidly identify the relatively poor, uncover hidden data patterns, and offer deeper predictive insights. Combining predictions from deep learning with explainable model analysis helps comprehensively understand the main factors and characteristics of relative poverty, supporting its management and intervention effectively [3-4].

Common machine learning algorithms like CART Tree can be used for classification and regression but require extensive data preprocessing and feature engineering. They are sensitive to outliers and noise, prone to overfitting or underfitting, and have limitations in handling nonlinear and complex data relationships. In contrast, DNN can capture associations between hidden features and manage complex nonlinear relationships, showing flexibility and generalization ability, which makes it suitable for relative poverty prediction. Therefore, building a relative poverty model that integrates DNN and SHapley Additive exPlanation (RP-DCSA model) is highly advantageous.

2. RELATED TECHNOLOGIES

2.1. DNN Model

A DNN is a multi - layer feedforward artificial neural network. Its key feature is that multiple hidden layers gradually extract hierarchical data features [5]. The basic DNN structure comprises an input layer, several hidden layers, and an output layer [6]. Neurons across layers are weight - connected and transmit signals via nonlinear activation functions. DNNs, deeper than shallow networks, can automatically learn complex data patterns without manual feature extraction [7]. They excel in function fitting and generalization, making them suitable for image recognition, speech processing, and natural language understanding. However, DNN training demands lots of labeled data and computational resources. Their "black - box" nature also causes poor interpretability. Despite this, DNNs, with outstanding performance in image classification, speech recognition, and machine translation, are a major driver of modern AI.

2.2. Model Interpretation

The SHAP model, based on game theory, explains machine learning by quantifying each feature's contribution to predictions [8]. Using the Shapley value from cooperative game theory, it calculates each feature's marginal contribution across all possible feature combinations, offering local or global output explanations [9]. The SHAP model unifies the mathematical frameworks of multiple interpretation methods (e.g., LIME and feature importance analysis), generating consistent feature contribution metrics for any machine learning model.

3. RELATIVE POVERTY IDENTIFICATION AND PREDICTION

3.1. Data Processing

The experimental dataset uses the 2020 China Family Panel Studies (CFPS2020) data, covering three types of questionnaires: family, adult, and child. The steps are as follows: first, remove invalid samples and irrelevant variables from the three databases and keep valid samples; second, merge the family and adult databases via family and personal codes; finally, combine the child data with the merged family data using personal codes to form a three - tier 25 - province dataset. Table 1 shows the dataset's description, identifying ten feature variables. Each data item has a category label (0 for non - relative poverty, 1 for relative poverty).

The Spearman correlation coefficient is a non - parametric statistic for measuring the monotonic relationship between two variables [10]. It assesses their correlation by calculating the Pearson correlation coefficient of their ranks (data ordered by magnitude). Its value ranges between - 1 and 1. When performing significance analysis on a feature set with the Spearman correlation coefficient, we can comprehensively evaluate the relationship between features and the target variable and determine the statistical significance of this relationship. The results for the CFPS2020 dataset are shown in Table 2. The four features with the lowest correlation coefficients are x1, x2, x4, and x5. Significance analysis using the Spearman correlation coefficient involves hypothesis testing to assess the significance of the correlation coefficient. This helps us understand whether the relationship between features and the target variable is practically significant. During this process, we calculate the p - value of the correlation coefficient, which is the probability of the correlation coefficient occurring under the null hypothesis. As shown in Table 2, for the CFPS2020 data, the p - values of x2 and x5 are greater than 0.05, indicating that these features do not significantly affect the target variable. Considering both the Spearman correlation coefficient and the p - value, we remove features x1, x2, x4, and x5, which represent adult education, child education, adult health, and medical insurance, respectively. The remaining six features are the most significant for relative poverty identification

and prediction. Their retention helps optimize model performance by removing redundant features, thereby enhancing the model's generalization ability and accuracy.

Table 1 Description of the Dataset

Dimension	Feature	Meaning	Description
Education	x1	Adult Education	Whether the population aged 16 and above has received education for more than 6 years
	x2	Child Education	Whether children aged 7 to 15 are not attending school
Health	x3	Adult Health	Whether adults self-rate their health as "unhealthy"
	x4	Child Health	Whether children (under 16) have visited a doctor three or more times in the past year due to illness
Living Standard	x5	Health Insurance	Whether members have health insurance
	x6	Cooking Fuel	Whether natural gas, liquefied gas, electricity, or solar energy can be used for cooking
	x7	Safe Drinking Water	Whether well water, tap water, or purified water can be used
	x8	Housing	Whether there is homeownership of housing
	x9	Assets	Whether the total value of durable goods and agricultural machinery is less than 4000 yuan
Income	x10	Annual Income	Whether the household's disposable income is below the relative poverty line standard

Table 2 Correlation Coefficients and p-values of the Dataset

Feature	X1	X2	X3	X4	X5	X6	x7	X8	X9	X10
Correlation Coefficient	0.001	0.002	-0.078	0.032	0.005	-0.297	-0.141	0.270	-0.665	-0.733
p-value	0.000	0.799	0.000	0.000	0.594	0.000	0.000	0.000	0.000	0.000

3.2. Construction of the Relative Poverty Identification and Prediction Model

This paper constructs the RP-DCSA model for relative poverty identification and classification,predicting relative poverty with DNN as the base model and quantifying feature importance via the SHAP framework. Using the CFPS2020 data, after preprocessing and merging, we create a three - tier dataset covering families, adults, and children across 25 provinces. We screen for features that significantly impact predictions by assessing feature correlations, streamlining the model and boosting its performance. Multiple machine learning models, including DNN, are built, trained, tested, and evaluated on relative poverty binary - classification tasks. SHAP is used to interpret prediction results, spot key features, and inform decision - making.

3.3. Prediction Results and Analysis

3.3.1. Evaluation Metrics

The prediction model in this study is built based on x3, x6, x7, x8, x9, x10, and y1. The classifier's predictions on the test dataset can be represented by the confusion matrix shown in Table 3.

Table 3 Confusion Matrix for Relative Poverty Prediction

—	Predicted Positive (Relative Poverty)	Predicted Negative (Not Relative Poverty)
Actual Positive	Number of correctly predicted classifications in positive samples (TP)	Number of incorrectly predicted classifications in positive samples (FN)
Actual Negative	Number of incorrectly predicted classifications in negative samples (FP)	Number of correctly predicted classifications in negative samples (TN)

To evaluate the model based on this matrix, the following classification performance metrics are used:

(1) Accuracy: It represents the percentage of correctly predicted samples out of the total samples. The formula for accuracy is as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

(2) Precision: It represents the percentage of correctly predicted positive samples out of all samples predicted as positive. The formula for precision is as follows:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

(3) Recall: It represents the percentage of correctly predicted positive samples out of all actual positive samples. The formula for recall is as follows:

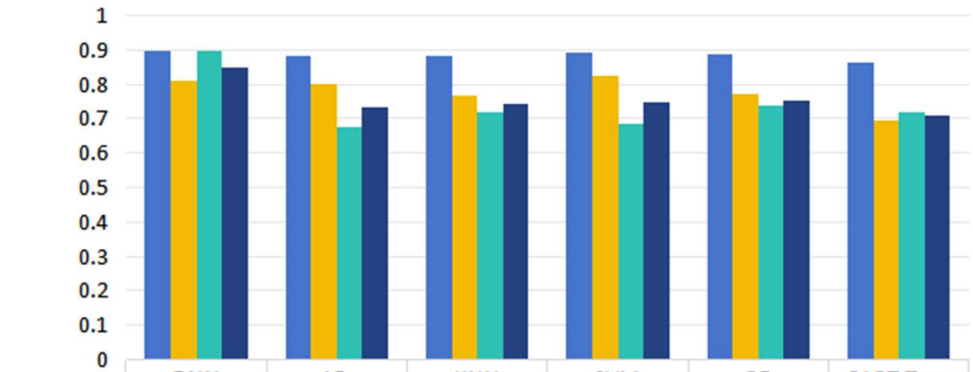
$$recall = \frac{TP}{TP + FN} \tag{3}$$

(4)F1 Score: It is a metric that combines precision and recall to provide a single value that balances both measures. The F1 Score reaches its best value at 1 and worst at 0. The formula for F1 Score is:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \times 100\% \tag{4}$$

3.3.2. Result Analysis

DNN is compared with five existing models:LR, KNN, SVM, RF, and CART Tree. 10 - fold cross - validation is used to ensure fairness and credibility in model comparison. Bar charts of accuracy, precision, recall, and F1 - score (Fig. 1) show that DNN has higher prediction accuracy with an accuracy of 89.55% and precision of 80.89%.



	DNN	LR	KNN	SVM	RF	CART Tree
accuracy	0.8955	0.8844	0.8844	0.8919	0.8872	0.8611
precision	0.8089	0.7981	0.7692	0.823	0.7699	0.6962
recall	0.8943	0.6773	0.72	0.6853	0.736	0.7211
f1	0.8494	0.7328	0.7438	0.7478	0.7526	0.7084

Fig.1 Performance Comparison of Various Models on the CFPS2020 Dataset

4. ANALYSIS OF RELATIVE POVERTY CHARACTERISTICS

SHAP summary plots offer a global interpretation of feature impacts on predictions, visualizing the magnitude and direction of these impacts. Fig. 2 presents the SHAP summary plot from the experiment. Annual income (x10) and cooking fuel (x6) are the most influential features in relative poverty identification, while safe drinking water, housing (x8), and assets are secondary factors with fluctuating importance across different years. This indicates that living standard variables significantly affect relative poverty identification, followed by health variables. Housing (x8) has an inhibiting effect on relative poverty predictions. Cooking fuel (x6), assets (x9), and annual income (x10) are positively correlated with relative poverty. Over time, the importance of adult health (x3) increases, while that of assets (x9) decreases, highlighting the growing significance of adult health in reducing the risk of relative poverty.

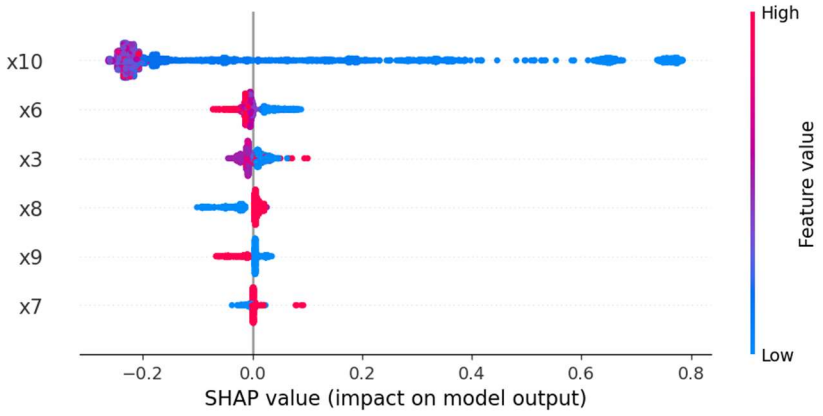


Fig. 2 SHAP Overview Diagram

5. CONCLUSION AND PROSPECT

This paper presents a high - performance RP-DCSA model for relative poverty prediction and identification. It uses the Spearman correlation coefficient for feature selection, enhancing training and inference speed. The model is flexible and analyzes features affecting relative poverty predictions, thus improving subsequent prediction results. The SHAP model aids in accurately understanding each feature's impact on the model, effectively guiding resource allocation and intervention in relative poverty governance. Experimental results show that our model excels in relative poverty prediction. Developing a model to identify and warn of potential relative poverty groups and analyzing features will greatly improve the understanding and intervention of relative poverty, providing significant support for its governance.

REFERENCES

[1] Tan,X.W.,Dong,M.,Ou.Y.X.,et al.(2024).New Advances in Anti-Poverty Research in the New Era: A Summary of the International Seminar on "Innovations in Anti-Poverty Theory". Chinese Rural Economy, 06, 173-184.

[2] Zhou,Z.H.,&Li,X.Y.(2024).Analysis of the Spirit of Poverty Alleviation and Common Prosperity in the New Era. Shanghai Economic Research, 424(01), 5-14.

[3] Chen,W.Q.,&Yang,Z.L.(2024). Research on the Modernization Path of Governance of Relative Poverty in Chinese Small-Scale Agriculture. Agricultural Economics,05, 102-104.

[4] Qiu,H.,&Du,Z.L.(2024). Multi-dimensional Dilemmas and Mechanism Construction of Governance of Relative Poverty. Academic Exchange,02, 143-157.

[5] Luo,Y.Q.,Liu,M.P.,Chen,S.W.,et al.(2025).Charging Load Forecasting and Optimal Scheduling Model Based on Deep Neural Network. Electronic Design Engineering,33(08),92 - 95.DOI:10.14022/j.issn1674 - 6236.2025.08.019.

[6] Qin,X.Y.,Luo,D.,Ye,C.(2025).Analysis of Factors Affecting Average Hospitalization Days Based on Deep Neural Network. Modern Hospital,25(03),388 - 392.

[7] Luo,H.,Li,P.(2025).Establishment of Enhanced CT Image Kidney Segmentation Model Based on Deep Neural Network. Chongqing Medical Journal,54(03),630 - 634.

[8] Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. Computers, Environment and Urban Systems, 96, 101845.

[9] Feng, D.C., Wang, W.J., Mangalathu, S., et al. (2021). Interpretable XGBoost-SHAP machine-learning model for shear strength prediction of squat RC walls. Journal of Structural Engineering, 147(11), 04021173.

[10] Pan,R.P.,Yang,H.,Xin,B.X.,et al.(2025).Monitoring and Early Warning Method for Feeder Power Failure Risk Based on Spearman Correlation Coefficient. China New Technologies and New Products,(05),137 - 139.DOI:10.13612/j.cnki.cntp.2025.05.026.