IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 23, 2022, accepted March 15, 2022, date of publication March 22, 2022, date of current version March 30, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3161574

Rare Potential Poor Household Identification With a Focus Embedded Logistic Regression

YAN-XUE WU[1], **ZHI-NENG HU**[2], **YUAN-YUAN WANG**[2], **AND FAN MIN**[3], **(Member, IEEE)**
[1]School of Computer, Sichuan Technology and Business University, Chengdu 611745, China
[2]Business School, Sichuan University, Chengdu 610065, China
[3]School of Computer Science, Southwest Petroleum University, Chengdu 610500, China

Corresponding author: Yan-Xue Wu (yeah_imwyx@163.com)

This work was supported in part by the Chinese National Funding of Social Science under Grant 18AGL022, and in part by the Fundamental Research Funds for the Central Universities under Grant 2021ZY-SX02.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the School of Computer, Sichuan Technology and Business University.

ABSTRACT With the rapid development of poverty alleviation in China, multidimensional poverty identification has always been challenging. This paper adopted a focus embedded logistic regression (FeLR) to solve two types of difficulties–the rarity and hard-distinguishability, of the potential poor household (PPH) identification. The PPH identification was decomposed into two subproblems–the potential re-poverty household (PRPH) identification, and the potential unidentified poor household (PUPH) identification. The FeLR embedded a focal loss to deal with the hard-distinguishability, and adopted a weighting technique to address the rarity. The sample weight exponent was extended to negative values to overlook the hard negative samples. This setting significantly improved the recall of PPHs, compared with that using traditional logistic regression. A few indicators were critical to the incidence of PPH, especially the household income per capita, medical expenses for chronic diseases, and house structure. Local policy makers are suggested to pay more attention to the crucial indicators to against the poverty contrapuntally.

INDEX TERMS Focal loss, logistic regression, potential poor households, rare events.

I. INTRODUCTION

Promoting rural development, eradicating poverty, and achieving common prosperity are all objectives that have been regularly targeted [1]. However, as poverty alleviation remains a major challenge for the global community, external interventions are vital, such as poverty reduction policies [2]. As part of the effort, China completed the nationwide registration of poor population in 2013, and implemented the Targeted Poverty Alleviation (TPA) strategy to move households out of poverty by 2020 [3]. Since then, the anti-poverty work has entered a crucial period, transferring the focus from "poverty alleviation" to "poverty prevention" [4].

During this process, a key issue was to prevent the poverty-alleviated household (i.e., the household who has been relieved from poverty, PAH) and the non-poor household (NPH) from falling into poverty again. PAHs referred to those that have reached the poverty-alleviated standard and have completed the poverty-alleviated procedure, while PHs were those scheduled to be out of poverty after 2017 (See Appendix A. for details). A small part of PAHs and NPHs were defined as *potential re-poverty households* (PRPHs) and *potential unidentified poor households* respectively, collectively referred to as *potential poor households* (PPHs). Both PAHs and NPHs were non-poor households, while PPHs could be seen as those that are sometimes poor with floating, unstable, dynamic characteristics, and the risk of falling back into poverty again [5]. Concretely, the PRPHs were only hidden in PAH and referred to households who were at high risk of returning to poverty from poverty-alleviated status. The PRPHs were judged by the local government to be PAHs, but their poverty-alleviated status was unstable. The PUPHs were only presented in NPH and were not pre-detected by the local government, and they were highly likely to be poverty households, and yet were omitted. Note that PPHs contained PRPHs and PUPHs, and both PRPH and PUPH can also be called PPH. But compared to PPH, the terms PRPH and PUPH were more precise. Yang *et al.* [6] suggested that the

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.

government should accelerate the establishment of a poverty identification and dynamic monitoring system. However, the identification of PPHs has not been resolved yet, or rather its two facets–rarity and hard-distinguishability.

Since the PRPHs only existed in PAHs, and PUPHs only existed in NPHs, the PPH identification problem could be divided into two individual binary classification problems, namely PRPHs classification and PUPHs classification. The PRPHs would be identified from the data consisting of PRPHs and PAHs, and the PUPHs would be identified from the data comprised of PUPHs and NPHs. To effectively identify PRPHs and PUPHs, this paper adopted a focus embedded logistic regression (FeLR). The FeLR encompassed a focal loss to tackle the hard-distinguishability, and used a weighting technique to resolve the rarity [7]. The sample weight exponent was extended to negative values, yielding a significantly higher recall of PRPH compared with that using traditional logistic regression.

The main contributions of this paper were: 1) A focus embedded logistic regression model was adopted for the rare potential poor household identification from the perspective of social computing; 2) The sample weight exponent was extended to negative values for hard-distinguishability, yielding a significant higher recall of potential poor households, compared with those using the methods of traditional logistic regression, weighted logistic regression and their varieties; 3) The differences and commonalities between the crucial indicators generated by FeLR and the traditional method were compared, and the reasons why the crucial factors obtained by the traditional method were not convincing were carefully analyzed; and 4) The crucial indicators impacting potential poor households were picked out by FeLR to help local policy makers combat poverty contrapuntally.

The remainder of this paper was organized as follows. Section II specified the research objectives and summarized the literature. Section III concerned the data collection and preprocessing. The model was demonstrated in Section III-B. Section IV presented the results and discussion. Section V drew key conclusions and envisaged future researches.

II. LITERATURE REVIEW

A. MULTIDIMENSIONAL POVERTY IDENTIFICATION

Multidimensional poverty could better reflect the real living standard compared with single dimensional poverty with income only [8]. Many measures of multi-dimensional poverty have been proposed, but selecting and integrating poverty indices remains challenging. Two strategies have been usually adopted to integrate the multidimensional poverty indices. One was to combine various factors into a comprehensive poverty index [9]. The other multidimensional poverty estimation method called the counting approach, was used to identify the number of deprived dimensions [10]. Simply put, they focused on designing or selecting some indicators that well reflect the real poverty.

To investigate the causes of poverty or effectively reduce poverty, many applications have been studied through statistical models. For example, Xu *et al.* [11] illustrated that "both settlements isolation and land use changes had an impact on poverty", through geographically weighted regression and Pearson correlation analysis. Xu *et al.* [11] suggested that the distance between settlements should be kept less than 5000m. Pathak *et al.* [12] adopted fixed effects models to estimate how bus transit access affected poverty. However, they did not examine individual or household movements in and out of the census tracts [12]. The quasi-experimental design showed that the microfinance had a mild, positive impact on reducing poverty [13] in North-eastern Mindanao, the Philippines. They also surveyed some microfinance household profiles from a micro perspective, gaining policymakers and MFIs valuable insights into the households' motivations to access microfinance [13]. Liao and Fei [14] analyzed the effectiveness of photovoltaic-based development intervention programs on poverty reduction, by examining their determinants with a spatial lag model. However, installing PV systems in some rural areas was difficult because of sparse population there, or rather inactive economy and low electricity demands. Adopting alternative measures of disability severity, Palmer*et al.* [15] illustrated that disability severity was positively associated with poverty at a health-demographic surveillance site in VietNam. Pasanen [16] analyzed the factors of poverty on household and village levels in Laos with two-level structural equation modelling, but they were unable to specify the dynamic between migration and poverty with cross-sectional data.

To accurately identify poverty, machine learning has been used to directly work out the coefficient of determination for regression, or the measures for classification, such as accuracy, recall, and F-score [17]. A random forest regression model estimated the household wealth index through multiple data sources [18]. Random forest also could contribute to better poverty predictions than multiple imputation did with variables selected by stepwise regression and Lasso regression [19]. Both studies barely analyzed the factors influencing poverty from a sociological perspective. Blumenstock *et al.* [20] predicted the wealth of mobile phone subscribers through regularized logistic regression with cross-validation. They, too, believed that the predicted attributes of millions of individuals could, in turn, accurately reconstruct wealth distributions across countries or infer asset distributions in micro-regions. The clipped Gaussian geo-classification (CGG-C) model could be applied to the poverty of household, in addition to existing methods [21]. A prediction map of poverty distribution was depicted through CGG-C model, although no practical policy recommendations were given.

B. IMBALANCED DATA CLASSIFICATION WITH MACHINE LEARNING

As far as poverty researches are concerned, imbalanced data classification tends to be neglected. However, imbalanced data (i.e., rare events data) classification has been prevalent in many other applications, such as medical diagnosis, fraud detection, and network intrusion detection [22]. In the past decades, many machine learning methods have been designed to improve the performance on imbalanced data. In practice, they could be divided into five categories, namely one-class learning, feature selection, cost-sensitive learning, data level approaches, and ensemble learning [23]. Machine learning holds reference significance for accurate poverty identification.

In terms of good interpretability, logistic regression has been a powerful classifier for imbalanced data [24]. Maalouf and Trafalis [25] proposed the rare event weighted kernel logistic regression (RE-WKLR), yielding good performances on the imbalanced data, small or large. The RE-WKLR algorithm took advantage of bias correction and the power of the kernel methods, given that the datasets were neither balanced nor linearly separable. Compared with other machine learning methods, rare events logistic regression was more robust on rare events data and relatively simple to be implemented [26]. Recent studies have shown that logistic regression is suitable for predicting major chronic diseases, where it yields as good performance as other machine learning approaches such as random forest, support vector machine, and neural network do [27]. Nusinovici *et al.* [27] also suggested that traditional regression models should continue to play a key role in disease risk prediction when using a limited set of simple clinical predictors, and they downplayed the role of other traditional machine learning techniques in major chronic diseases prediction. That seemed reasonable, because by comparing logistic regression with 5 traditional machine learning algorithms, they concluded that logistic regression was indeed more interpretable than some other machine learning methods.

III. DATA AND METHODOLOGY

A. DATA

A lot of preliminary work was done in the data collection stage. The data used in this paper would not have been possible without the efforts of all the 63 members of the Third-party Poverty Assessment Team of Sichuan University in 2017 1. First, the team members gave out different forms of questionnaires to three types of households (i.e. PH, PAH, and NPH). Second, the members visited southwestern China, and did field survey there.[1] They collected PH, PAH, and NPH data from 20 administrative villages via stratified random sampling. For villages with a full coverage of PHs, NPHs were randomly selected by 1:1, whereas for sampling villages, both PHs and NPHs were randomly selected to conduct questionnaires.

Finally, a total of 1009 valid questionnaires were obtained, from 74 PHs, 427 PAHs and 508 NPHs. Based on the collected data, the Third-party Poverty Assessment Team held discussions with relevant cadres of the local government, and interviewed the cadres on the township level. Guided by the Provincial Technical Consultant, 6 PAHs and 9 NPHs were suspected to be the PHs, thus defined as potential re-poverty households (PRPHs), and potential unidentified poor households (PUPHs), respectively. In other words, PRPHs/PUPHs might actually be PHs, but they could be wrongly determined by the local government as PAHs/NPHs. Note that PRPHs only existed in PAHs, PUPHs only existed in NPHs. Plus, both PRPHs and PUPHs are pretty less (Fig. 2). PRPHs accounted for only 1.4% of PAHs, and PUPHs only 1.77% of NPHs.

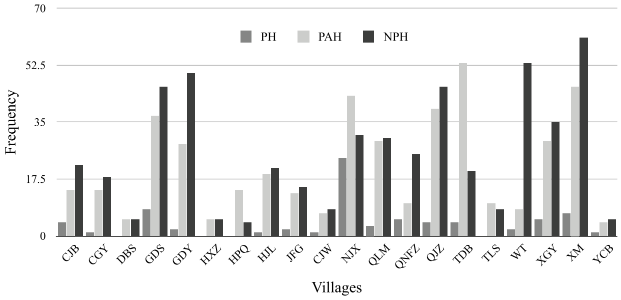


FIGURE 1. Data frequency of PHs, PAHs, and NPHs in different villages.

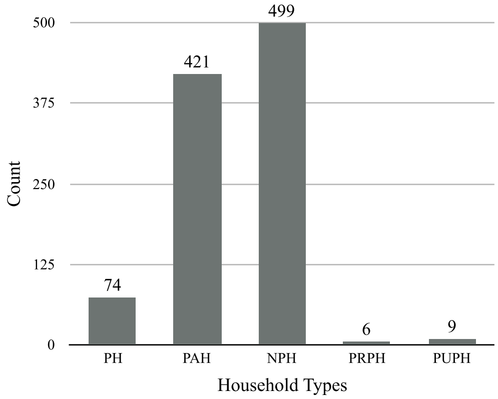


FIGURE 2. Number of different types of households.

The original indicators of questionnaires for PHs, PAHs, and NPHs were interlinked, but not completely the same with each other. For instance, some indicators in the questionnaires of PHs and PAHs were meaningless to NPH, such as "special poverty alleviation measures" and "industrial poverty alleviation measures". The questionnaires of PHs and PAHs had a total of 249 indicators, whereas those of NPHs had 124 indicators. To make indicator systems concise, this research only chose indicators that might be useful and representative in the counterpart of different indicator systems, but left out contents with little impact or no proper indicators at all. The indicator system contained 4 top-level perspectives, namely "Household demographic characteristics", "Basic needs", "Safe water & Domestic electricity & Radio signal", and "Debt & Income" (Table 1).

[1] Informed consent was obtained.

These indicators were combined to measure the multidimensional rural poverty at the household level (Table 1). "Household demographic characteristics" contained 9 indicators related to household population. Most of them were statistical indicators that reflected the population of different groups in a household. There were evidences of a direct or indirect relationship between most of these indicators and poverty of household level. For instances, McBride and Nichols [17] adopted out-of-sample validation and stochastic ensemble to improve the out-of-sample performance of poverty targeting tools, where they adopted the age of householder, and the number of household members. To predict poverty of household level, Puurbalanta [21] introduced the age of householder, the educational years of householder, and the number of family members. Ntsalaze and Ikhide [28] indicated that the educational years of householder were associated with multidimensional poverty, where they also introduced the age of householder, the number of the juveniles under 16, and the number of family members to predict the household's deprivation score. Rupasingha and Goetz [29] indicated that less the quantity of workforce might exacerbate poverty. Some literature illustrated that these indicators might be associated with some policy-alleviated policies, such as microfinance [13] and Dibao [30]. Gao *et al.* [30] also introduced the number of elders over 60, the number of children under 18, and the employment status of householder. "Basic needs" laid out the need of food and clothing, basic medical care, compulsory education, and safe housing. The 28 indicators of "Basic needs" comprehensively measured the basic material demand of a household (see Appendix A. for more details). "Safe water & Domestic electricity & Radio signal" and "Debt & Income" had 4 and 8 indicators, respectively.

The value of all the binary variables was either of [Yes, No], except those with missing values (i.e., "house is safe", "house meets the needs of production and life", "houses for humans and livestock are separated", "be able to afford medical expenses for chronic disease"), which were filled to "Unknown". In the preprocessing stage, all the categorical variables were decomposed to dummy types. Since different variables had different scales, each continuous or dummy variable was preprocessed with z-scores normalization. As the classifications of PAH/PRPH and NPH/PUPH were independent of each other, z-scores were computed separately for them.

B. METHODOLOGY

Let π_i be the prediction probability of PRPH or PUPH, given by

$$\pi_i = \frac{1}{1+\exp(-\mathbf{x}_i\boldsymbol{\beta})}, \tag{1}$$

where the unknown parameter $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^k$, β_0 was a scalar constant term, and $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_{k-1})^{\mathrm{T}} \in \mathbb{R}^{(k-1)}$ was a vector whose elements corresponded to explanatory variables.

To solve the rarity and hard-distinguishability of PPHs, the log-likelihood of FeLR was given by

$$\log L_w^F(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_1(1-\pi_i)^{\gamma} Y_i \log \pi_i + w_0\pi_i^{\gamma}(1-Y_i)\log(1-\pi_i) + A(\boldsymbol{\beta}), \tag{2}$$

where $w_0 = \frac{1-\tau}{1-\bar{y}}, w_1 = \frac{\tau}{\bar{y}}$ compensated for differences in the sample ($\bar{y}$) and population (τ) fractions of ones induced by choice-based sampling [24], γ adjusted the relative importance of hard (i.e., hard-distinguishable) and easy samples [7], and $A(\boldsymbol{\beta})$ was the penalized function. Here, let $A(\boldsymbol{\beta})$ be the $\ell_1(\boldsymbol{\beta})$ to do variable selection and ensure better interpretability, and three other types of penalized functions were used to do comparison study, namely Ridge (ℓ_2), ElasticNet (ℓ_{12}), and Firth-type (ℓ_{Firth}) (see Appendix B. for details). For the identification of PPHs, $Y_i = 1$ or 0 meant that the household was PRPH (PUPH) or PAH (NPH).

Lin *et al.* [7] argued that the sample weight exponent γ should be set to a positive. Therefore, γ was firstly set from 0.1 to 0.5 with step 0.1, that was, $[0.1, 0.2, 0.3, 0.4, 0.5]$. However, it did not improve the micro-R of PRPH or PUPH, which was where attention should be paid. Even with the increase of γ, the micro-R of PRPH decreased unexpectedly, whereas that of PUPH remained almost unchanged. An unconventional manner of γ with negative values ranging from -0.1 to -0.5 with step -0.1 was adopted, yielding quite good results (subsection IV-B).

From the perspective of machine learning, 3 measures, namely micro-P, micro-R, and micro-F_1 were adopted to evaluate the classification results of PRPH and PUPH. From the statistical point of view, 5 indices were adopted to test the goodness-of-fit, including deviance D_1, D_2, and three statistical pseudo R^2, namely McFaddden's R^2 (R^2_{mf}), Cox-Snell's R^2 (R^2_{cs}), and Nagelkerke's R^2 (R^2_{nk}) (see Appendix C. for details).

IV. RESULTS AND DISCUSSION

This section mainly discussed the effectiveness of FeLR and the determinants of PRPHs and PUPHs. To illustrate the effectiveness of FeLR, it was compared with the traditional logistic regression and the penalized counterpart. The impact of γ setting was discussed in detail, where $\gamma < 0$ was adequately explained. Via the summary statistics of the sample, the anticipated determinants of PPHs were illustrated. Finally, the goodness-of-fit of the model was examined through various of statistics and the determinants of PRPHs and PUPHs were analyzed.

A. COMPARISON BETWEEN FeLR AND OTHER MODELS ON PRPHs AND PUPHs

To illustrate the effectiveness of the FeLR, two groups of models were constructed for comparison. The difference between the two groups of models lied in whether it was focus embedded or not (Table 2).

TABLE 1. Indicator system.

Top-level Perspective	Second-level Perspective	Indicator (Variable)	Type
Household demographic characteristics	-	Age of the householder	Continuous
		Education level of the householder. Any of [Primary school or below, Middle school, Vocational school, High school, Bachelor or above]	Categorical
		Number of family members	Continuous
		Number of the elders over 60	Continuous
		Number of the juveniles under 16	Continuous
		Quantity of workforce	Continuous
		Number of agricultural members	Continuous
		Number of migrant workers	Continuous
		Number of permanent residents	Continuous
Basic needs	Basic needs for food	Lack of food	Binary
		Frequency of eating nourishing food. Any of [During festivals, Occasionally, Often, Frequently]	Categorical
		No worry about food	Binary
	Basic needs for clothing	Change clothes daily	Binary
		There are seasonal clothes throughout the year	Binary
		Main source of clothes. Any of [Self-purchase, From sons and daughters, Donation, Others]	Categorical
		No worry about wearing	Binary
	Safe housing	Number of houses	Continuous
		Housing structure. Any of [None, Cave dwelling, Adobe, Bamboo, Brick wood, Brick, RC, Others]	Categorical
		House is safe	Binary
		House meets the needs of production and life	Binary
		Houses for humans and livestock are separated	Binary
		The owner of the current house. Any of [Self, Sons and daughters, Relatives, Renting]	Categorical
		Enjoy the housing subsidy policies	Binary
		Housing construction area	Continuous
		Government subsidy funds for housing construction	Continuous
		Self-raised funds for housing construction	Continuous
		Debt for housing construction	Continuous
	Compulsory education	There are children in compulsory education	Binary
	Basic medical care	The type of medical insurance covered. Any of [None, One type, Two types, Three types, Four types]	Categorical
		Level of disability. Any of [None, Level 1, Level 2, Level 3, Level 4, Level 5, Level 6, Level 7]	Categorical
		There are long-term chronic patients	Binary
		Type of chronic disease. Any of [None, ENT, Cardiovascular, Lung, Liver, Gastrointestinal, Kidney, Rheumatism, Orthopedic, Nervous system, Cancer, Others]	Categorical
		Able to afford medical expenses for chronic disease	Binary
		Number of critically ill patients	Continuous
		Type of critical illness. Any of [None, ENT, Cardiovascular, Lung, Liver, Gastrointestinal, Kidney, Rheumatism, Orthopedic, Nervous system, Cancer, Others]	Categorical
		Self-raised funds for the treatment of critical illness	Continuous
		Reimbursement ratio for the treatment of critical illness	Continuous
Safe water & Domestic electricity & Radio signal	Safe water	Guaranteed water source	Binary
		The water resource. Any of [Tap water, Well water, Spring, River, Water tank, Water cellar, Others]	Categorical
	Domestic electricity	Voltage is stable in daily life	Binary
	Radio signal	There is sound radio and television signal	Binary
Debt & Income	Household debt	Type of debt. Any of [None, House construction, Education, Disease, Others]	Categorical
		Type of creditor. Any of [None, Individual, Financial institutions, Others]	Categorical
	Household income	Household business income	Continuous
		Wage income	Continuous
		Property income	Continuous
		Transfer income	Continuous
		Primary income source. Any of [Work, Aquaculture, Crop farming, Self-management, Relatives, Government subsidies, Others]	Categorical
		Household income per capita exceeds 3300 CNY	Binary

TABLE 2. **The models used for comparison.**

Group	Model Name	Weighting	Focus Embedded	ℓ_1	ℓ_2	ℓ_{Firth}
Group I	Baseline	–	–	–	–	–
	WLR	✓	–	–	–	–
	WLR-ℓ_1	✓	–	✓	–	–
	WLR-ℓ_2	✓	–	–	✓	–
	WLR-ℓ_{12}	✓	–	✓	✓	–
	LR-ℓ_{Firth}	–	–	–	–	✓
Group II	F-Baseline	–	✓	–	–	–
	F-WLR	✓	✓	–	–	–
	F-WLR-ℓ_1	✓	✓	✓	–	–
	F-WLR-ℓ_2	✓	✓	–	✓	–
	F-WLR-ℓ_{12}	✓	✓	✓	✓	–
	F-LR-ℓ_{Firth}	–	✓	–	–	✓

TABLE 3. **Classification results of different models for the identification of PRPHs ($\gamma = -0.5$).**

Model	$\overline{\text{TP}}$	$\overline{\text{FN}}$	$\overline{\text{FP}}$	$\overline{\text{TN}}$	micro-P	micro-R	micro-F_1
Baseline	107	133	27	213	**0.7985**	0.4458	0.5722
F-Baseline	160	80	42	198	0.7921	**0.6667**	**0.7240**
WLR	172	68	50	190	0.7748	0.7167	0.7446
F-WLR	240	0	56	184	**0.8108**	**1**	**0.8955**
WLR-ℓ_1	144	96	36	204	0.8000	0.6000	0.6857
F-WLR-ℓ_1	240	0	55	185	**0.8136**	**1**	**0.8972**
WLR-ℓ_2	175	65	46	194	0.7919	0.7292	0.7593
F-WLR-ℓ_2	240	0	55	185	**0.8136**	**1**	**0.8972**
WLR-ℓ_{12}	169	71	41	199	0.8048	0.7042	0.7511
F-WLR-ℓ_{12}	240	0	55	185	**0.8136**	**1**	**0.8972**
LR-ℓ_{Firth}	110	130	28	212	**0.7971**	0.4583	0.5820
F-LR-ℓ_{Firth}	160	80	42	198	0.7921	**0.6667**	**0.7240**

For PRPHs, the micro-R and micro-F_1 of all the models in Group II were better than those of Group I (Table 3). The micro-P of nearly all the models in Group II were better than those of Group I. Only the micro-P of F-Baseline and F-LR-ℓ_{Firth} were slightly worse than those of Baseline and LR-l_{Firth}, respectively, whereas the micro-R of them were much better than those of corresponding methods. Specifically, the micro-R of F-WLR, F-WLR-ℓ_1, F-WLR-ℓ_2, and F-WLR-ℓ_{12} are 1, which meant that all PRPHs were recalled.

Intuitively, the precision-recall curves of all the models in Group II mostly covered those of the models in Group I (Fig. 3). Especially when micro-R was nearly 0.2, the two curves showed clear differences. When micro-R $=$ 1.0, micro-P decreased continuously of any focus embedded method, whereas those of traditional methods were relatively stable. This phenomenon indicated the different $\overline{\text{FP}}$ between focus embedded methods and others (see Table 3), due to the different propensity of models w.r.t. the change of γ (see Subsection IV-B).

For PUPHs, the micro-R of all the methods were not so good, whereas the F-Baseline still performed best on all the three measures (Table 4). Although the micro-P of all the focus embedded methods were worse than those of others, except for the F-Baseline, they were achieved by reducing $\overline{\text{FP}}$, but not increasing $\overline{\text{TP}}$. The micro-R of WLR, WLR-ℓ_1, WLR-ℓ_2, WLR-ℓ_{12}, and the corresponding focus embedded methods were unchanged, due to three outliers in PUPHs.

TABLE 4. **Classification results of different models for the identification of PUPHs ($\gamma = -0.5$).**

Model	$\overline{\text{TP}}$	$\overline{\text{FN}}$	$\overline{\text{FP}}$	$\overline{\text{TN}}$	micro-P	micro-R	micro-F_1
Baseline	199	161	31	329	0.8562	0.5528	0.6718
F-Baseline	251	109	38	322	**0.8685**	**0.6972**	**0.7735**
WLR	240	120	48	312	**0.8333**	0.6667	**0.7407**
F-WLR	240	120	59	301	0.8026	0.6667	0.7287
WLR-ℓ_1	240	120	42	318	**0.8511**	0.6667	**0.7477**
F-WLR-ℓ_1	240	120	57	303	0.8081	0.6667	0.7306
WLR-ℓ_2	240	120	44	316	**0.8451**	0.6667	**0.7454**
F-WLR-ℓ_2	240	120	59	301	0.8027	0.6667	0.7284
WLR-ℓ_{12}	240	120	44	316	**0.8451**	0.6667	**0.7454**
F-WLR-ℓ_{12}	240	120	58	302	0.8054	0.6667	0.7295
LR-ℓ_{Firth}	199	161	101	259	**0.6633**	**0.5528**	**0.6030**
F-LR-ℓ_{Firth}	192	168	115	245	0.6254	0.5333	0.5757

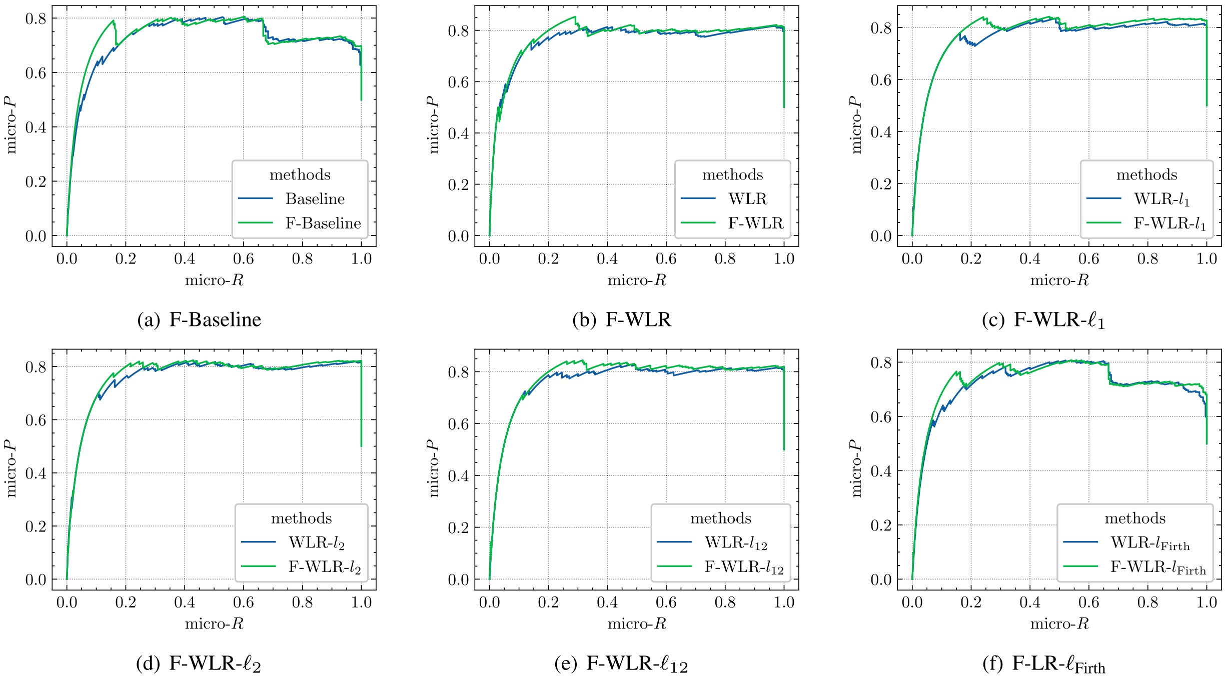
Generally, the precision-recall curves of the models in Group II could cover those of the models in Group I, except for the F-LR-ℓ_{Firth} (Fig. 4). When micro-R was around 0.667, the micro-P of most methods saw a drastic drop. Three outliers in PUPHs were quite different from the rest. Although outliers could not be recalled while maintaining a relatively high precision, most models in Group II still covered those of the models in Group I.

B. THE IMPACT OF γ SETTING on the RECALL OF PRPHs AND PUPHs

The results (Table 3 and Table 4) were explained through analyzing the impact of different γ settings on the classification results. If $\gamma = 0$, the model was the same as if with no focus embedded. The micro-R of PRPHs gradually decreased from 1.0 to around 0.5 with the increase of γ, whereas the change of micro-P was not obvious of most focus embedded methods, except for the F-Baseline and F-LR-ℓ_{Firth} (Fig. 5(a) and Fig. 5(b)). The smaller the γ, the more likely it was to correctly classify more PRPHs. The $\overline{\text{FP}}$ increased with the $\overline{\text{TP}}$, due to the inconspicuous change of micro-P.

However, the identification results of PUPHs were very different from those of PRPHs (Fig. 6). Obviously, the micro-R of most methods for PUPHs had not changed at all regardless of γ, but F-Baseline achieved the max micro-R (Fig. 6(b)). Although there might be outliers in PUPHs that could not be recalled under the combined effect of weighting and focus embedding, the F-Baseline model still performed best on micro-R when $\gamma = -0.5$. The micro-P of PUPHs increased slightly with γ (Fig. 6(a)). The reason was not the increase in $\overline{\text{TP}}$, but the decrease in $\overline{\text{FP}}$, which was the same as in the case of PRPHs. This result was consistent with the analyses on Table 3 and Table 4.

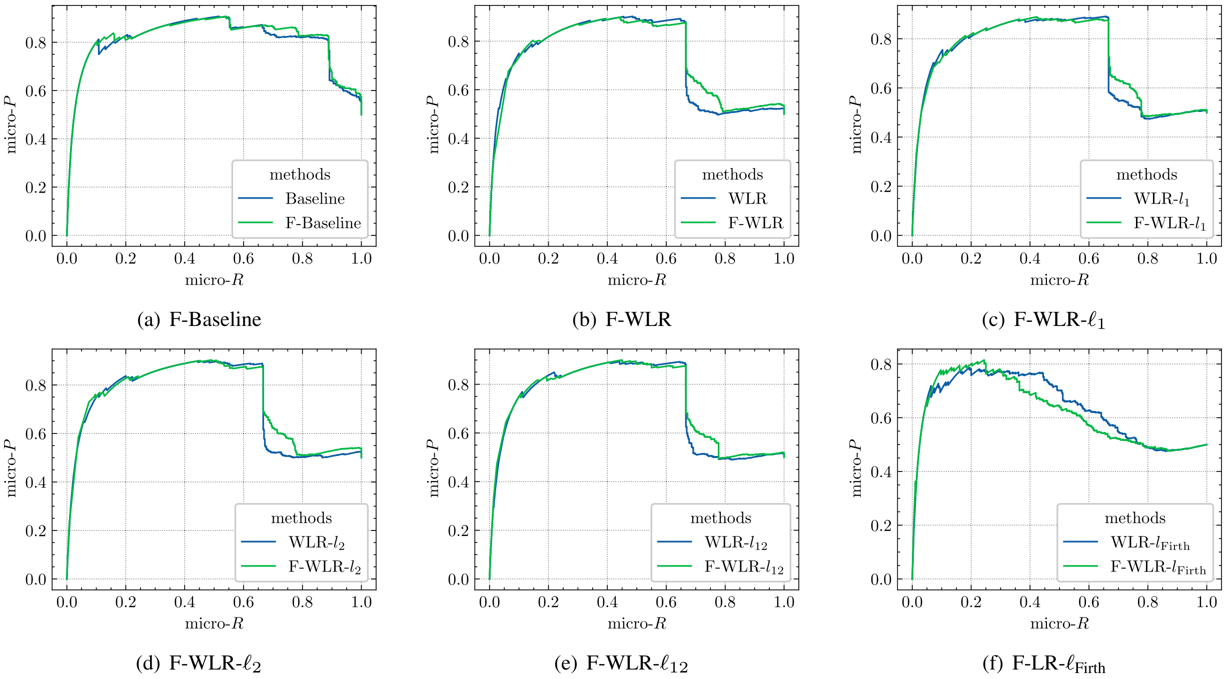
In the context of FeLR, the sample weight exponent γ was to measure the degree of relative importance between hard and easy samples [7]. An unconventional manner on γ that enabled negative values was adopted, whereas Lin *et al.* [7] argued that γ should be greater than 0. The blue circle points represented the negative class, and the red square points represented the positive class. Here the number of the positive was much less than that of the negative (Fig. 7).



(a) F-Baseline (b) F-WLR (c) F-WLR-ℓ_1

(d) F-WLR-ℓ_2 (e) F-WLR-ℓ_{12} (f) F-LR-ℓ_{Firth}

FIGURE 3. **Precision-recall curves of different models for the identification of PRPHs.**



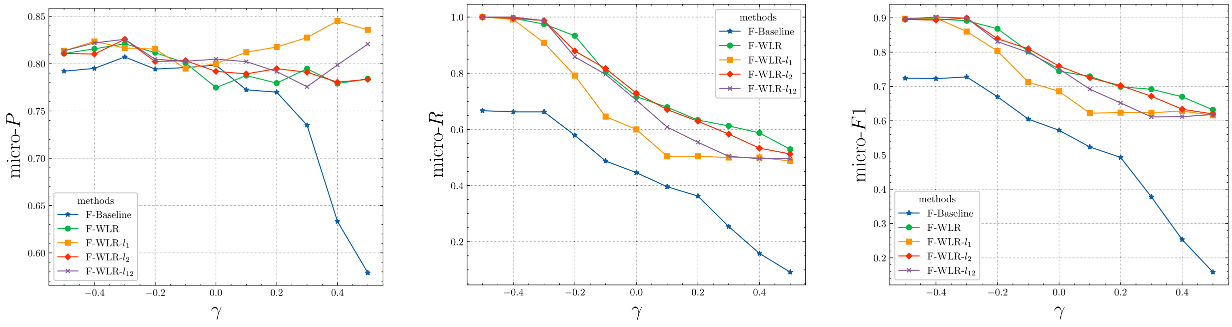
(a) F-Baseline (b) F-WLR (c) F-WLR-ℓ_1

(d) F-WLR-ℓ_2 (e) F-WLR-ℓ_{12} (f) F-LR-ℓ_{Firth}

FIGURE 4. **Precision-recall curves of different models for the identification of PUPHs.**

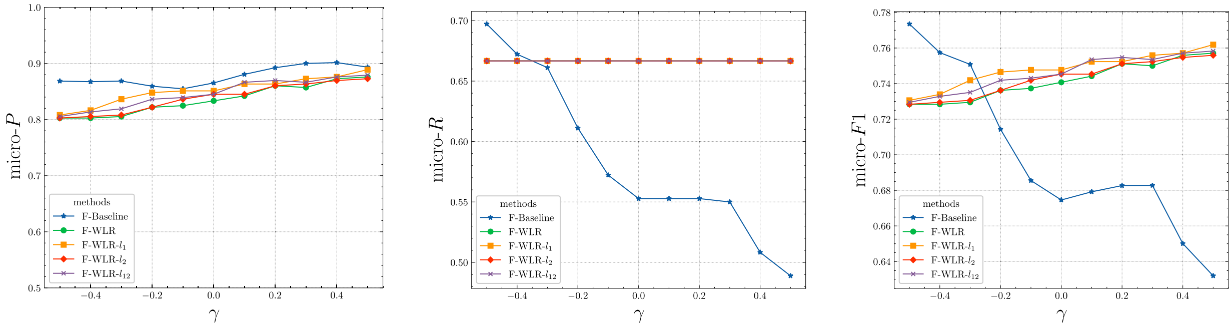
The three vertical dotted lines represented different classification hyperplanes. The samples between the three lines were considered to be hard-distinguishable. If $\gamma = 0$, obviously the FeLR reduced to classic logistic regression, and all samples had the same weight. Obviously, the class weight would dominantly impact the classification boundary (the green dotted line).

If $\gamma > 0$, the FeLR tended to learn a classification boundary that was more tolerant to the negative class with hard samples, of which the classification probability was close to 0.5. The degree of relative importance of hard samples would be greater than that of easy samples. Consequently, the sample weight might replace the class weight to dominantly impact the classification boundary. Since the classifier tended



(a) micro-P (b) micro-R (c) micro-F_1

FIGURE 5. Classification results w.r.t. the change of γ for the identification of PRPHs.



(a) micro-P (b) micro-R (c) micro-F_1

FIGURE 6. Classification results w.r.t. the change of γ for the identification of PUPH.

to classify more negative hard samples correctly, more positive hard samples would be misclassified (the yellow dotted line), in contrast to the green dotted line.

However, $\gamma < 0$ *reversed* the difference of relative weight between hard and easy samples. The results were that the relative weight of easy samples would be greater than that of hard samples. The classifier would be *out-of-focus* on negative hard samples, which might be ignored and the class weight would so dominantly impact the classification boundary. Due to the greater weight of positive class, the classification boundary would prefer to tolerate more positive hard samples (the blue dotted line). Naturally, the recall of positive class might improve when $\gamma < 0$, while more negative samples would be misclassified, yielding an increased $\overline{\text{FP}}$ (see Table 3 and Table 4).

C. ANTICIPATED DETERMINANTS OF PRPHs AND PUPHs

Before the FeLR was conducted, the anticipated determinants of PRPHs of PUPHs had been clarified. Although PPHs were identified with a total of 49 indicators based on 4 top-level perspectives, some indicators might have little effect on the identification results. According to some previous studies and our preliminary research, the determinants of PPHs might be a small subset of these variables (Table 5 and Table 6).

When it came to household demographic characteristics, the ages of the householders of PPHs did not differ from the sample to a large extent. The householders' education levels among PPHs were lower than or equal to middle school, and the mean education level of PRPHs and PUPHs (1.17 and 1.11) were lower than what the data was (1.32). Evidence showed that people with less formal education or work experiences were more likely to be living in poverty [31]. Less the quantity of workforce might exacerbate poverty [29], and all the PPHs had a labor force of no more than 3 people, whereas the maximum of this variable was 6 in the data. Similarly, the number of migrant workers could directly reflect the earning power, and in PRPHs and PUPHs, the mean of this variable (0.33 and 0.44) did not reach the overall mean level (1.2). DeNavas-Walt and Proctor [31] illustrated that a smaller proportion of the elderly ($\geq$ 65 years old) suffered from poverty compared to those who were younger (from 18 to 64 years old), which might be due to the improvement of the welfare of the elderly. But it did not seem to make an obvious difference on the number of elderly between the PPHs and the sample.

For basic needs, an intuitive feature of absolute poverty was the shortage of food or clothing [32]. Nevertheless, in the context of relative poverty, the basic needs of food or clothing did not show distinguishability on PPHs. In terms of the

TABLE 5. The descriptive statistics of numeric variables.

Variable	All of the sample				PRPHs				PUPHs			
	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
Household demographic characteristics												
Age of the householder	26	97	61.43	11.74	54	72	65.17	6.18	41	69	61.56	9.22
Number of family members	1	10	3.60	1.72	1	4	2.50	1.05	1	6	3.33	1.50
Number of the elders over 60	0	8	1.10	0.85	0	2	1.33	0.82	0	2	1.11	0.93
Number of the juveniles under 16	0	5	0.51	0.77	0	0	0	0	0	3	0.56	1.01
(∗,⋆) Quantity of workforce	0	6	2.27	1.39	0	2	0.83	0.75	0	3	2	1.12
Number of agricultural members	0	4	1.10	0.87	0	2	0.83	0.75	0	2	1.44	0.73
(∗,⋆) Number of migrant workers	0	5	1.20	1.12	0	1	0.33	0.52	0	2	0.44	0.73
Number of permanent residents	0	9	2.56	1.58	1	4	2.50	1.05	0	6	2.89	1.69
Basic needs												
Number of houses	0	4	1.08	0.37	1	1	1	0	0	2	1.11	0.60
Housing construction area	0	3000	121.32	125.68	80	240	145.50	61.66	80	200	127.56	30.90
(∗,⋆) Government subsidy funds for housing construction	0	200000	8196.28	16418.31	0	8500	4333.33	3855.73	0	20000	4111.11	7003.47
(∗,⋆) Self-raised funds for housing construction	0	400000	15832.27	42270.69	0	34500	8083.33	13328.23	0	200000	23888.89	66144.62
(∗,⋆) Debt for housing construction	0	350000	3803.27	17946.64	0	34500	5750	14084.57	0	70000	12000	23259.41
(∗,⋆) Number of critically ill patients	0	2	0.16	0.37	0	2	0.83	0.75	0	2	0.33	0.71
Self-raised funds for the treatment of critical illness	0	200000	1573.23	9526.45	0	94000	17481.83	37592.02	0	20955	2361.67	6973.21
Reimbursement ratio for the treatment of critical illness	0	100	10.49	27.59	0	100	40.73	45.87	0	56	8	18.76
Safe water & Domestic electricity & Radio signal												
Debt & Income												
(∗,⋆) Household business income	0	600000	3854.03	21250.98	0	1240	733.67	594.75	0	5936	2111.78	1738.11
(∗,⋆) Wage income	0	500000	28518.45	36882.60	0	45000	10524.17	17944.83	0	32000	9044.44	11393.76
(∗,⋆) Property income	0	1000000	2354.20	44492.28	0	1308	431.67	523.70	0	2200	244.44	733.33
(∗,⋆) Transfer income	0	156527.55	8069.28	13326.83	1812	35570.09	13765.68	12003.44	0	6000	1827.22	1960.61

Variables labelled with "∗" and "⋆" might be the determinants of PRPHs and PUPHs, respectively. The sample also contained 74 PHs.

TABLE 6. The descriptive statistics of categorical variables.

Variable	value(% of the sample)	value(% of the PRPHs)	value(% of the PUPHs)
Household demographic characteristics			
(∗,⋆) Education level of the householder. 1-Primary school or below; 2-Middle school; 3-Vocational school; 4-High school; 5-Bachelor or above	1(73.0),2(24.4),3(0.3),4(2.1),5(0.2)	1(83.3),2(16.7)	1(88.9),2(11.1)
Basic needs			
Lack of food. 0-No; 1-Yes	0(98.7),1(1.3)	0(100.0)	0(88.9),1(11.1)
Frequency of eating nutritious food. 1-During festivals; 2-Occasionally; 3-Often; 4-Frequently	1(0.8),2(6.5),3(17.8),4(74.8)	2(16.7),3(50),4(33.3)	2(22.2),3(44.4),4(33.3)
No worry about food. 0-No; 1-Yes	0(0.8),1(0.2)	1(100.0)	0(11.1),1(88.9)
Change clothes daily. 0-No; 1-Yes	0(0.5),1(99.5)	1(100.0)	1(100.0)
There are seasonal clothes throughout the year. 0-No; 1-Yes	0(0.3),1(99.7)	1(100.0)	1(100.0)
Main source of clothes. 1-Self-purchase; 2-Purchase by sons and daughters; 3-Donation; 4-Others	1(65.1),2(31.5),3(2.2),4(1.2)	1(50.0),2(33.3),3(16.7)	1(55.6),2(44.4)
No worry about wearing. 0-No; 1-Yes	0(0.8),1(99.2)	1(100.0)	1(100.0)
(⋆) House structure. 0-None; 1-Cave dwelling; 2-Adobe; 3-Bamboo; 4-Brick wood; 5-Brick; 6-RC; 7-Others	0(1.9),1(0.1),2(2.3),3(0.2),4(41.1), 5(39.3),6(7.4),7(7.6)	4(66.7),5(16.7),6(16.7)	0(11.1),2(33.3),4(55.6)
House is safe. 0-No; 1-Yes; 2-Unknown	0(2.5),1(94.0),2(3.6)	1(100.0)	0(33.3),1(66.7)
House meets the needs of production and life. 0-No; 1-Yes; 2-Unknown	0(2.1),1(94.9),2(3.0)	1(100.0)	0(33.3),1(66.7)
Houses for humans and livestock are separated. 0-No; 1-Yes; 2-Unknown	0(2.1),1(94.4),2(3.5)	1(100.0)	0(11.1),2(88.9)
The owner of the current house. 1-Self; 2-Sons and daughters; 3-Relatives; 4-Renting	1(96.2),2(1.1),3(1.4),4(1.3)	1(100.0)	1(100.0)
Enjoy the housing subsidy policies. 0-No; 1-Yes	0(41.9),1(58.1)	0(16.7),1(83.3)	0(66.7),1(33.3)
There are children in compulsory education. 0-No; 1-Yes	0(71.4),1(28.6)	0(100.0)	0(66.7),1(33.3)
(⋆) The type of medical insurance covered. 0-None; 1-One type; 2-Two types; 3-Three types; 4-Four types	0(2.2),1(69.1),2(18.2),3(5.3)	1(50.0),2(33.3),4(16.7)	0(22.2),1(55.6),2(22.2)
Level of disability. 0-None; 1-Level 1; 2-Level 2; 3-Level 3; 4-Level 4; 5-Level 5; 6-Level 6; 7-Level 7	0(84.1),1(3.2),2(7.2),3(4.4),4(0.9),6(0.1),7(0.1)	0(33.3),2(50.0),4(16.7)	0(77.8),1(11.1),4(11.1)
(∗,⋆) There are long-term chronic patients. 0-No; 1-Yes	0(58.4),1(41.6)	0(33.3),1(66.7)	0(11.1),1(88.9)
Type of chronic disease. 0-None; 1-ENT; 2-Cardiovascular; 3-Lung; 4-Liver; 5-Gastrointestinal; 6-Kidney; 7-Rheumatism; 8-Orthopedic; 9-Nervous system; 10-Cancer; 11-Others	0(35.8),1(1.4),2(21.7),3(7.6),4(2.4),5(4.0), 6(6.8),7(6.7),8(5.8),9(2.8),10(0.5)	0(16.7),2(16.7),3(16.7),7(16.7),9(16.7)	0(11.1),2(22.2),4(11.1),5(11.1), 6(11.1),8(11.1),9(11.1),10(11.1)
(∗,⋆) Able to afford medical expenses for chronic disease. 0-No; 1-Yes; 2-Unknown	0(4.1),1(56.9),2(39.0)	0(33.3),1(33.3),2(33.3)	0(22.2),1(66.7),2(11.1)
Type of critical illness. 0-None; 1-ENT; 2-Cardiovascular; 3-Lung; 4-Liver; 5-Gastrointestinal; 6-Kidney; 7-Rheumatism; 8-Orthopedic; 9-Nervous system; 10-Cancer; 11-Others	0(84.7),1(0.3),2(3.9),3(0.6),4(0.2),5(0.8), 6(1.3),7(0.3),8(2.2),9(1.2),10(2.5),11(2.1)	0(33.3),2(16.7),8(16.7),10(16.7),11(16.7)	0(77.8),6(11.1),10(11.1)
Safe water & Domestic electricity & Radio signal			
Guaranteed water source. 0-No; 1-Yes	0(1.9),1(98.1)	1(100.0)	0(11.1),1(88.9)
The water resource. 1-Tap water; 2-Well water; 3-Spring; 4-River; 5-Water tank; 6-Water cellar;	1(38.3),2(55.1),3(5.0),4(0.4),4(0.4),4(0.4)	1(33.3),2(66.7)	1(22.2),2(66.7),5(11.1)
Voltage is stable in daily life. 0-No; 1-Yes	0(1.4),1(98.6)	0(16.7),1(83.3)	1(100.0)
There is sound radio and television signal. 0-No; 1-Yes	1(1.0),2(99.0)	1(100.0)	0(11.1),1(88.9)
Debt & Income			
(∗) Type of debt. 0-None; 1-House construction; 2-Education; 3-Disease; 4-Others	0(78.7),1(10.8),2(1.4),3(5.8),4(3.3)	0(33.3),3(66.7)	0(55.6),1(22.2),2(11.1),3(11.1)
(∗) Type of creditor. 0-None; 1-Individual; 2-Financial institutions; 3-Others	0(79.2),1(17.0),2(3.4),3(0.4)	0(33.3),1(66.7)	0(55.6),1(33.3),2(11.1)
Primary income source. 1-Work; 2-Aquaculture; 3-Crop farming; 4-Self-management; 5-Relatives; 6-Government subsidies; 7-Others	1(37.9),2(11.3),3(36.2),4(2.0), 5(3.0),6(8.8),7(0.9)	2(33.3),3(16.7),6(50.0)	1(22.2),2(11.1),3(55.6),5(11.1)
(∗,⋆) Household income per capita exceeds 3300 CNY. 0-No; 1-Yes	0(4.7),1(95.3)	0(16.7),1(83.3)	0(66.7),1(33.3)

Variables labelled with "∗" and "⋆" might be the determinants of PRPHs and PUPHs, respectively. The sample also contained 74 PHs.

basic needs of houses, most households (88.4%) had 1 house. But the average quality of houses among PUPHs (2.89) was worse than that of the sample (4.65). The mean values of the government subsidy funds for housing construction among PRPHs and PUPHs (4333.33 and 4111.11 CNY) were lower than those of the sample (8196.28 CNY). The mean values of the debt for housing construction among PRPHs and PUPHs (5750 and 12000 CNY) were higher than the mean level of the data (3803.27 CNY), and Ntsalaze and Ikhide [28] also stated that high household indebtedness was a severe social scourge. All of the PUPHs did not have complete medical insurance, possibly out of cost concerns [33]. 41.6% of the households had long-term chronic patients, and this proportion was even higher in PRPHs and PUPHs (66.7% and 88.9%). Likewise, there were higher proportions among PRPHs and PUPHs (50% and 25%) that could not afford the medical expenses for chronic disease, while this proportion was only 6.7% of the sample, except for "Unknown". The major disease would bring a heavy financial burden to a household, and respectively, and PRPHs (66.7%) and PUPHs (22.2%) had 1 critically ill patient at least, which were higher than that in the sample (16.2%).

From the looks of it, "Safe water & Domestic electricity & Radio signal" did not show any clear difference among PRPHs, PUPHs and the sample, however, Sullivan [34] did illustrate the causal relationship between water, electricity and poverty, respectively. As for the debt, the main types of debt of PRPHs and PUPHs were different from those of the sample. Specifically, 66.7% of PRPHs were in debt due to diseases, while this proportion was only 5.8% in the sample. Furthermore, creditors of all disease-stricken PRPHs were individual, which implied that these PRPHs might be unable to repay the loan to financial institutions. The mean value of household business income, wage income, and property income of PRPHs and PUPHs were much lower than those of the sample (Table 5). This phenomenon was intuitive, whereas the mean transfer income of PRPHs was higher than that of the sample. Considering the variable "Primary income source", the transfer income of 50% of PRPHs was mainly from government subsidies, while no PUPH received subsidies from government, or a minor proportion did. The household income per capita was an important indicator to distinguish PPHs and others. Although those households whose income per capita was less than 3300 CNY only accounted for only 1.98% ($\frac{20}{1009} \times 100\%$) of the total, up to 5% ($\frac{1}{20} \times 100\%$), 5% ($\frac{1}{20} \times 100\%$), 60% ($\frac{12}{20} \times 100\%$) and 30% ($\frac{6}{20} \times 100\%$) of them were from PAHs, NPHs, PRPHs and PUPHs, respectively. While in all of PAHs, NPHs, PRPHs and PUPHs, those whose income per capita was less than 3300 CNY were accounted for 0.24% ($\frac{1}{421} \times 100\%$), 2.40% ($\frac{12}{499} \times 100\%$), 16.67% ($\frac{1}{6} \times 100\%$) and 66.67% ($\frac{6}{9} \times 100\%$), respectively.

In summary, the anticipated determinants of PRPHs and PUPHs did show similarities and differences as listed below, starting with similarities:

1) a lower education level, which limited their job options [31];
2) fewer migrant workers and lower the government subsidy funds for housing construction, which directly capped their income [29];
3) higher debts for housing construction and diseases, which added to their financial burden [28];
4) a higher percentage of long-term chronic patients, and a higher percentage of inability to afford the medical expenses;
5) a higher percentage of living with a minimum of 1 critically ill patient;
6) lower incomes from businesses, wages and properties;
7) a higher percentage of a lower income per capita (less than 3300 CNY annually).

In terms of differences, the average house quality among PRPHs was not significantly lower than that of the samples, quite the opposite for PUPHs. In fact, house quality was not showing dynamic changes, so the local government believed that PRPHs should be PAHs. However, it was clear that the local government had left out PUPHs and considered this group of households as NPHs. This was confirmed in the exchanges with the local government. Therefore, the average housing quality did not account for the difference between PAHs and PRPHs as expected, but rather for the difference between NPHs and PUPHs. Although the PRPHs had a higher mean transfer income than that of PUPHs, the transfer income of 50% of PRPHs mainly came from government subsidies, while no PUPHs received subsidies from government. This was because the local government tilted the subsidies to the poverty households as much as possible. Giving subsidies to poverty households did serve as a direct way out of poverty, thus the local government believed that subsidized PRPH should be PAH. This was because the local government gave subsidy preferences to the poverty households as much as possible, which served as a direct way to alleviate poverty, and the local government believed that the subsidized PRPHs were PAHs. However, some poverty households were left out (i.e. PUPHs), and the subsidies were actually not given to them.

D. DETERMINANTS OF PRPHs GENERATED BY FeLR

F-WLR-ℓ_1 recalled all the PRPHs, and the micro-P was also high (81.36%). $D_1 = 254.639$ ($p = 0.939$) showed the high goodness-of-fit of the full model. $D_2 = 1691.064$ ($p = 0.000$) showed that the full model was significantly different from the null model, and explanatory variables ($\boldsymbol{\beta}_1$) or a subset of them significantly contributed to predicting PRPHs. The value of R^2_{mf}, R^2_{cs} and R^2_{nk} were 0.869, 0.981 and 0.991, respectively, which also indicated the good fit.

A total of 9 indicators from the 3 perspectives significantly influenced re-poverty (Table 7). The education level of the householder did not show the significant influence on re-poverty, which differed from the anticipation. That apart, those households with less earners might be at a higher risk of

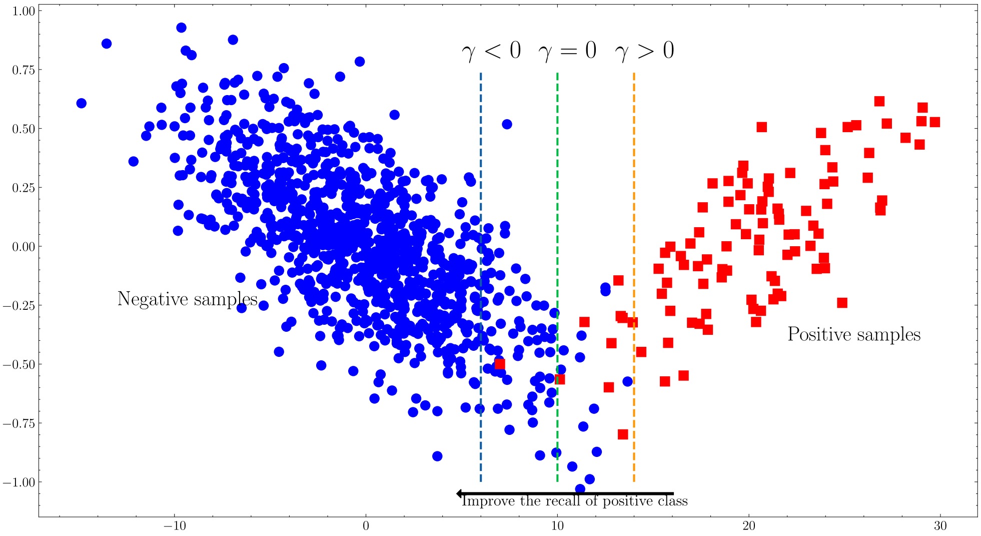


FIGURE 7. **Classification boundary w.r.t. the change of gamma.**

falling into poverty again. The results were consistent with the anticipated inference and similar to the previous study [29].

Back to basic needs. The lack of government subsidy funds for housing construction led to a significantly higher risk of re-poverty. Subsidies to households would directly alleviate the re-poverty, which went in line with the anticipation and the study [35]. Households with more critically ill patients were more likely to fall back to poverty. Evidence showed that disease was closely related to poverty [36]. Along with more self-raised funds against critical illness might come heavier debt, which could be also illustrated by the debt from disease. A household that could not afford the medical expenses for chronic diseases had a poor financial resilience, because the chronic disease treatment entailed long-term expenditures. The result was similar to some previous studies [37], and Liu *et al.* [38] reported that "higher medical expenditures have become an important cause of poverty in rural China, increasing the number of the poor households by 44.3%".

There were relatively low risks of re-poverty to those who were not in debt, which was in accord with the study [28]. Higher household income per capita was followed by a lower risk of re-poverty. A similar study [39] showed that the "participation in supermarket channels was associated with a 48% gain in average household income", shedding a light on poverty reduction. Due to the instability of household income per capita among PRPHs, they might be neglected by the local government, which was more of the case for PUPHs.

E. DETERMINANTS OF PUPHs GENERATED BY FeLR

The micro-R and micro-P of F-WLR-ℓ_1 were 66.67% and 80.81%, respectively. 3 particular PUPHs could not be recalled by F-WLR-ℓ_1, as well as in most other models. $D_1 = 296.634$ ($p = 0.998$) showed the high goodness-of-fit of the full model. $D_2 = 1882.697$ ($p = 0.000$) showed that the full model was significantly different with the null model, and explanatory variables ($\boldsymbol{\beta}_1$) or a subset of them significantly contributed to predicting PUPHs. The value of R^2_{mf}, R^2_{cs} and R^2_{nk} were 0.864, 0.976 and 0.989, respectively, which also indicated the good fit.

In spite of some differences between the estimates of PUPHs and PRPHs (Table 8), the indicators that significantly contributed to predicting PUPHs were also intuitive. The number of migrant workers and the number of permanent residents affected the prediction of PUPHs quite the opposite. The migrant workers and permanent residents of a household should be viewed in terms of *income* and *expenditure*. The more the migrant workers, the higher the income. The more the permanent residents, the greater the financial burden. The result indicated that jobs were critical to reducing poverty, and Bell and Newitt [40] suggested decent work should be integrated into the international development agenda to eradicate poverty. A higher education level might promise stabler earning skills or jobs, thus bringing down the risk of falling into poverty, which has gained wide recognition [41]. Having one inferior house or not even one could substantially contribute to poverty, which accorded with Hanratty's work [42] that illustrated that poverty rate had strong positive impacts on homelessness. Households lacking adequate medical insurances were easily ignored by local policymakers, whereas those with all four types of medical insurances were less susceptible to poverty, which was inconsistent with the study [43].

Those households without the disabled held less possibilities to become PUPHs; on the contrary, households with seriously disabled people were more likely to be PUPHs. The result tallied with the study [44]. Likewise, the number of critically ill patients also contributed to predicting PUPHs. Among PUPHs, up to 66.67% of them were below the absolute poverty line (i.e., 3300 CNY per capita income), and they were inevitably neglected by the local government.

TABLE 7. Estimates of F-WLR-ℓ_1 for the identifications of PRPHs.

Variables	β	SE	Sig.	exp(β)
Household demographic characteristics				
Quantity of workforce	-0.189	0.070	0.007***	0.828
Basic needs				
Government subsidy funds for housing construction	-0.170	0.057	0.003***	0.844
Number of critically ill patients	0.247	0.106	0.020**	1.280
Self-raised funds for the treatment of critical illness	0.288	0.113	0.011**	1.330
Be able to afford medical expenses for chronic disease	-0.607	0.072	0.000***	0.545
Type of critical illness-Others	0.215	0.106	0.043**	1.240
Debt & Income				
Type of debt-None	-0.183	0.063	0.004***	0.833
Type of debt-Disease	0.266	0.047	0.000***	1.310
Source of debt-None	-0.185	0.064	0.004***	0.831
Source of debt-Individual	0.211	0.054	0.000***	1.240
Household income per capita exceeds 3300 CNY	-0.421	0.084	0.000***	0.656
Constant=-1.425 $D_1 = 254.639$ ($p = 0.939$) $D_2 = 1691.064$ ($p = 0.000$)	$R^2_{mf} = 0.869$	$R^2_{cs} = 0.981$	$R^2_{nk} = 0.991$	

*, **, *** indicate statistical significance at the 10%, 5%, and 1% level. It does not show those statistical insignificant variables with smaller weights.

TABLE 8. Estimates of F-WLR-ℓ_1 for the identification PUPH.

Variables	β	SE	Sig.	exp(β)
Household demographic characteristics				
Number of migrant workers	-0.181	0.026	0.000***	0.834
Number of permanent residents	0.168	0.043	0.000***	1.180
Education level of the householder-Vocational school	-0.173	0.049	0.000***	0.841
Education level of the householder-Bachelor or above	-0.170	0.064	0.008***	0.844
Basic needs				
House structure-None	0.324	0.111	0.004***	1.380
House structure-Adobe	0.187	0.040	0.000***	1.210
House structure-Brick	-0.186	0.039	0.000***	0.831
The type of medical insurance covered-Four types	-0.217	0.063	0.001***	0.805
Level of disability-None	-0.307	0.095	0.001***	0.736
Level of disability-Level 1	0.378	0.135	0.005***	1.460
Number of critically ill patients	0.208	0.089	0.020**	1.230
Type of chronic disease-Cancer	0.279	0.099	0.005***	1.320
Type of chronic disease-Nervous system	0.241	0.091	0.008***	1.270
Debt & Income				
Household income per capita exceeds 3300 CNY	-0.736	0.053	0.000***	0.479
Constant=-1.439 $D_1 = 296.654$ ($p = 0.998$) $D_2 = 1882.697$ ($p = 0.000$)	$R^2_{mf} = 0.864$	$R^2_{cs} = 0.976$	$R^2_{nk} = 0.989$	

*, **, *** indicate statistical significance at the 10%, 5%, and 1% level. It does not show those statistical insignificant variables with smaller weights.

However, the micro-R of most methods was fixed as 0.667 regardless the change of γ, due to three outliers in PUPHs (Fig. 4(b)). The prediction probabilities of the three particular PUPHs were close to 0, thus misclassified to NPHs in all the different train-test-splits. Those households who had critically ill patients, or the disabled were not necessarily PUPHs, because 10.02% of NPHs had critically ill patients, and 10.42% of NPHs had disabled people, which matched analysis of most other variables (Fig. 8). The household incomes per capita of all the three misclassified samples exceeded 3300 CNY. Not only did the household income per capita directly determine the absolute poverty, but it was a primary indicator to reflect the relative poverty. There was a logic that the three samples did not fall into the scope of PUPHs, but NPHs. That said, the household income per capita did not necessarily determine a PUPH, because there were still 2.4% of NPHs with income per capita less than 3300 CNY. Local policymakers are suggested to keep close tabs on not only the household income per capita, but the remaining indicators (Fig. 8).

F. COMPARISON OF DETERMINANTS ON PRPHs AND PUPHs OF BOTH FeLR AND BASELINE

Although the classification performance of Baseline method (i.e. traditional logistic regression) was not as effective as FeLR, it was necessary to analyze the differences between the determinants generated by Baseline and FeLR.

For PRPH identification, the micro-R and micro-P of Baseline were 44.58% and 79.85%, respectively (Table 3). $D_1 = 214.171$ ($p = 1.000$) showed the high goodness-of-fit of the full model, however, $D_2 = 11.633$ ($p = 1.000$) showed that there was no significantly difference between the full model and the null model (Table 9). The value of R^2_{mf}, R^2_{cs} and

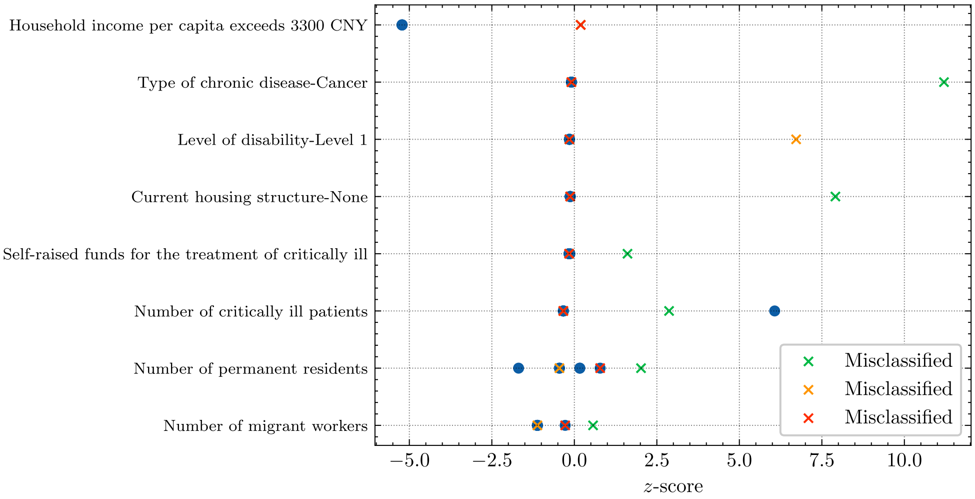


FIGURE 8. ***z*-score of PUPHs on 8 variables (three samples labeled with "×" are misclassified to NPH).**

TABLE 9. Estimates of baseline for the identifications of PRPHs.

Variables	β	SE	Sig.	exp(β)
Household demographic characteristics				
Number of family members	-0.0899	0.0152	0.000***	0.914
Basic needs				
Number of critically ill patients	0.498	0.18	0.006***	1.645
Self-raised funds for the treatment of critical illness	0.172	0.0402	0.000***	1.188
Be able to afford medical expenses for chronic disease	0.179	0.0239	0.000***	1.196
Type of critical illness-Cancer	-0.134	0.0386	0.001***	0.875
Type of critical illness-Cardiovascular	-0.189	0.0676	0.005***	0.828
Type of critical illness-Liver	-0.0913	0.0134	0.000***	0.913
Type of critical illness-Nervous system	-0.0954	0.038	0.012**	0.909
Constant=-1.26 $D_1 = 214.171$ ($p = 1.000$) $D_2 = 11.633$ ($p = 1.000$)	$R^2_{mf} = 0.052$	$R^2_{cs} = 0.027$	$R^2_{nk} = 0.066$	

*, **, *** indicate statistical significance at the 10%, 5%, and 1% level. It does not show those statistical insignificant variables with smaller weights.

R^2_{nk} were 0.052, 0.027 and 0.066, respectively, which also showed the bad fit of Baseline.

The determinants of PRPHs generated by Baseline and FeLR were very different (Table 9 and Table 7). According to the classification results of Baseline, the number of family members was a determinant. Although the max number of family members in all PRPHs was 4, while that among all of the sample is 10 (Table 5), 71.26% (i.e. $\frac{300}{421}$) of PAHs had less than or equal to 4 family members. Therefore, the fact that the number of family members was a determinant of PRPHs could not be reasonably explained. Similar to the estimates of F-WLR-ℓ_1 (Table 7), Baseline also revealed that the number of critically ill patients and the self-raised funds for the treatment of critical illness were the determinants of PRPHs. In contrast, Baseline indicated that a household was more likely to be PRPH, if they could afford medical expenses for chronic disease, which did not make common sense. It was also a contradiction that the coefficients of several disease types were negative, while that of the number of critically ill patients was positive. In addition, Baseline did not show the household income per capita as a determinant of PRPH. These unexplained factors also suggested that the Baseline was not suitable for identifying PRPH.

For PUPH identification, the micro-*R* and micro-*P* of Baseline were 55.28% and 85.62%, respectively (Table 4). $D_1 = 253.655$ ($p = 1.000$) showed the high goodness-of-fit of the full model, but similar to PRPH identification, $D_2 = 21.064$ ($p = 1.000$) showed that there was no significantly difference between the full model and the null model (Table 10). The value of R^2_{mf}, R^2_{cs} and R^2_{nk} were 0.077, 0.041 and 0.097, respectively, which also showed the bad fit of Baseline.

There were some differences between the determinants of PUPHs generated by Baseline and FeLR (Table 10 and Table 8). For example, Baseline did not reveal the number of migrant workers and the basic medical insurance as the determinants of PUPH. In all of the sample, there were only 8 households used river or water tank as the water source, but due to the poor fit, Baseline still indicated the water source as the determinants of PUPH. The coefficients of identifying PUPH had the same contradiction as that of identifying PRPH, i.e., the coefficients of several disease types were negative, while that of the number of critically ill patients was

TABLE 10. Estimates of baseline for the identifications of PUPHs.

Variables	β	SE	Sig.	exp(β)
Household demographic characteristics				
Education level of the householder-Bachelor or above	-0.085	0.0092	0.000***	0.919
Basic needs				
Number of critically ill patients	0.867	0.285	0.002***	2.38
Frequency of eating nourishing food-During festivals	-0.138	0.0267	0.000***	0.871
House structure-None	0.113	0.0223	0.000***	1.12
Level of disability-Level 4	0.0916	0.0148**	0.000***	1.1
Type of chronic disease-Cancer	0.0956	0.0197	0.000***	1.1
Type of critical illness-Cardiovascular	-0.124	0.0607	0.041**	0.883
Type of critical illness-Gastrointestinal	-0.081	0.0322	0.012**	0.922
Type of critical illness-Rheumatism	-0.0984	0.0328	0.003***	0.906
Type of critical illness-Orthopedic	-0.128	0.0562	0.023**	0.88
Type of critical illness-Nervous system	-0.128	0.0462	0.006***	0.88
Type of critical illness-Cancer	-0.16	0.0819	0.051*	0.852
Type of critical illness-Others	-0.169	0.0755	0.025**	0.845
The water resource-River	-0.087	0.0123	0.000***	0.917
The water resource-Water tank	0.0926	0.00914	0.000***	1.1
Debt & Income				
Household income per capita exceeds 3300 CNY	-0.220	0.0754	0.004	0.803
Constant=-1.25 $D_1 = 253.665$ ($p = 1.000$) $D_2 = 21.064$ ($p = 1.000$)	$R^2_{mf} = 0.077$	$R^2_{cs} = 0.041$	$R^2_{nk} = 0.097$	

*, **, *** indicate statistical significance at the 10%, 5%, and 1% level. It does not show those statistical insignificant variables with smaller weights.

positive. Additionally, the determinants of PUPH generated by Baseline and that generated by FeLR showed some similarities. Baseline revealed that house structure and household income per capita affected the occurrence of PUPH as well, but the absolute values of these two coefficients fitted by FeLR were larger (Table 8 and Table 10), suggesting that FeLR expressed the determinants of PUPH more accurately.

V. CONCLUSION

This paper studied the rare potential poor household (PPH) identification from the perspective of social computing. PPHs, comprised of non-poor households (NPHs) and poverty-alleviated households (PAHs), were stricken with a higher risk of re-poverty. Since PPHs were rare and hard-distinguishable, a focus embedded logistic regression (FeLR) was designed with a view to the identification. By extending the sample weight exponent to negative values, FeLR overlooked negative samples that were hard-distinguishable. Consequently, the recall of PRPHs was significantly improved compared with those of traditional logistic regression, weighted logistic regression and their varieties. Furthermore, FeLR not only showed its excellent fitting performance, but also picked out more reasonable crucial factors affecting potential poverty, compared with that of traditional logistic regression. Unimproved recall of PUPH was due to 3 outliers, whose values on "household income per capita" exceeded the absolute poverty line-3300 CNY. 9 indicators significantly contributed to the incidence of the PRPH and PUPH respectively, of which "medical expenses for chronic disease" and "house structure" stood out respectively, and "household income per capita" jointly.

In summary, compared with traditional logistic regression and other methods used in this paper, FeLR has the following advantages:

1) FeLR is a generalization of traditional logistic regression. By setting equal class weights and sample weight exponent-γ to 0, FeLR degenerates into basic logistic regression.
2) FeLR is more suitable to handle rare event classification and hard-distinguished problem (i.e., PPH classification in this paper). The classification accuracy can be improved by setting γ to positive, and the recall can be improved by setting γ to negative (Fig. 7).
3) The results of various hypothesis tests indicate that the FeLR method has a higher goodness-of-fit for PPH classification.
4) Compared with traditional method, FeLR picked out more accurate and crucial factors affecting potential poverty.

To keep up the anti-poverty efforts, local policymakers are suggested to pay attention to those households 1) Who still struggle in absolute poverty (i.e., the income per capita was below the poverty line); 2) Who have critically ill patients but did not have adequate access to medical insurance; 3) Who are in debt due to disease; 4) Who are stricken with serious disability; or 5) Who have no house, or an inferior house. An early warning mechanism is recommended to predict the potential poverty, and improve the effectiveness of anti-poverty.

There were some limitations and shortcomings in this paper. Compared with other related literature, there were too many indicators used in this paper, and the correlation between these indicators was not tested in this paper, so multicollinearity might arise between indicators. In addition, for discrete variables such as housing structure, chronic disease type, and disease type, some of the values taken might be too few, and we used one-hot encoding to produce data that might be sparse, and these values should probably be combined and thus simplify the model. Further research shall focus on two things: 1) Shore up potential poverty investigation. The analytic framework of potential poverty will be constructed to systematically evaluate the control strategies against potential poverty. 2) Try to refine indicators that hold a causal relationship with potential poverty. With that, multicollinearity, endogeneity and heteroscedasticity caused by outliers of the data will be analyzed.

APPENDIX A
CRITERIA FOR MANUALLY DETERMINING THAT POOR HOUSEHOLDS HAVE BEEN LIFTED OUT OF POVERTY

Poverty alleviation was based on the rural household registration population as a unit. In accordance with *"Thirteenth Five-Year" Poverty Alleviation Plan* proposed by State Council of the People's Republic of China, the criteria used to manually determine the households out of poverty were mainly as follows:

1) Household income per capita. The poverty-alleviated household income per capita exceeds 3300 CNY annually.
2) Adequate food. The households have the ability to meet their basic food needs and supply certain nutritious foods such as meat, eggs, and soy products to a point of self-sufficiency.
3) Adequate clothing. The households can afford their own clothes or use their relatives' help to gain access to seasonal clothes..
4) Guaranteed compulsory education. Children who have reached the age of six are sent to school by their parents or other legal guardians to receive and complete their compulsory education.
5) Guaranteed medical insurance. The households with seriously ill patients receive a proper amount of reimbursement.
6) Guaranteed housing security. Their houses meet the corresponding standards set by the Housing and Construction Department.
7) Guaranteed water source. The households have easy access to domestic water.
8) Guaranteed electricity. Living electricity in the households meets the demand of lighting and basic home appliances.
9) Guaranteed Sound radio and television signal. The households have broadcast television at home or are able to pick up broadcast television signal.

In this paper, in addition to the above-mentioned criteria, we also used indicators such as demographic characteristics and various income characteristics.

APPENDIX B
METHOD DESCRIPTION

Since the PRPH only existed in PAH, and the PUPH only existed in NPH, naturally they were treated as two independent binary classification problems. Each of the household was naturally treated as a single outcome variable Y_i ($i = 1, \ldots, n$) following a Bernoulli probability function that took on the value 1 with probability π_i, and 0 with the probability $1 - \pi_i$. For the identification of PRPHs, $Y_i = 1$ or 0 meant that the household was PRPH or PAH. For that of PUPHs, $Y_i = 1$ or 0 meant that the household was PUPH or NPH. The prediction probability of PPH π_i was given by

$$\pi_i = \frac{1}{1 + \exp(-\mathbf{x}_i \boldsymbol{\beta})}. \tag{3}$$

The Bernoulli had probability function $P(Y_i = y_i|\pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$. The unknown parameter $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^k$, where β_0 was a scalar constant term, and $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_{k-1})^{\mathrm{T}} \in \mathbb{R}^{(k-1)}$ was a vector whose elements corresponded to explanatory variables. The parameters were estimated by maximum likelihood (MLE). By taking logs on likelihood function and using Eq. (3), the log-likelihood was simplified as

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \log(\pi_i) + (1 - Y_i)\log(1 - \pi_i). \tag{4}$$

Due to the rarity of PRPH and PUPH, traditional logistic regression might not be satisfactory. A weighting strategy compensated for differences in the sample ($\bar{y}$) and population (τ) fractions of ones induced by choice-based sampling. Instead of maximizing Eq. (4), weighting technique maximized weighted log-likelihood as

$$\log L_w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_1 Y_i \log(\pi_i) + w_0(1 - Y_i)\log(1 - \pi_i), \tag{5}$$

in which $w_1 = \frac{\tau}{\bar{y}}$, and $w_0 = \frac{1-\tau}{1-\bar{y}}$.

To ensure good interpretability, it was necessary to do variable selection for the identification of PRPH and PUPH, and ℓ_1 penalization selected the important variables. The penalized likelihood regression imposed a penalty to the likelihood, which essentially was to add a prior information on $\boldsymbol{\beta}$:

$$\log L_p(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) + A(\boldsymbol{\beta}). \tag{6}$$

Four typical penalized formats were listed as follows.

1) Lasso (ℓ_1): $A(\boldsymbol{\beta}) = -\lambda_1 \sum_{i=1}^{k-1} |\beta_i|$. With Laplacian prior on $\boldsymbol{\beta}$ to perform variable selection;
2) Ridge (ℓ_2): $A(\boldsymbol{\beta}) = -\lambda_2 \sum_{i=1}^{k-1} \beta_i^2$. With Gaussian prior on $\boldsymbol{\beta}$ to avoid extreme estimates and stabilize variance;
3) ElasticNet (ℓ_{12}): $A(\boldsymbol{\beta}) = -\lambda_{12}(\sum_{i=1}^{k-1} \rho|\beta_i| + (1 - \rho)\beta_i^2)$. With both Gaussian and Laplacian prior on $\boldsymbol{\beta}_1$ to integrate the two components;
4) Firth-type (ℓ_{Firth}): $A(\boldsymbol{\beta}) = \frac{1}{2}\log\det(I(\boldsymbol{\beta}))$. With Jeffreys prior on $\boldsymbol{\beta}$ to correct small-sample bias in $\boldsymbol{\beta}$. $I(\boldsymbol{\beta})$ is the Fisher information matrix, and $\det(\cdot)$ is the determinant of a matrix.

Parameters $\lambda_1, \lambda_2, \lambda_{12} > 0$ were the coefficients of corresponding penalized items, and $\rho \in [0, 1]$ in ElasticNet adjusted the two components. Specifically, if $\rho = 0$, ElasticNet would be reduced to the Ridge, and if $\rho = 1$, it would be reduced to the Lasso. Via adding penalized item on Eq. (5), the weighted penalized log-likelihood was given by

$$\log L_{pw}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_1 Y_i \log(\pi_i) + w_0(1-Y_i)\log(1-\pi_i) + A(\boldsymbol{\beta}). \qquad (7)$$

The weighting technique only tackled the rarity of PRPHs and PUPHs. However, there was not only imbalanced class, but also hard-distinguishability to deal with. Hard samples were usually located near the classification boundary, whose prediction probabilities π_i were around 0.5. They might exist in not only PRPH/PUPH, but also PAH/NPH. The focal loss technique worked with logistic regression to address this problem, hence the focus embedded logistic regression (FeLR) was designed. It originated from computer vision domain to address the hard-distinguishability problem in object detection. The focal loss was given by

$$\text{Loss}^F(\boldsymbol{\beta}) = -\sum_{i=1}^{n}(1-\pi_i)^\gamma Y_i \log \pi_i + \pi_i^\gamma (1-Y_i) \times \log(1-\pi_i), \qquad (8)$$

where the sample weight exponent γ adjusted the relative importance of hard and easy samples.

If $\gamma > 0$, Eq. (8) would focus on hard samples by elevating the corresponding losses, while reducing those of easy samples. Specifically, $\gamma = 0$ meant the model with no focus embedded. Because the loss function was the negative of log-likelihood, the focus embedded log-likelihood was derived as

$$L^F(\boldsymbol{\beta}) = \sum_{i=1}^{n}(1-\pi_i)^\gamma Y_i \log \pi_i + \pi_i^\gamma (1-Y_i)\log(1-\pi_i). \qquad (9)$$

Focus embedded log-likelihood could be treated as a generalization of the traditional type assigning different weights on different samples. Combined with the class weight technique, the focus embedded weighted log-likelihood was given by

$$L_w^F(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_1(1-\pi_i)^\gamma Y_i \log \pi_i + w_0 \pi_i^\gamma (1-Y_i) \times \log(1-\pi_i). \qquad (10)$$

Using logistic function (Eq. (3)), the estimate of $\boldsymbol{\beta}$ of FeLR with penalized item was to solve

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{\text{argmax}}\ L_w^F(\boldsymbol{\beta}) + A(\boldsymbol{\beta}) \\ &= \underset{\boldsymbol{\beta}}{\text{argmax}}\ \sum_{i=1}^{n}[w_1(1-\pi_i)^\gamma Y_i \log \pi_i \\ &\quad + w_0 \pi_i^\gamma (1-Y_i)\log(1-\pi_i)] + A(\boldsymbol{\beta}) \\ &= \underset{\boldsymbol{\beta}}{\text{argmax}}\ \sum_{i=1}^{n}\{w_1 Y_i[1+\exp(\mathbf{x}_i\boldsymbol{\beta})]^{-\gamma}\log[1+\exp(-\mathbf{x}_i\boldsymbol{\beta})] \\ &\quad - w_0(1-Y_i)[1+\exp(-\mathbf{x}_i\boldsymbol{\beta})]^{\gamma}\log[1+\exp(\mathbf{x}_i\boldsymbol{\beta})]\} \\ &\quad + A(\boldsymbol{\beta}). \end{aligned} \qquad (11)$$

APPENDIX C
EVALUATION MEASURES DESCRIPTION

Since PRPHs and PUPHs were rare, leave-one-out (LOO) and cross-validation (CV) techniques were used to divide the samples into training (in-sample) data and testing (out-of-sample) data. For the identification of PRPHs and PUPHs, LOO process was repeated 240 times and 360 times respectively, due to the different number of positive samples. In each train-test-split, the model with different $\hat{\boldsymbol{\beta}}$ learned from different training data was evaluated on testing data, yielding a confusion matrix, and then the sum of different confusion matrices (Table 11).

TABLE 11. Sum of confusion matrices of different train-test-splits.

Actual Class	Prediction Class	
	Positive	Negative
Positive	$\overline{\text{TP}}$ (Sum of True Positive)	$\overline{\text{FN}}$ (Sum of False Negative)
Negative	$\overline{\text{FP}}$ (Sum of False Positive)	$\overline{\text{TN}}$ (Sum of True Negative)

Naturally, for the identification of PRPHs, $\overline{\text{TP}}+\overline{\text{FN}} = \overline{\text{FP}}+\overline{\text{TN}} = 240$, which was 360 for PUPHs. Three measures were adopted for estimating global performance of the classifier, namely micro-precision (micro-P), micro-recall (micro-R), and micro-F_1 score:

$$\begin{aligned}\text{micro-}P &= \frac{\overline{\text{TP}}}{\overline{\text{TP}}+\overline{\text{FP}}}, \\ \text{micro-}R &= \frac{\overline{\text{TP}}}{\overline{\text{TP}}+\overline{\text{FN}}}, \\ \text{micro-}F_1 &= \frac{2\cdot\text{micro-}P\cdot\text{micro-}R}{\text{micro-}P+\text{micro-}R}. \end{aligned} \qquad (12)$$

From the statistical point of view, 5 indices were adopted to test the goodness-of-fit, including deviance D_1, D_2, and three statistical pseudo R^2, namely McFaddden's R^2 (R^2_{mf}), Cox-Snell's R^2 (R^2_{cs}), and Nagelkerke's R^2 (R^2_{nk}):

$$\begin{aligned} D_1 &= -2(\log L_F - \log L_S) = -2\log L_F \sim \chi^2(n-k), \\ D_2 &= -2(\log L_0 - \log L_F) \sim \chi^2(k-1), \\ R^2_{\text{mf}} &= 1 - \frac{\log L_F}{\log L_0}, \\ R^2_{\text{cs}} &= 1 - \left(\frac{L_0}{L_F}\right)^{\frac{2}{n}} = 1 - \exp[\frac{2}{n}(\log L_0 - \log L_F)], \\ R^2_{\text{nk}} &= \frac{1-\left(\frac{L_0}{L_F}\right)^{\frac{2}{n}}}{1-L_0^{\frac{2}{n}}} = \frac{1-\exp[\frac{2}{n}(\log L_0 - \log L_F)]}{1-\exp[\frac{2}{n}\log L_0]}, \end{aligned} \qquad (13)$$

where L_F and L_0 were the likelihood functions for the full model and intercept-only model, respectively, and L_S was the saturated model with $\log L_S = 0$. Naturally, a smaller D_1 indicated a better fit and the full model was closer to the saturated model, whereas a bigger D_2 represented a better fit and the independent variables were more interpretable. Due to CV, the mean values of these indices in different models were finally used.

ACKNOWLEDGMENT

The authors would like to thank those who provided constructive comments and feedback to help improve this paper and also would like to thank Leon, from Guangzhou, Guangdong, China, for editing the English text of the draft of this manuscript.

REFERENCES

[1] A. Mani, S. Mullainathan, E. Shafir, and J. Zhao, "Poverty impedes cognitive function," *Science*, vol. 341, no. 6149, pp. 976–980, Aug. 2013.

[2] J. Haushofer and E. Fehr, "On the psychology of poverty," *Science*, vol. 344, no. 6186, pp. 862–867, May 2014.

[3] J.-Y. Wang, Y.-F. Chen, and M.-C. Yan, "Research on the targeted measures of poverty alleviation and its innovative ways in China," *Bull. Chin. Acad. Sci.*, vol. 31, no. 3, pp. 289–295, 2016.

[4] Q. Deng, E. Li, and P. Zhang, "Livelihood sustainability and dynamic mechanisms of rural households out of poverty: An empirical analysis of Hua County, Henan Province, China," *Habitat Int.*, vol. 99, May 2020, Art. no. 102160.

[5] S. Murphy and P. P. Walsh, "Social protection beyond the bottom billion," IZA Discuss. Paper, IZA-Inst. Labor Econ., Bonn, Germany, Tech. Rep. 8376, 2014, pp. 1–29. [Online]. Available: https://ssrn.com/abstract=2481565

[6] Y.-Y. Yang, Y.-S. Liu, and Z.-W. Zhang, "Study on policy innovation and suggestions of targeted poverty alleviation based on typical investigation," *Bull. Chin. Acad. Sci.*, vol. 31, no. 3, pp. 337–345, 2016.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[8] B. Halleröd, "Sour grapes: Relative deprivation, adaptive preferences and the measurement of poverty," *J. Social Policy*, vol. 35, no. 3, pp. 371–390, Jul. 2006.

[9] P. Annoni, R. Bruggemann, and L. Carlsen, "A multidimensional view on poverty in the European union by partial order theory," *J. Appl. Statist.*, vol. 42, no. 3, pp. 535–554, Mar. 2015.

[10] A. B. Atkinson, "Multidimensional deprivation: Contrasting social welfare and counting approaches," *J. Econ. Inequality*, vol. 1, no. 1, pp. 51–65, 2003.

[11] J. Xu, J. Song, B. Li, D. Liu, D. Wei, and X. Cao, "Do settlements isolation and land use changes affect poverty? Evidence from a mountainous province of China," *J. Rural Stud.*, vol. 76, pp. 163–172, May 2020.

[12] R. Pathak, C. K. Wyczalkowski, and X. Huang, "Public transit access and the changing spatial distribution of poverty," *Regional Sci. Urban Econ.*, vol. 66, pp. 198–212, Sep. 2017.

[13] F. W. Agbola, A. Acupan, and A. Mahmood, "Does microfinance reduce poverty? New evidence from Northeastern Mindanao, the Philippines," *J. Rural Stud.*, vol. 50, pp. 159–171, Feb. 2017.

[14] C. Liao and D. Fei, "Poverty reduction through photovoltaic-based development intervention in China: Potentials and constraints," *World Develop.*, vol. 122, pp. 1–10, Oct. 2019.

[15] M. G. Palmer, N. T. M. Thuy, Q. T. N. Quyen, D. S. Duy, H. V. Huynh, and H. L. Berry, "Disability measures as an indicator of poverty: A case study from Viet Nam," *J. Int. Develop.*, vol. 24, pp. 53–68, Jan. 2012.

[16] T. Pasanen, "Multidimensional poverty in laos: Analysis on household and village levels," *J. Int. Develop.*, vol. 29, no. 6, pp. 714–728, Aug. 2017.

[17] L. McBride and A. Nichols, "Retooling poverty targeting using out-of-sample validation and machine learning," *World Bank Econ. Rev.*, vol. 32, no. 3, pp. 531–550, 2018.

[18] X.-Z. Zhao, B.-L. Yu, Y. Liu, Z.-Q. Chen, Q.-X. Li, C.-X. Wang, and J.-P. Wu, "Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh," *Remote Sens.*, vol. 11, no. 4, pp. 1–18, 2019.

[19] T. P. Sohnesen and N. Stender, "Is random forest a superior methodology for predicting poverty? An empirical assessment," *Poverty Public Policy*, vol. 9, no. 1, pp. 118–133, Mar. 2017.

[20] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, vol. 350, no. 6264, pp. 1073–1076, Nov. 2015.

[21] R. Puurbalanta, "A clipped Gaussian geo-classification model for poverty mapping," *J. Appl. Statist.*, vol. 48, no. 10, pp. 1882–1895, 2021.

[22] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.

[23] Y.-X. Wu, X.-Y. Min, F. Min, and M. Wang, "Cost-sensitive active learning with a label uniform distribution model," *Int. J. Approx. Reasoning*, vol. 105, pp. 49–65, Feb. 2019.

[24] G. King and L.-C. Zeng, "Logistic regression in rare events data," *Social Sci. Electron. Publishing*, vol. 9, no. 2, pp. 137–163, 2001.

[25] M. Maalouf and T. B. Trafalis, "Robust weighted kernel logistic regression in imbalanced and rare events data," *Comput. Statist. Data Anal.*, vol. 55, no. 1, pp. 168–183, 2011.

[26] Y. Xiong and R. Zuo, "GIS-based rare events logistic regression for mineral prospectivity mapping," *Comput. Geosci.*, vol. 111, pp. 18–25, Feb. 2018.

[27] S. Nusinovici, Y. C. Tham, M. Y. Chak Yan, D. S. Wei Ting, J. Li, C. Sabanayagam, T. Y. Wong, and C.-Y. Cheng, "Logistic regression was as good as machine learning for predicting major chronic diseases," *J. Clin. Epidemiol.*, vol. 122, pp. 56–69, Jun. 2020.

[28] L. Ntsalaze and S. Ikhide, "The threshold effects of household indebtedness on multidimensional poverty," *Int. J. Social Econ.*, vol. 44, no. 11, pp. 1471–1488, Nov. 2017.

[29] A. Rupasingha and S. J. Goetz, "Social and political forces as determinants of poverty: A spatial analysis," *J. Socio-Econ.*, vol. 36, no. 4, pp. 650–671, Aug. 2007.

[30] Q. Gao, S. Yang, and S. Li, "Welfare, targeting, and anti-poverty effectiveness: The case of urban China," *Quart. Rev. Econ. Finance*, vol. 56, pp. 30–42, May 2015.

[31] C. DeNavas-Walt and D. B. Proctor. (2015). *Income and Poverty in the United States: 2014*. [Online]. Available: https://www.census.gov/-content/dam/Census/library/publications/2015/demo/p60-252.pdf

[32] J. Bradshaw. (2000). *Poverty and Social Exclusion in Britain*. [Online]. Available: http://eprints.whiterose.ac.U.K./74234/1/Document.pdf

[33] D. Khullar and D. A. Chokshi, "Health, income, & poverty: Where we are & what could help," *Health Affair*, vol. 10, pp. 1–6, Oct. 2018.

[34] C. Sullivan, "Calculating a water poverty index," *World Develop.*, vol. 30, no. 7, pp. 1195–1210, Jul. 2002.

[35] J. Hagen-Zanker, F. Bastagli, L. Harman, V. Barca, G. Sturge, and T. Schmidt. (2016). *Understanding the Impact of Cash Transfers: The Evidence (Briefing)*. [Online]. Available: https://www.odi.org/-sites/odi.org.U.K./files/resource-documents/10748.pdf

[36] Y.-H. Yu, W. Luo, M.-X. He, X. Yang, B. Liu, Y. Guo, G. Thornicroft, C. L. W. Chan, and M.-S. Ran, "Household poverty in people with severe mental illness in rural China: 1994–2015," *BJPsych Open*, vol. 6, no. 5, pp. 1–8, Sep. 2020.

[37] H. M. E. Vanagt, K. Stronks, and J. P. Mackenbach, "Chronic illness and poverty in The Netherlands," *Eur. J. Public Health*, vol. 10, no. 3, pp. 197–200, Sep. 2000.

[38] Y.-L. Liu, K.-Q. Rao, and W. C. Hsiao, "Medical expenditure and rural impoverishment in China," *J. Health, Population Nutrition*, vol. 21, no. 3, pp. 216–222, 2003.

[39] E. J. O. Rao and M. Qaim, "Supermarkets, farm household income, and poverty: Insights from Kenya," *World Develop.*, vol. 39, no. 5, pp. 784–796, May 2011.

[40] S. Bell and K. Newitt, "Decent work and poverty eradication: Literature review and two-country study," Ergon Assoc., London, U.K., Tech. Rep. 1, Jan. 2010, pp. 1–70. [Online]. Available: https://www.ituc-csi.org/IMG/pdf/Decent-work-and-poverty-eradication-literature-review-and-two-country-study_Full-report.pdf

[41] J. B. G. Tilak, "Education and poverty," *J. Hum. Develop.*, vol. 3, no. 2, pp. 191–207, 2002.

[42] M. Hanratty, "Do local economic conditions affect homelessness? Impact of area housing market factors, unemployment, and poverty on community homeless rates," *Housing Policy Debate*, vol. 27, no. 4, pp. 640–655, Jul. 2017.

[43] M.-Y. Ma, Y. Li, N.-S. Wang, Q.-H. Wu, L.-H. Shan, M.-L. Jiao, X.-L. Fu, H. Li, T. Sun, B. Yi, W.-X. Tian, Q. Xia, B.-G. Shi, Y.-H. Hao, H. Yin, N. Ning, L.-J. Gao, L.-B. Liang, and J.-H. Wang, "Does the medical insurance system really achieved the effect of povert alleviation for the middle-aged and elderly people in China? Characteristics of vulnerable groups and failure links," *BMC Public Health*, vol. 20, pp. 1–15, Dec. 2020.

[44] M. Pinilla-Roncancio, "Disability and poverty: Two related conditions. A review of the literature," *Revista de la Facultad de Medicina*, vol. 63, no. 3, pp. 113–123, Oct. 2015.

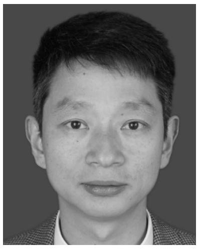
YAN-XUE WU received the B.S. and M.S. degrees from the School of Computer Science, Southwest Petroleum University, Chengdu, China, in 2016 and 2019, respectively.

His current research interests include machine learning, data mining, and causal inference. He is an Anonymous Reviewer of IEEE Access.

ZHI-NENG HU received the Ph.D. degree in management from Sichuan University, Chengdu, China, in 2005.

He is currently the Vice Director of the Institute of Information Decision Making and the Head of the Department of Management Science and Systems Science, Sichuan University. He has published seven official books in the Science Press and more than 30 journal articles. His current research interests include management science and systems science. He is a member of the Council of the Systems Engineering Society of China (SESC) and the Vice Secretary General of the Systems Engineering Society of Sichuan, China (SESSC).

YUAN-YUAN WANG received the M.S. degree in management science and engineering from Chongqing Jiaotong University, Chongqing, China, in 2018. She is currently pursuing the Ph.D. degree with Sichuan University, Sichuan, China.



FAN MIN (Member, IEEE) received the M.S. and Ph.D. degrees from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2000 and 2003, respectively.

He visited the University of Vermont, Burlington, Vermont, from 2008 to 2009. He is currently a Professor with Southwest Petroleum University, Chengdu. He has published more than 100 refereed articles in various journals and conferences, including *Information Sciences*, the *International Journal of Approximate Reasoning*, and *Knowledge-Based Systems*. His current research interests include data mining, recommender systems, active learning, and granular computing.