Predictive Modeling of Energy Poverty with Machine Learning Ensembles: Strategic Insights from Socioeconomic Determinants for Effective Policy Implementation

Article *in* International Journal of Energy Research · June 2024

DOI: 10.1155/2024/9411326

Sidique Gawusu
Nanjing University of Science and Technology
62 PUBLICATIONS **905** CITATIONS
SEE PROFILE

Seidu Abdulai Jamatutu
Nanjing University of Science and Technology
25 PUBLICATIONS **490** CITATIONS
SEE PROFILE

Abubakari Ahmed
SD Dombo University of Business and Integrated Development Studies
116 PUBLICATIONS **2,761** CITATIONS
SEE PROFILE

Wiley
International Journal of Energy Research
Volume 2024, Article ID 9411326, 18 pages
https://doi.org/10.1155/2024/9411326

Research Article

Predictive Modeling of Energy Poverty with Machine Learning Ensembles: Strategic Insights from Socioeconomic Determinants for Effective Policy Implementation

Sidique Gawusu,[1] **Seidu Abdulai Jamatutu**,[2] **and Abubakari Ahmed**[3]

[1] *Whiting School of Engineering, Johns Hopkins University, Baltimore 21211, MD, USA*
[2] *School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China*
[3] *Department of Urban Design and Infrastructure Studies, Faculty of Planning and Land Management, SD Dombo University of Business and Integrated Development Studies, Bamahu-Wa, Ghana*

Correspondence should be addressed to Sidique Gawusu; gawususidique@gmail.com, Seidu Abdulai Jamatutu; sasajamat78@gmail.com and Abubakari Ahmed; abukson1987@gmail.com

Received 22 February 2024; Revised 25 May 2024; Accepted 1 June 2024

Academic Editor: Manohar Mishra

This study aims to identify the key predictors of the multidimensional energy poverty index (MEPI) by employing advanced machine learning (ML) ensemble methods. Traditional energy poverty research often relies on conventional statistical techniques, which limits the understanding of complex socioeconomic factors. To address this gap, we propose an approach using three distinct ML ensemble models: extreme gradient boosting (XGBoost)-random forest (RF), XGBoost-multiple linear regression (MLR), and XGBoost-artificial neural network (ANN). These models are applied to a comprehensive dataset encompassing various socioeconomic indicators. The findings demonstrate that the XGBoost-RF ensemble achieves exceptional accuracy and reliability, with a root mean squared error (RMSE) of 0.041, an R-squared (R^2) of 0.975, and a Pearson correlation coefficient of 0.992. The XGBoost-MLR ensemble shows superior generalizability, maintaining a consistent R^2 of 0.845 across both the testing and training phases. The XGBoost-ANN model balances complexity with predictive capability, achieving an RMSE of 0.056, an R^2 of 0.954 in the testing phase, and an R^2 of 0.799 in training. Significantly, the study identifies "Education," "Food Consumption Score (FCS)," "Household Food Insecurity Access Scale (HFIA)," and "Dietary Diversity Score (DDS)" as critical predictors of MEPI. These results highlight the intricate relationship between energy poverty and factors related to food security and education. By integrating the insights from these ML models with policy initiatives, this study offers a promising new approach to addressing energy poverty. It highlights the importance of education, food security, and socioeconomic factors in crafting effective policy interventions.

1. Introduction

Energy poverty is a critical global issue with far-reaching consequences for health, education, and quality of life. Despite its profound impact, many people around the world still lack access to reliable and affordable energy, a fundamental component of economic growth and social welfare [1, 2]. This study aims to deepen the understanding of energy poverty by utilizing advanced ML techniques to predict the multidimensional energy poverty index (MEPI).

To grasp the full scope of energy poverty, it is essential to look beyond mere economic indicators. Intersecting factors such as educational levels, food security, and household demographics are believed to play pivotal roles in shaping access to energy [3, 4]. In response to this complexity, this research introduces an innovative use of ML ensembles, combining the strengths of extreme gradient boosting (XGBoost) with random forest (RF), multiple linear regression (MLR), and artificial neural network (ANN) models. This fusion of techniques is designed to reveal the interplay between socioeconomic factors and their influence on energy poverty [5, 6].

The scholarly pursuit of energy poverty prediction is gaining momentum, with an increasing number of studies employing various feature selection and regression methods. Notably, Qurat-ul-Ann and Mirza [7] utilized logistic regression to investigate the determinants of multidimensional energy poverty, discovering that male-headed households are at a higher risk. Yet, this risk diminishes with the increase in the age and education level of the household head, as well as with the household's geographic latitude. Abbas et al. [8] further leveraged machine learning (ML) to identify key socioeconomic factors that contribute to severe multidimensional energy poverty, such as household wealth, home size and ownership, marital status of the main breadwinner, and geographical location.

The objectives of this research are threefold: to enhance the effectiveness of policymaking in the area of energy poverty, to rigorously evaluate the robustness, accuracy, and generalizability of the ML models, and to apply advanced ML techniques to uncover deep insights into the predictors of the MEPI. The study seeks to answer the following questions:

(i) How can advanced ML models enhance policymaking by identifying key predictors of MEPI and providing precise data-driven insights for targeted interventions?

(ii) What measures ensure the robustness, accuracy, and generalizability of ML models across diverse contexts, making the insights they generate universally applicable?

(iii) How does the application of ML in energy poverty research transcend the capabilities of traditional analytical techniques, offering more dynamic tools for policymakers?

The paper is organized as follows: Section 2 offers a concise review of the literature on energy poverty through the lens of ML. Section 3 describes the methodology employed and provides an exposition of the predictive modeling techniques utilized. Section 5, titled "Results and Discussion," delves into the findings and their analysis. The study concludes with Section 6, "Conclusions and Policy Implications," where we summarize the findings and offer recommendations for policy formulation.

2. Related Literature

In contemporary society, characterized by rapid technological advancements, the persistence of energy poverty remains a salient challenge. Despite the global focus on development and innovation, a significant segment of the world's population still grapples with this issue, a reality that transcends the conventional divide between developed and developing countries [9, 10]. This phenomenon is especially pronounced in low-income and rural communities. Energy poverty can be measured through two main lenses: objective and subjective indicators. The former encompasses quantifiable metrics like household income and energy expenditure, used by external evaluators to assess the extent of energy poverty [11]. In contrast, subjective indicators revolve around individual perceptions or satisfaction levels concerning energy costs, reflecting the personal assessment of the energy expenditure required to maintain a basic standard of living [11].

Over recent years, there has been a surge in innovative techniques aimed at understanding and addressing energy poverty. Numerous studies have delved deep into its prevalence and the underlying causes across various global contexts [12, 13]. A notable development in this research trajectory is the incorporation of AI in the analytical process [14]. AI's integration has not only facilitated a more nuanced understanding but has also provided tools to potentially alleviate energy scarcity [15, 16].

For instance, a study by Legendre and Ricci [17] employed a logistic regression model to explore the demographic and socioeconomic determinants of energy poverty. Their findings emphasized that certain groups, such as individuals with higher education, those living with a partner, and homeowners, are less likely to experience energy poverty. On the other hand, retired seniors living solo, inhabitants of older buildings, or those dwelling in expansive residences are more vulnerable to energy poverty.

In a similar vein, van Hove et al. [15] harnessed the power of ML to decipher the multifaceted contributors to energy poverty in Europe. By training a gradient-boosting classifier on a range of socioeconomic characteristics, they discerned both overarching and region-specific predictors of energy poverty, providing a granular understanding of this complex issue. Rajić et al. [18] further expanded the research horizon by introducing a sophisticated AI-driven model grounded in socioeconomic determinants to study energy poverty. Their innovative approach focused on predicting electricity and heat consumption based on 15 critical factors. A standout feature of their methodology was the development of a unique data optimization framework, which streamlined the process of selecting the most relevant variables from raw datasets. Similarly, Romero et al. [19] and Abbas et al. [20] undertook extensive logistic regression and Tobit model analyses, respectively, to unearth the various factors exacerbating or mitigating energy poverty. Their comprehensive studies shed light on the socioeconomic dynamics, such as household composition, education levels, and housing status, that play a pivotal role in determining energy poverty vulnerability. Another study by Hurst et al. [21] capitalized on the data from gas smart meters to predict household poverty states. Utilizing decision trees and cloud analytics, they achieved remarkable accuracy in determining socioeconomic conditions, further exemplifying the potential of ML in addressing energy poverty.

The vast body of literature on energy poverty offers a tapestry of insights, strategies, and findings. These collectively underscore the significance and relevance of ongoing research in this domain. However, gaps in knowledge and understanding persist, signaling uncharted territories that warrant exploration. This study will not only complement but also enrich the current knowledge base. The data collection techniques and analytical methods are highlighted, ensuring a holistic approach while integrating diverse research paradigms.

3. Methodology and Data

Data preprocessing is a critical and initial statistical step in developing a precise ML model. The current model incorporates various kinds of variables, such as categorical, binary,

Table 1: Variables of the predictive model and descriptions.

#	Mean	Standard deviation	Skewness	Kurtosis	Coefficient of variation	Mean absolute deviation	Minimum	Median	Maximum
Input									
Age	36.41	10.85	1.33	1.91	0.30	8.22	20.00	34.00	86.00
Sex	0.54	0.50	−0.18	−1.97	0.92	0.50	0.00	1.00	1.00
Household size	5.01	2.57	1.17	3.36	0.51	1.88	1.00	5.00	21.00
Monthly income	1,597.83	1,372.59	2.62	13.30	0.86	930.58	0.00	1,400.00	13,000.00
FCS	84.77	15.80	−0.86	0.26	0.19	12.63	22.00	89.00	112.00
HFIA	1.38	0.74	1.94	2.90	0.54	0.57	1.00	1.00	4.00
DDS	12.07	2.00	−1.19	1.10	0.17	1.56	4.00	13.00	14.00
Married	0.54	0.50	−0.14	−1.98	0.93	0.50	0.00	1.00	1.00
Employed	0.87	0.34	−2.22	2.92	0.39	0.22	0.00	1.00	1.00
Education	0.83	0.38	−1.77	1.13	0.45	0.28	0.00	1.00	1.00
Output									
MPEI	0.29	0.26	0.06	−1.35	0.88	0.24	0.00	0.40	0.99

and continuous. As part of data cleaning, it is crucial to preprocess and transform data to stop large numeric values from overpowering smaller ones [22, 23]. This eliminates biases, noise, anomalies, outliers, or skewness, which could otherwise lead to inaccuracies [24]. Table 1 presents a detailed description and statistical analysis of the variables employed in this study.

3.1. Study Area and Data Source. The study was conducted in Wa, a secondary urban center located in the northwestern part of Ghana's semi-arid region. Geographically, Wa is situated between latitudes 1°40′N and 2°45′N and longitudes 9°32′W and 10°20′W, covering an area of 579.86 km^2 [25]. Historical demographic data indicate a significant increase in Wa's population, growing from 8,000 residents in 1,880 [26] to 71,340 in 2010, as reported by the Ghana Statistical Service [27], with an estimated rise to 125,479 by 2017. From 1986 to 2016, there was substantial urban expansion in Wa, with built-up areas increasing from 3.7 to 29.2 km^2 [25].

Wa faces various socioeconomic challenges, particularly in employment and energy utilization. Within this urban setting, 91.5% of the economically active population, totaling 34,984 out of 38,239 people, are employed [27]. Regarding energy access, over 81% of the residents have electricity for lighting, which suggests considerable urban infrastructure development [27]. Nonetheless, the remaining population relies on alternative lighting sources, including flashlights, private electric generators, and solar or kerosene lamps.

Regarding cooking energy sources, 65% of Wa's residents primarily utilize charcoal for cooking needs [27]. With the introduction and growing acceptance of LPG, its use has increased to over 18% for cooking [27]. Still, many inhabitants continue to rely on traditional fuels such as firewood, kerosene, and crop residues. In terms of spatial arrangements for cooking, 50% of the population uses verandas, while 22% cook in open spaces within their residential compounds [27]. Some residents, lacking designated cooking areas, are forced to cook in spaces within their living quarters that lack modern kitchen facilities [27, 28]. These practices point to significant concerns about indoor pollution, health implications, and gender-related issues, reflecting the complex interaction between socioeconomic factors and living conditions [29, 30].

This research utilizes original data obtained from 776 households residing in various neighborhoods within the Wa Municipality of the Upper West Region (Figure 1). Figure 1 is reproduced from Gawusu et al. [14]. These neighborhoods were grouped into different zones, core, and fringe zonal areas. The dataset was compiled through a variety of rigorous data collection strategies, ensuring a robust and representative sample of the population in Wa. These strategies included structured surveys, direct interviews, and observational methods, each tailored to effectively gather data on energy usage, socioeconomic status, and other pertinent variables. The survey instruments were developed based on established guidelines in the literature and were pre-tested to ensure reliability and validity in the context of Wa's unique socioeconomic and cultural landscape.

Data collection was conducted over a period of 6 months, involving a team of trained local researchers who were familiar with the regional dialects and cultural nuances, which significantly enhanced the accuracy and depth of the data obtained. This primary data was then verified and processed to ensure consistency and accuracy before analysis. As a result, the dataset offers a comprehensive view of the demographic, social, economic, and health characteristics of the households, including information about their possessions, and spatial variables. The data encompasses all the relevant variables required to assess various indicators of multidimensional energy poverty at the household level. Each variable within the dataset was checked for accuracy and completeness, ensuring the integrity of the findings.

3.1.1. Derivation of MEPI. In this study, the MEPI was computed by considering seven indicators that measure energy deprivation [32]. These indicators encompass the following aspects: access to lighting, access to modern cooking fuel, access to a clean indoor environment, and access to household

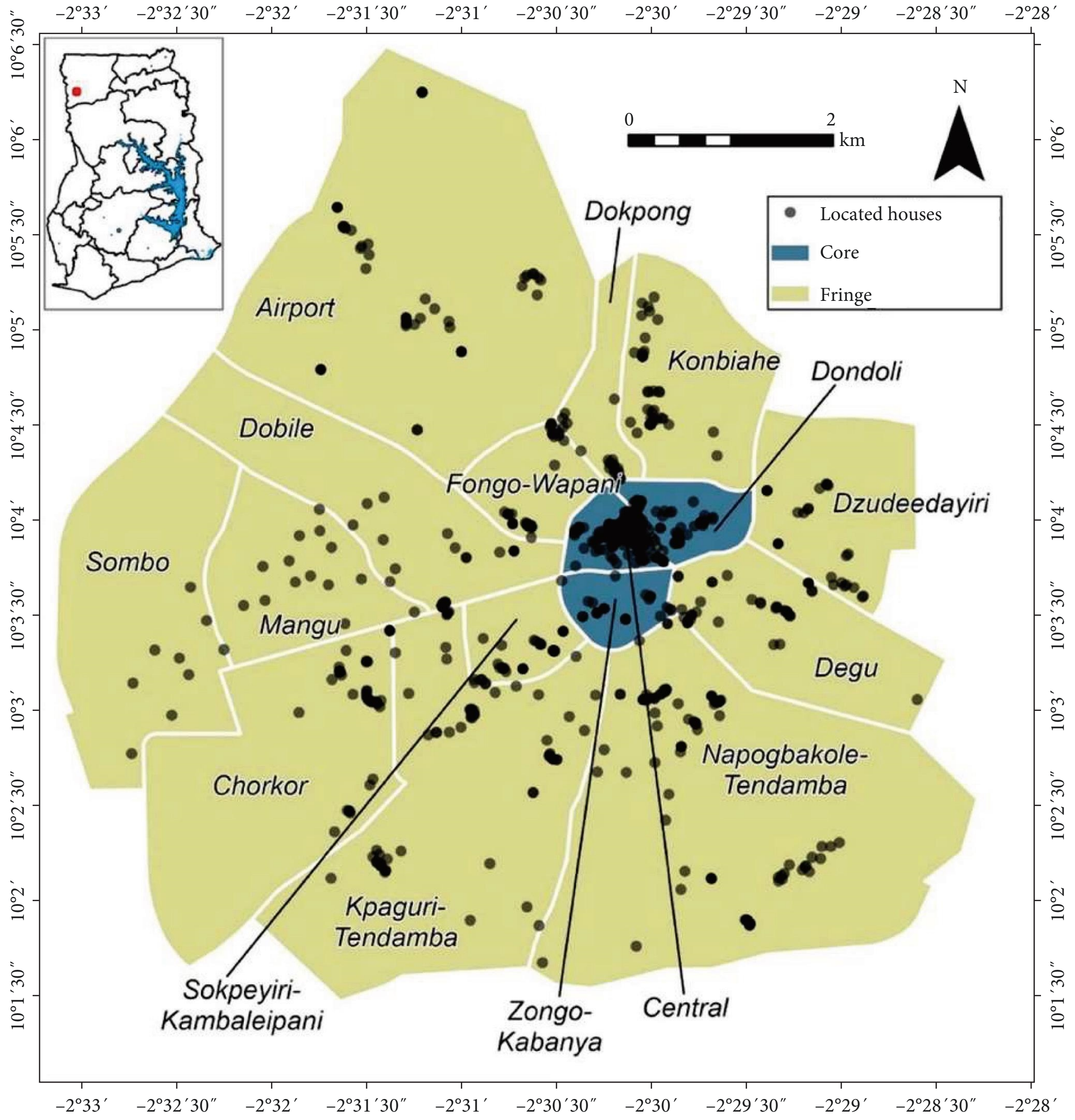


Figure 1: Context map of Wa with studied core and fringe households [31].

amenities such as refrigeration, recreation, communication, and space cooling. These indicators are considered to capture the multidimensional nature of energy. MEPI captures the various energy deprivations that can impact individuals or households. It consists of five dimensions that represent essential energy services, with seven indicators used to measure these deprivations. A household is considered energy-poor if the combination of deprivations exceeds a predefined threshold. In this study, the threshold set requires a household to be deprived in at least 30% of the weighted indicators to be considered multidimensionally poor.

3.2. Feature Selection. The feature selection narrows down the count of predictors when building a predictive model, primarily by removing input variables that show a weak correlation strength with the output variable [33]. Having irrelevant input variables in the predictive model introduces inefficiencies and slows down the training, validation, and testing procedures by taking up substantial system memory [34].

The choice of features was guided by their relevance to the MEPI, the target variable. These features include critical demographic indicators, socioeconomic factors, and nutritional metrics. Each feature was selected based on its established connection to poverty outcomes, as documented in various studies on poverty assessment. We analyzed statistical properties such as mean, standard deviation, skewness, correlation coefficients, and kurtosis for each feature to ensure a robust analysis framework, enhancing the understanding of each variable's impact on poverty.

Consequently, it is beneficial to decrease features to boost the overall performance and efficiency of the model and, as a result, reduce its computational cost [35, 36]. The choice between a supervised or unsupervised feature selection method depends on the nature of the data and target variables. A supervised method is utilized in this study because of the nature of the target variable; alternatively, in scenarios without a response variable, an unsupervised method is adopted [37].

The Pearson correlation coefficient (PCC) is widely employed if the output variable is continuous, while the mutual information ranking method is chosen if the response variable is categorical or binary [38]. Given the diverse nature of input and output variables (continuous, binary, and categorical) from survey data, this study utilizes both methods to select the most relevant features to build the final predictive mode. The PCC is mostly used in statistical analysis to eliminate predictors that are not correlated when developing a model [39, 40].

$$R_i = \frac{\operatorname{cov}(X_i, Y)}{\sqrt{\operatorname{var}(X_i, \operatorname{var}(Y))}}, \tag{1}$$

$$R_i = \frac{\sum_i (x_i - \overline{x}_i)}{\sqrt{\sum_i (x_i - \overline{x}_i)^2 (y_i - \overline{y}_i)^2}}. \tag{2}$$

Equation (1) utilizes the terms cov to represent covariance and var to indicate variance. On the other hand, Equation (2) employs $\overline{x}_i$ and $\overline{y}_i$ to represent the means of variables X and Y, respectively, while R_i denotes the correlation value that measures the strength of the linear relationship between X and Y, ranging from −1 to 1. This equation establishes a ranking criterion that assesses the quality of the linear association between the variables [39].

The feature selection process in the predictive modeling aimed to identify relevant predictors for MEPI estimation. Given the complex nature of the dataset, it was initially hypothesized that other parameters might serve as predictors for MEPI. To evaluate the potential of these variables as features, we conducted a preliminary correlation analysis, as shown in Figure 2.

In selecting predictors for the ensemble predictive modeling, we considered a range of variables, including age, sex, household size, monthly income, FCS, HFIA, DDS, marital status, employment status, and education level. While some of these predictors exhibit weak or negative correlations with the target variable, MEPI, their inclusion is justified by a comprehensive understanding of the multifaceted nature of energy poverty. First, the presence of weak correlations does not negate the potential predictive power of these variables when combined with others in an ensemble model, where the interaction between predictors can reveal complex, nonlinear relationships not evident in simple correlation analyses [41]. Therefore, weak correlations do not inherently disqualify variables from inclusion in the predictive models [42]. The complexity of MEPI often defies simple linear relationships. Consequently, ensemble models like XGBoost can capture complex interactions between features that may not be evident in a correlation matrix. ML models can unearth patterns within the data that go beyond what is revealed by correlation alone, particularly when interactions between variables are considered.

Negative correlations, particularly those observed with monthly income, education, FCS, and DDS, are theoretically informative, suggesting that higher socioeconomic status and better food security are inversely related to energy poverty, an insight valuable for policy formulation and intervention design. Moreover, the ensemble modeling approach, by aggregating predictions from multiple models, can leverage the unique contributions of each predictor, regardless of its correlation with MEPI, thereby enhancing model robustness and predictive accuracy.

The inclusion of a diverse set of predictors also allows for a holistic analysis of energy poverty, acknowledging it as a multidimensional issue influenced by a variety of socioeconomic, demographic, and household factors, thereby aligning the methodology with the complex reality of the phenomena under study.

Moreover, given the weak correlations observed, we incorporated dimensionality reduction techniques to refine the feature set. Methods such as principal component analysis and feature importance scores from preliminary model runs were employed to reduce model complexity and prevent overfitting [43]. These techniques can help in retaining only the most informative features, potentially including those with weak individual correlations but strong collective influence on the target variable.

3.3. Predictive Modeling Techniques. In this study, we have carefully selected a suite of ML ensembles to investigate the MEPI. The choice to use specific ML models-XGBoost-RF, XGBoost-MLR, and XGBoost-ANN; was driven by the aim to achieve high accuracy in predicting MEPI and to effectively capture the complex, nonlinear relationships that are characteristic of data on energy poverty.

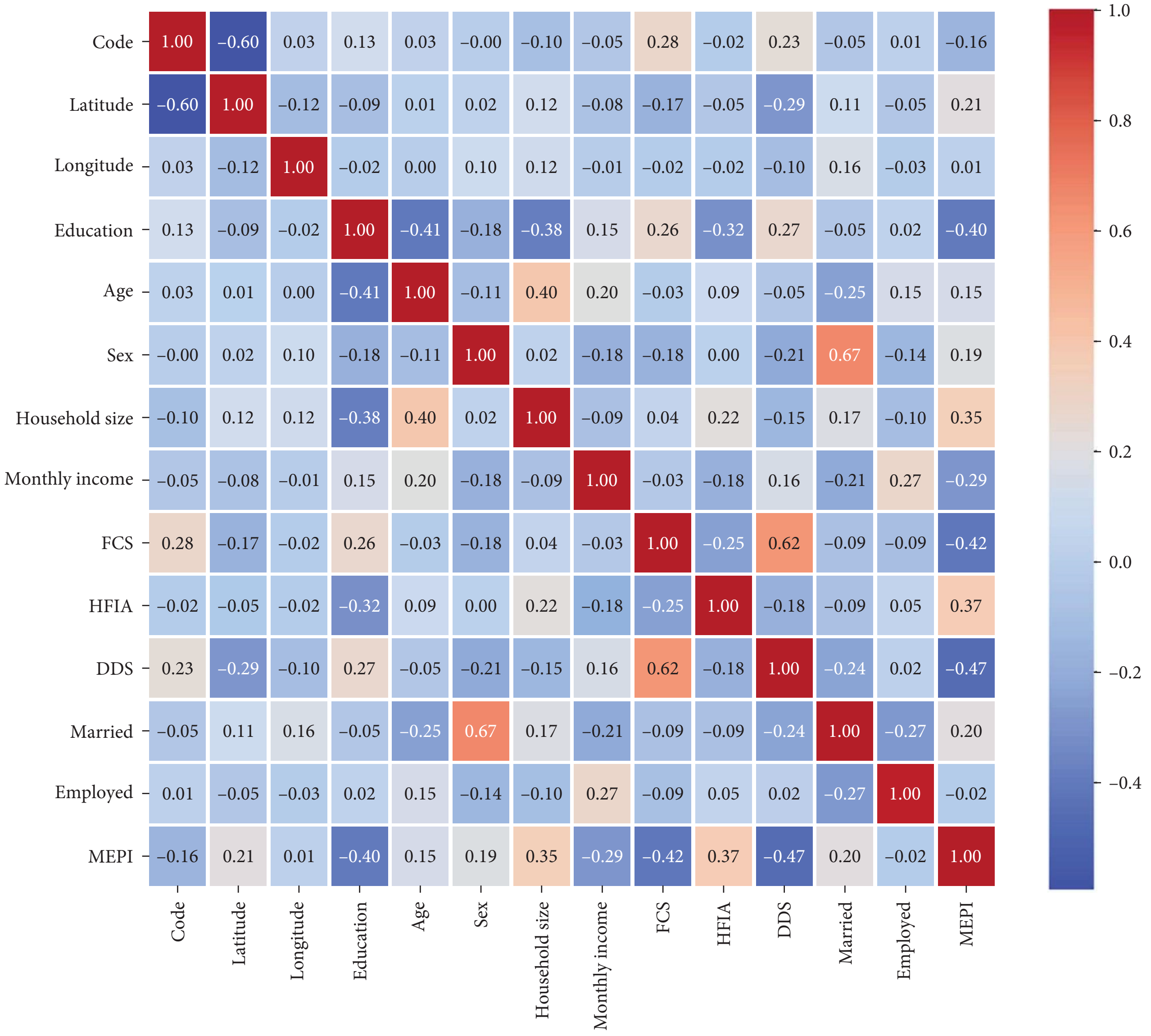
The rationale behind selecting these particular ensemble models was based on the unique properties of the dataset and the intricate nature of energy poverty. Each model combines the strengths of XGBoost with other advanced ML techniques, offering unique benefits tailored to different aspects of the analysis:

(i) XGBoost-MLR is chosen for its ability to provide insights into the generalizability of the model across various contexts, which is essential for applying the findings to broader scenarios.

(ii) XGBoost-RF is utilized for its robustness and reliability, making it ideal for handling diverse data types and complex data structures within the dataset.

(iii) XGBoost-ANN leverages the capabilities of ANNs to delve deeper into the dataset, identifying subtle patterns and interactions that simpler models might overlook.

This approach not only enhances the accuracy of the predictions but also provides comprehensive insights into the effectiveness of various interventions to combat energy poverty. By integrating these diverse ML techniques, the



	Code	Latitude	Longitude	Education	Age	Sex	Household size	Monthly income	FCS	HFIA	DDS	Married	Employed	MEPI
Code	1.00	−0.60	0.03	0.13	0.03	−0.00	−0.10	−0.05	0.28	−0.02	0.23	−0.05	0.01	−0.16
Latitude	−0.60	1.00	−0.12	−0.09	0.01	0.02	0.12	−0.08	−0.17	−0.05	−0.29	0.11	−0.05	0.21
Longitude	0.03	−0.12	1.00	−0.02	0.00	0.10	0.12	−0.01	−0.02	−0.02	−0.10	0.16	−0.03	0.01
Education	0.13	−0.09	−0.02	1.00	−0.41	−0.18	−0.38	0.15	0.26	−0.32	0.27	−0.05	0.02	−0.40
Age	0.03	0.01	0.00	−0.41	1.00	−0.11	0.40	0.20	−0.03	0.09	−0.05	−0.25	0.15	0.15
Sex	−0.00	0.02	0.10	−0.18	−0.11	1.00	0.02	−0.18	−0.18	0.00	−0.21	0.67	−0.14	0.19
Household size	−0.10	0.12	0.12	−0.38	0.40	0.02	1.00	−0.09	0.04	0.22	−0.15	0.17	−0.10	0.35
Monthly income	−0.05	−0.08	−0.01	0.15	0.20	−0.18	−0.09	1.00	−0.03	−0.18	0.16	−0.21	0.27	−0.29
FCS	0.28	−0.17	−0.02	0.26	−0.03	−0.18	0.04	−0.03	1.00	−0.25	0.62	−0.09	−0.09	−0.42
HFIA	−0.02	−0.05	−0.02	−0.32	0.09	0.00	0.22	−0.18	−0.25	1.00	−0.18	−0.09	0.05	0.37
DDS	0.23	−0.29	−0.10	0.27	−0.05	−0.21	−0.15	0.16	0.62	−0.18	1.00	−0.24	0.02	−0.47
Married	−0.05	0.11	0.16	−0.05	−0.25	0.67	0.17	−0.21	−0.09	−0.09	−0.24	1.00	−0.27	0.20
Employed	0.01	−0.05	−0.03	0.02	0.15	−0.14	−0.10	0.27	−0.09	0.05	0.02	−0.27	1.00	−0.02
MEPI	−0.16	0.21	0.01	−0.40	0.15	0.19	0.35	−0.29	−0.42	0.37	−0.47	0.20	−0.02	1.00

Figure 2: Correlation analysis for the predictors and MEPI.

study gains a holistic understanding of the dynamics at play, which aids in formulating more effective and precisely targeted policy measures.

3.3.1. MLR. MLR is a statistical method that employs numerous independent variables to forecast the value of a dependent variable [44]. The primary objective of MLR is to establish the linear relationship between the independent variables and the dependent variable [45]. The MLR model is defined as follows [46]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon, \qquad (3)$$

where Y is the dependent variable, $X_1, X_2, \ldots., X_n$ are the independent variables, $\beta_0, \beta_1, \beta_2, \ldots.., \beta_n$ are the parameters of the model, and ε is the error term.

MLR operates on a few crucial assumptions, such as linearity, independence, homoscedasticity, and normality. It is necessary to verify these assumptions to ensure the validity and reliability of the analysis outcomes.

3.3.2. RF. RF [47] is a widely used and adaptable machine ML technique [48, 49] capable of carrying out both regression and classification tasks. It exemplifies an ensemble learning method, where a collective of weak models amalgamate to construct a robust model. In the RF, multiple decision trees are generated, and each tree provides a "vote" for a specific class. The forest opts for the class receiving the most votes, and in the case of regression, it computes the average of the outputs from different trees [50].

3.3.3. XGBoost. XGBoost enhances the gradient boosting framework by integrating regularization techniques to curb overfitting, thus ensuring robust model performance [51]. It operates on the principle of sequentially building decision trees, where each tree corrects errors made by the previous ones, converging toward a powerful composite model. The core of XGBoost's success lies in its ability to optimize complex loss functions, incorporating both the predictions' accuracy and the model's simplicity through regularization.

Central to XGBoost's approach is its objective function, which is an amalgamation of a differentiable loss function and a regularization term. The objective function is defined as follows:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{4}$$

where l denotes the loss function assessing the discrepancy between the predicted $(\widehat{y}_i)$ and actual (y_i) values over n instances, K represents the number of trees, and Ω signifies the regularization term. The regularization term $\Omega(f)$ is pivotal in controlling the model's complexity, formulated as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2, \tag{5}$$

where T is the total number of leaves in the tree, w embodies the leaf weights, γ adjusts the complexity based on the number of leaves, and λ is the L2 regularization term on the leaf weights, mitigating overfitting by penalizing large weights.

The optimization of the objective function leverages both the gradient (g_i) and the Hessian (h_i), the first and second-order partial derivatives of the loss function with respect to the prediction. These derivatives guide the update steps in the model, ensuring that each addition effectively reduces the loss function:

$$g_i = \partial_{\widehat{y}^{(t-1)}} l\left(y_i, \widehat{y}^{(t-1)}\right), \tag{6}$$

$$h_i = \partial^2_{\widehat{y}^{(t-1)}} l\left(y_i, \widehat{y}^{(t-1)}\right). \tag{7}$$

This use of second-order information allows for more nuanced model adjustments, particularly in the presence of complex loss landscapes. XGBoost employs a novel tree construction algorithm that grows trees to their maximum depth and then prunes them using a gain calculation derived from the loss function. This method contrasts with traditional greedy algorithms, enabling the identification and removal of splits that contribute minimally to the model's predictive power.

3.3.4. Underlying Principle of the ANN Model. The MLP is a class of feedforward ANN, notable for its multiple layers of interconnected neurons, consisting of input, hidden, and output layers (Figure 3). These layers are intricately linked via neurons, each holding information in the form of weights. The input layer initiates the information for computation, the hidden layer(s) carry out computational duties, and the output layer conducts predictions and categorization [52], as highlighted in Equation (8).

$$y_{li} = f_{li}(z_{li}); z_{li} = \sum_{j=1}^{n_{l-1}} w(l-1)_j, liY(l-1)_j + b_{li}, \tag{8}$$

where lth are layers, ith are neurons, y_{li} denotes output, f_{li} is the activation function, w is the weight, and b_{li} are the biases. The activation function bridges the nodes of a layer, transmitting information to the nodes in the subsequent layer. The neural computations occurring in the hidden and output layers can be elucidated via Equations (9) and (10), respectively. In these equations, $b^{(1)}$ and $b^{(2)}$ represent the biases, $W^{(1)}$ and $W^{(2)}$ denote the weight vectors, and G and $s$are the activation functions. $\Phi = \{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\}$ is set for the parameters to learn [52].

$$o(x) = G\left(b^{(2)} + W^{(2)}h(x)\right), \tag{9}$$

$$h(x) = \Phi(x) = s\left(b^{(1)} + W^{(1)}x\right). \tag{10}$$

The training phase involves a set of predictors and output patterns used to train the model. The outputs generated by the neural network in response to input variables are the dependent variables. As each pattern is read, the network produces an output from the input data, which is then compared with the correct output. In the event of any discrepancies or errors, the connection weights automatically adjust their course to minimize these errors. Python programing language, augmented by TensorFlow, was used to construct and train the models, underscoring the significance of state-of-the-art software tools in the predictive modeling process. Figure 4 highlights the complex modeling process employed in this study.

3.4. Regression Model Accuracy. The evaluation of predictive outcomes across various scenarios presented in this document was conducted employing a suite of statistical measures, including mean absolute percentage error, root mean squared error (RMSE), R-squared (R^2), PCC, Nash–Sutcliffe coefficient (NSE), and Willmott's index (WI). Several researchers, including [53, 54], have advocated for these performance indicators as essential tools for addressing diverse challenges:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}, \tag{11}$$

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{12}$$

$$\text{NSE} = 1 - \left[\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}\right], \tag{13}$$

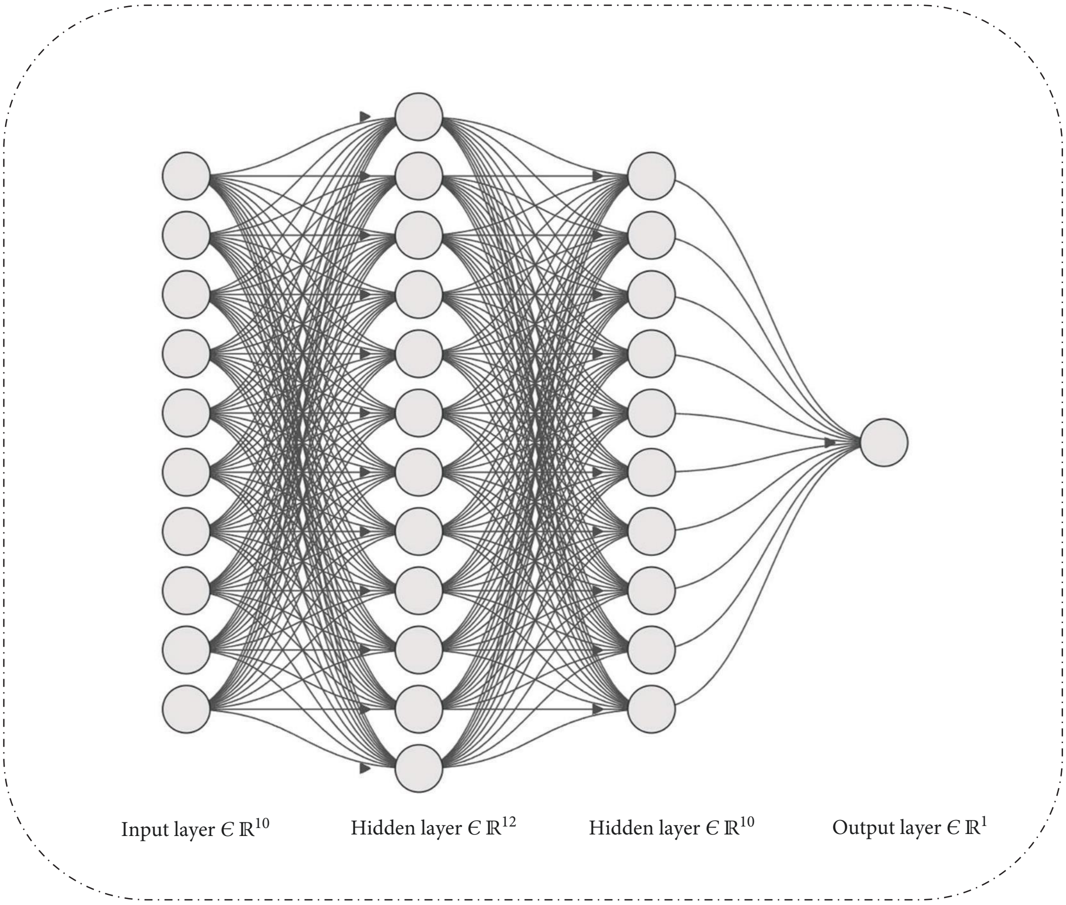


Figure 3: Artificial neural network architecture.

$$\mathrm{WI} = 1 - \left[\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(|y_i - \overline{y}| + |y_i - \hat{y}_i|)^2}\right], \quad (14)$$

$$\mathrm{PCC} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \quad (15)$$

where y_i = actual value, $\hat{y}_i$ = predicted value, n = total number of data points, x_i and y_i are the paired data values, $\overline{x}$ and $\overline{y}$ are the means of actual values.

3.5. Model Development and Configuration

3.5.1. Data Preprocessing. The dataset was divided into a training set, constituting 70% of the data, for model training and a testing set, making up the remaining 30%, to evaluate the models' predictive performance. To ensure an equitable influence of all features during the learning process, data scaling was performed to normalize the variable values, assigning them within the range of 0–1 [22, 55]. This normalization aligns with best practices for preprocessing data when utilizing ML algorithms [56, 57].

3.5.2. Hyperparameter Tuning. To ensure the robustness and reliability of the model evaluations, we employed k-fold cross-validation across various training data subsets. This method was particularly important for validating the performance of the ensemble configurations, as it helped mitigate the risk of overfitting by exposing each model to multiple training scenarios [58, 59].

For navigating this complex hyperparameter space, we utilized the GridSearchCV method, which systematically explored various combinations of hyperparameters to identify the most effective settings for each model within the ensemble. The search was not only focused on individual model performance but also on how well the models integrated and complemented each other in the ensemble framework.

The specific hyperparameters tuned for each model are as follows:

(i) XGBoost: We adjusted parameters such as n_estimators, learning_rate, max_depth, subsample, and colsample_bytree. These parameters are crucial as they control aspects such as the number of trees built, the depth of each tree, and the rate at which the model learns. For the learning_rate, we explored a range from 0.01 to 0.1, adjusting incrementally based on model performance during cross-validation. This range was chosen to balance the tradeoff

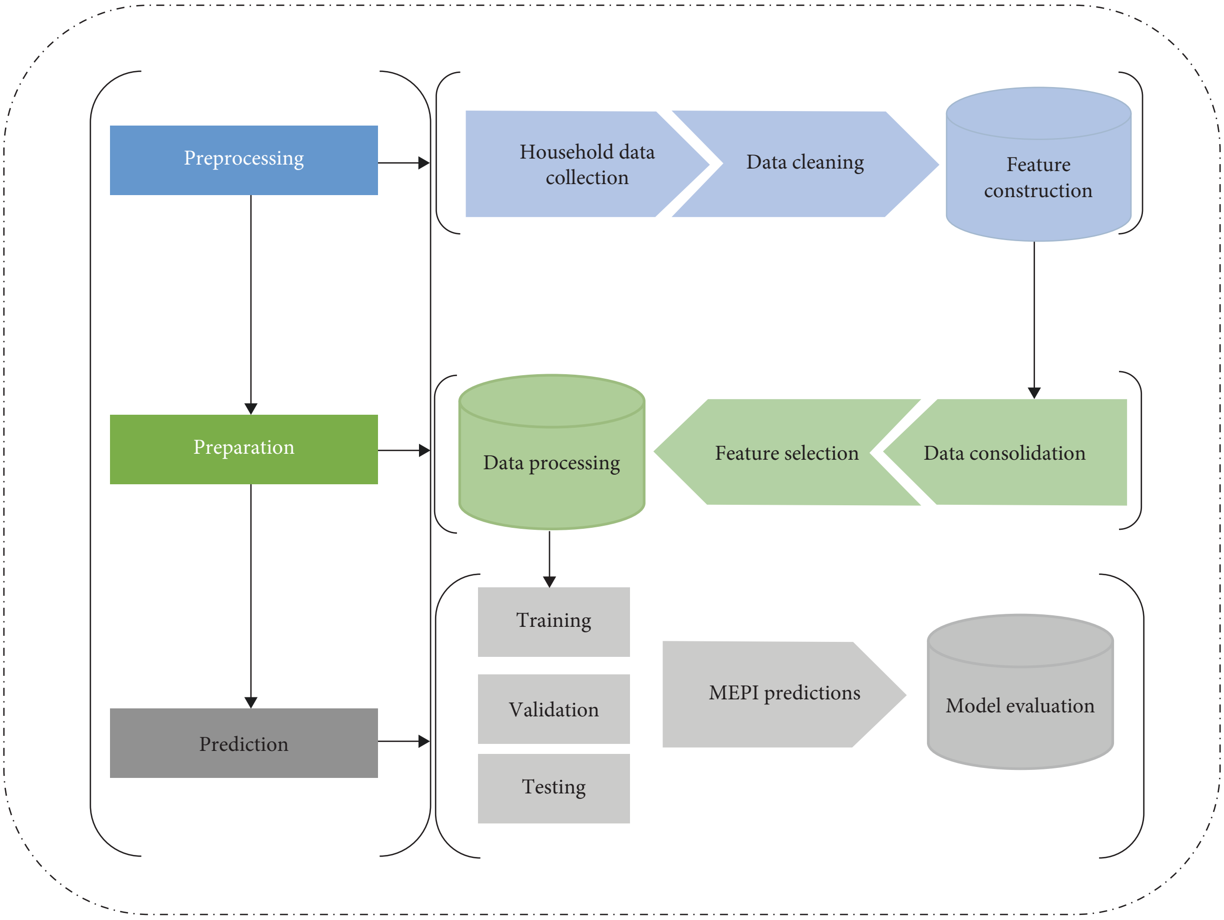


FIGURE 4: The ML prediction process.

between convergence speed and the risk of overshooting the optimal parameter space.

(ii) RF: We focused on n_estimators, max_features, max_depth, min_samples_split, and min_samples_leaf. These parameters influence the number of trees in the forest, the complexity of those trees, and how they handle the splitting of nodes and leaves within each tree. The n_estimators were set from 100 to 500, and max_depth ranged from 5 to 30, among others.

(iii) ANN: Parameters like learning_rate, as well as the number of layers and the number of neurons per layer, were tuned. These affect the network's learning speed and its overall architecture, which are pivotal for the model's capacity to process and predict from complex datasets. The learning_rate varied from 0.001 to 0.01, and the layers and neurons were adjusted within specified ranges to find the most effective structure.

(iv) MLR: The tuning focuses more on selecting relevant features and employing regularization techniques rather than traditional hyperparameter tuning. This approach helps in managing overfitting and enhancing the predictive accuracy of the model.

After the hyperparameter optimization, the ensemble models were rigorously evaluated on a testing set using a comprehensive suite of performance metrics (Equations (11), (12), (13), (14), and (15)). These metrics were selected to assess various aspects of model performance, including accuracy, correlation, and prediction precision, thus providing a holistic view of the effectiveness of the tuning strategy. This refined tuning process and the careful consideration of intermodel interactions have significantly enhanced the predictive power of the ensemble models, ensuring that the integration of outputs from different models is optimized to achieve superior predictive performance.

4. Results and Discussion

4.1. Descriptive Analysis. From Table 1, the mean MEPI shows an average value of 0.29 with a standard deviation of 0.26, indicating a moderate level of variation in energy poverty among the households studied. The coefficient of variation of 0.88 highlights a relatively high degree of variability relative to the mean, suggesting that energy poverty levels vary significantly across different households. The skewness and kurtosis values close to zero and negative, respectively, imply a distribution that is slightly skewed to the right and has a flatter peak compared to a normal distribution, indicating that while most households experience lower levels of energy poverty, there is a tail of households

Table 2: ANOVA results for selected predictors of the multidimensional energy poverty index.

Source	Sum of squares	df (degree of freedom)	F-statistic	p-Value
Education	0.37	1	10.15	0.0015
Household size	1.40	1	38.17	1.05e-09
Monthly income	1.71	1	46.59	1.79e−11
Employed	0.12	1	3.37	0.067
Age	0.07	1	1.78	0.182
Sex	0.0001	1	0.0036	0.952
HFIA	1.00	1	27.30	2.25e−07
DDS	0.77	1	20.87	5.72e-06
Married	0.19	1	5.11	0.024
FCS	1.61	1	43.82	6.77e−11
Residual	36.54	768	—	—

experiencing higher levels. The mean absolute deviation of 0.24 further supports the presence of variability in energy poverty experiences. The range of MEPI values, from 0.00 to 0.99, underscores the broad spectrum of energy poverty, from none to severe, within the population. This statistical analysis of MEPI sheds light on the complex nature of energy poverty, emphasizing the need for targeted interventions to address the varying levels of energy poverty across different households.

From Table 2, the nutritional indicators, namely the FCS, HFIA, and DDS, demonstrate a profound impact on MEPI. The FCS, with an F-statistic of 43.82 and a p-value below 0.00001, indicates that the quality of food consumption is a critical element in the context of energy poverty. Similarly, HFIA, which measures access to food, is also significantly associated with MEPI, evidenced by an F-statistic of 27.30 and a p-value of 2.25e-07. The DDS further corroborates the significance of nutritional diversity, showing a strong relationship with MEPI (F-statistic of 20.87, p-value of 5.72e-06).

Among the socioeconomic predictors, monthly income and household size remain highly significant, with F-statistics of 46.59 and 38.17, respectively, both indicating significant effects on MEPI. This highlights the direct influence of economic capacity and household dynamics on energy management within homes. Education also shows a significant effect (F-statistic of 10.15, p-value of 0.0015), suggesting that educational attainment can play a crucial role in enabling better energy decisions and access.

Marital status shows a moderate but statistically significant impact on MEPI (F-statistic of 5.11, p-value of 0.024), pointing to the potential social support systems associated with marital relationships that could affect energy poverty dynamics as reported by [60, 61]. In contrast, variables such as employment, age, and sex show diminished significance in this model, with sex showing almost no effect (p-value of 0.952), indicating that these factors may be less pivotal in the context of energy poverty compared to nutritional and economic indicators.

4.2. ML Models Predictions of MEPI. The performance metrics presented in Table 3 provide a comprehensive evaluation of the ML ensemble models, XGBoost-RF, XGBoost-MLR, and XGBoost-ANN, estimating the MEPI during the testing phase. These metrics include RMSE, R^2, PCC, NSE, and WI, which collectively offer insights into each model's accuracy, predictive power, and overall performance.

Table 3: Performance metrics of estimate MEPI using three ML ensemble models in the testing phase.

ML model	RMSE	R^2	PCC	NSE	WI
XGBoost-RF	0.041	0.975	0.992	0.975	0.993
XGBoost-MLR	0.101	0.845	0.939	0.845	0.948
XGBoost-ANN	0.056	0.954	0.979	0.845	0.947

The XGBoost-RF ensemble model exhibits superior performance across all evaluated metrics, with an RMSE of 0.041, an R^2 of 0.975, a PCC of 0.992, an NSE of 0.975, and a WI of 0.993. These results indicate a high level of accuracy and correlation between the observed and predicted MEPI values, showcasing the model's effectiveness in capturing the complex relationships inherent in the predictors of energy poverty. The strong performance of the XGBoost-RF model can be attributed to the synergistic combination of XGBoost's gradient boosting and RF's ensemble learning, which effectively reduces overfitting and enhances generalization [47, 62].

On the other hand, the XGBoost-MLR model shows comparatively lower performance, with an RMSE of 0.101, an R^2 of 0.845, a PCC of 0.939, an NSE of 0.845, and a WI of 0.948. While still demonstrating a good level of predictive capability, the XGBoost-MLR model's lower metrics reflect the challenges of linear regression models in capturing nonlinear relationships within complex datasets [63].

The XGBoost-ANN model achieves intermediate performance between the XGBoost-RF and XGBoost-MLR models, with an RMSE of 0.056, an R^2 of 0.954, a PCC of 0.979, an NSE of 0.845, and a WI of 0.947. The relatively high performance of the XGBoost-ANN ensemble underscores the ANN's capability to model complex, nonlinear relationships through its network of interconnected neurons and layers [64]. However, the slight discrepancy in NSE compared to its R^2 and PCC values could indicate potential overfitting or sensitivity to the distribution of the training data, which is a known challenge in deep learning models [65, 66].

Table 4: Performance metrics of estimate MEPI using three ML ensemble models in the training phase.

ML model	RMSE	R^2	PCC	NSE	WI
XGBoost-RF	0.126	0.796	0.684	0.796	0.812
XGBoost-MLR	0.113	0.808	0.715	0.808	0.824
XGBoost-ANN	0.121	0.799	0.667	0.799	0.715

Table 4 presents a detailed analysis of the performance of the three ML ensemble models in estimating MEPI during the training phase. The Taylor diagrams illustrated in Figures S1, S2, and S3, as detailed in the Supplementary Materials, also act as a graphical method for evaluating the effectiveness of different XGBoost ensemble models throughout their training and testing stages. By consolidating three key metrics-standard deviations, correlation, and centered root mean square difference-into a unified chart, Taylor diagrams offer a thorough perspective on the accuracy of the models' forecasts in relation to the observed MEPI values.

The XGBoost-RF model yields an RMSE of 0.126, an R^2 value of 0.796, a PCC of 0.684, an NSE of 0.796, and a WI of 0.812. These metrics indicate a reasonably good fit of the model to the testing data, suggesting effective learning and prediction capabilities. However, the lower PCC compared to the training phase metrics might indicate a reduced correlation between predicted and actual MEPI values in the testing phase, hinting at potential overfitting issues or limitations in capturing the variance in unseen data [67, 68].

The XGBoost-MLR model demonstrates slightly better generalization performance than the XGBoost-RF model, with an RMSE of 0.113, an R^2 of 0.808, a PCC of 0.715, an NSE of 0.808, and a WI of 0.824. The improvement in R^2 and NSE metrics over the RF model suggests that the linear combination in the MLR model may have captured the underlying relationships in the MEPI estimation more consistently across the training and testing phases. This could be attributed to the MLR model's ability to maintain its simplicity and avoid overfitting, thus preserving its predictive accuracy on unseen data [63].

The XGBoost-ANN model records an RMSE of 0.121, an R^2 of 0.799, a PCC of 0.667, an NSE of 0.799, and a WI of 0.715. While the model shows competitive RMSE, R^2, and NSE values, the lower PCC and particularly lower WI suggest challenges in accurately matching the patterns and variability in the testing dataset. The ANN component's complexity and capacity for nonlinear modeling might contribute to discrepancies in performance between training and testing, potentially indicating overfitting [69, 70].

In Figure 5, XGBoost-MLR seems to have a fair number of predictions close to the line of best fit on the training data, indicating good performance on the training set. However, the spread of points is wider in the testing data, suggesting that the model may not generalize as well on unseen data.

XGBoost-RF in Figure 6 shows a similar trend with a tighter cluster around the line of best fit for the training data, implying a better fit than the MLR model. However, XGBoost-ANN, in Figure 7, appears to perform comparably on the training data with a concentration of points near the line of best fit, but like the other models, it shows a broader spread in the testing data. This could indicate that while the model has learned the training data well, it may not be capturing the underlying pattern effectively enough to predict the testing data accurately.

While the predictive accuracy was generally satisfactory, certain combinations underperformed, possibly due to interactions between correlated variables. It is crucial to benchmark this predictive accuracy against the limited existing research in this area. For instance, the findings of the study differ significantly from what has been reported by Longa et al. [71], who developed ML models that achieved up to 80% accuracy in predicting energy poverty risk. Their research highlights income as the most significant predictor; however, incorporating additional socioeconomic factors is crucial for attaining high prediction reliability. Their findings suggest the intricate mechanisms that underlie energy poverty. van Hove et al. [15] also successfully employed a gradient-boosting classifier trained on a collection of socioeconomic features hypothesized to predict energy poverty across a diverse range of countries. Analysis of the classifier's internal model yielded new insights into the complexities of energy poverty, which aligns with the findings of the current study. The study confirms that in addition to income—the primary driver—household size also serves as a significant predictor.

Ensemble methods, such as RFs and boosting algorithms, have been widely praised for their superior predictive performance in various complex scenarios, including energy prediction and poverty mapping [72, 73]. These methods are known for reducing variance and bias, respectively, leading to more robust models. As shown in the testing phase, the XGBoost-RF model displayed exceptional accuracy (RMSE: 0.041, R^2: 0.975), aligning with the literature that suggests the efficacy of hybrid ensemble techniques in handling complex, nonlinear datasets [74]. However, the performance drop in the training phase (RMSE: 0.126, R^2: 0.796) potentially indicates overfitting, a common issue noted in highly complex models [67]. This suggests that while XGBoost-RF is powerful for generalization, it requires careful tuning to balance training performance and prevent overfitting, as discussed by Kotsiantis [75].

The performance limitations of linear regression models in handling nonlinear relationships are well-documented [76, 77]. These models often struggle in complex predictive tasks where interactions among predictors are non-linear [76, 78]. The XGBoost-MLR model's lower performance metrics in both phases (testing: RMSE: 0.101, R^2: 0.845; training: RMSE: 0.113, R^2: 0.808) illustrate the challenges linear components face in complex environments. This finding reinforces the argument that while MLR can provide insights into linear associations, it is less effective for datasets requiring the modeling of complex, multidimensional interactions, as seen in MEPI estimations.

ANNs are celebrated for their ability to approximate any nonlinear function [79, 80]. However, their performance can be highly variable and dependent on sufficient training data and proper architecture setup [81, 82]. The middle-ground performance of the XGBoost-ANN model in the testing

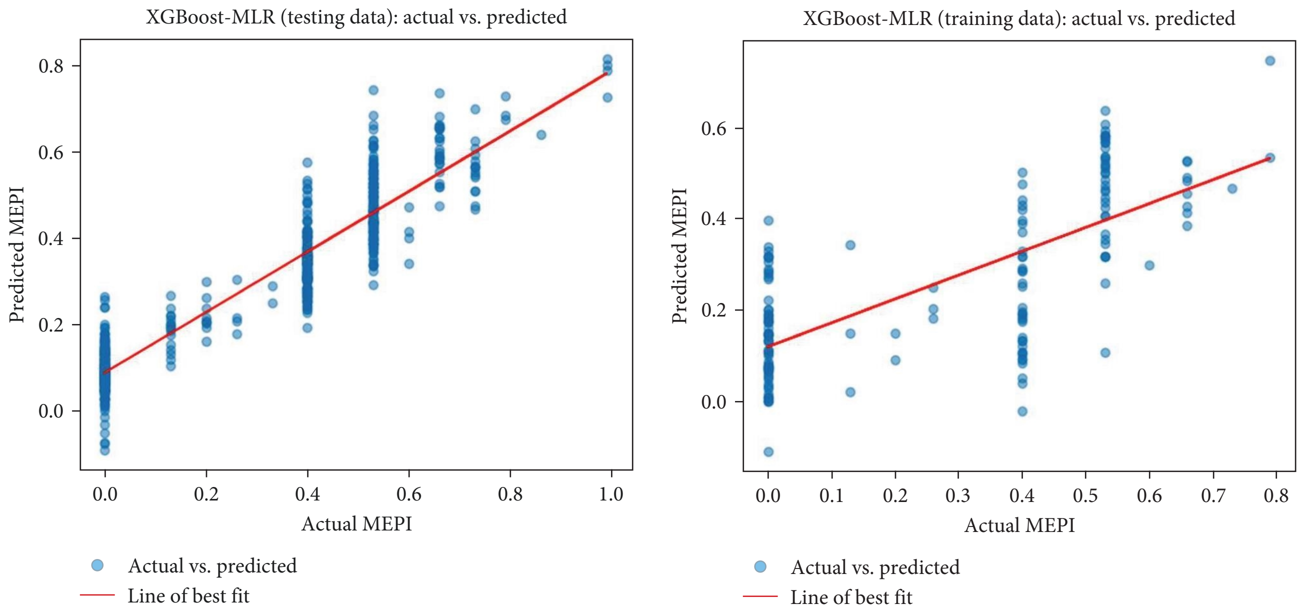


Figure 5: XGBoost-MLR ensemble model predictions for the testing and training datasets.

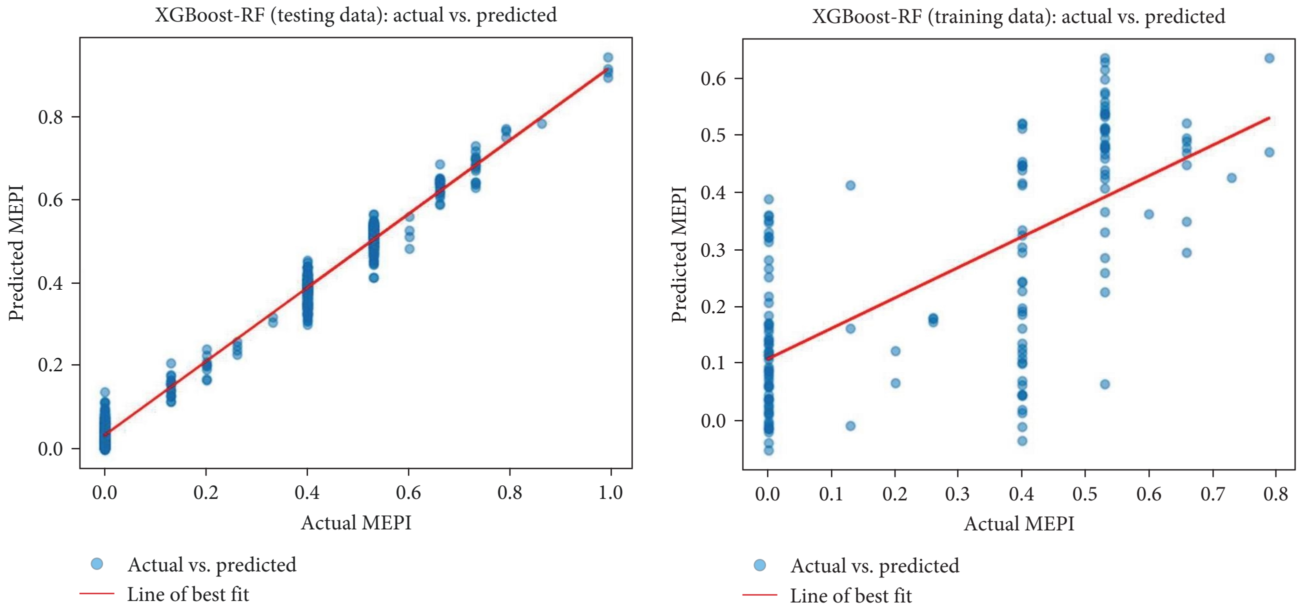


Figure 6: XGBoost-RF ensemble model predictions for the testing and training datasets.

phase (RMSE: 0.056, R^2: 0.954) suggests a strong capability but with notable limitations in generalization, as seen by its performance drop in the training phase (RMSE: 0.121, R^2: 0.799). This aligns with critiques in the literature about ANNs' sensitivity to overfitting and the need for large data sets and meticulous tuning [81, 82].

4.3. Feature Importance. Figure 8 presents a compelling narrative regarding the determinants of energy poverty as measured by the MEPI. It reveals the hierarchy of variables impacting energy poverty, with "Education" firmly at the pinnacle. This singular prominence underscores the critical role of education in energy poverty alleviation efforts. Such a dominant influence suggests that educational attainment may significantly empower individuals to navigate and mitigate the challenges associated with energy scarcity. Following the "Education" feature are "FCS," "HFIA," and "DDS." These parameters collectively form a cluster of nutrition-related indicators. Their substantive contribution to the model's predictive power underscores the intricate nexus between nutritional security and energy access. This interplay may reflect the broader socioeconomic canvas, where the ability to secure adequate

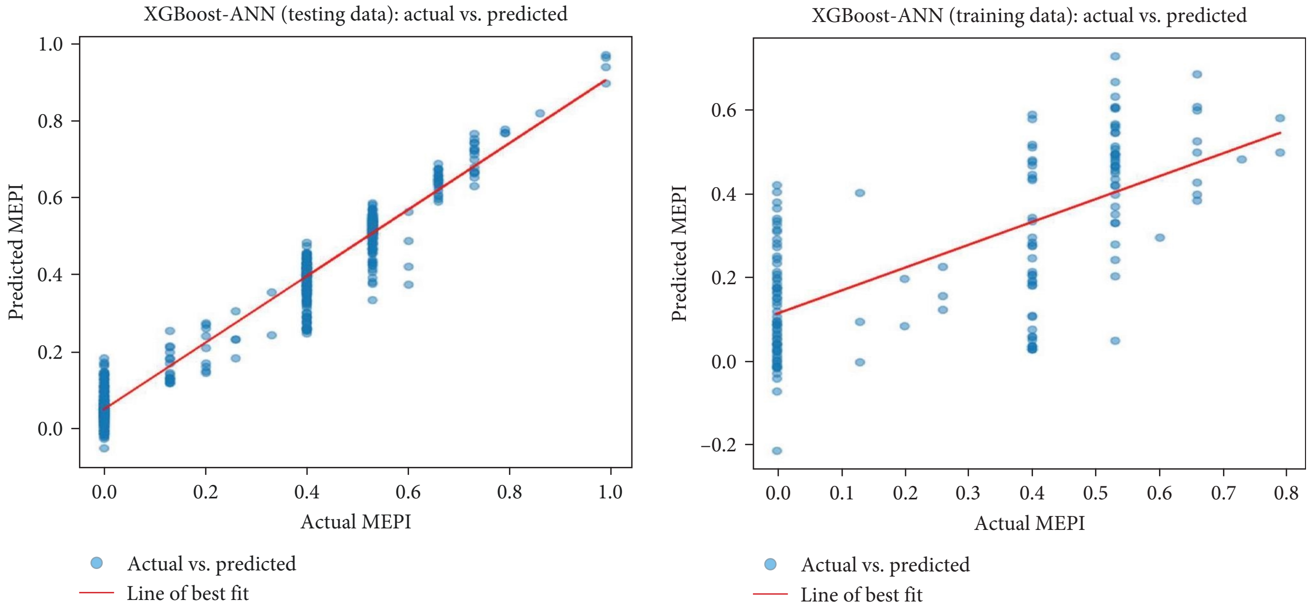


FIGURE 7: XGBoost-ANN ensemble model predictions for the training and testing datasets.

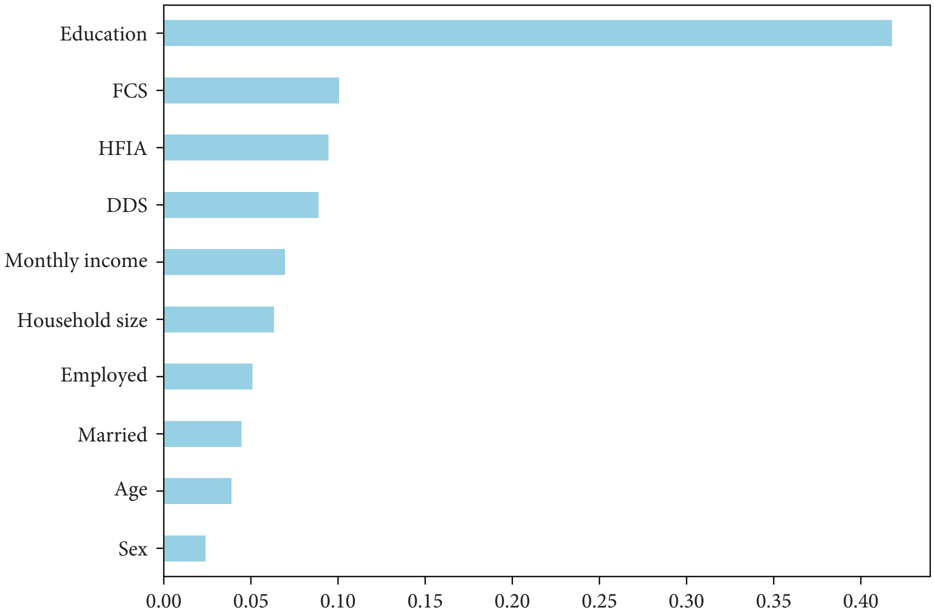


FIGURE 8: Feature importance for the predictions.

and diverse food options intertwines with the capacity to access and utilize energy resources.

"Monthly Income" and "Household Size" are identified as moderate influencers in the context of energy poverty, presenting a deviation from numerous studies documented in the existing literature [15, 71], which assert that income and household size are significant predictors of energy poverty. The model's sensitivity to these factors reaffirms that financial resources and household demographics shape energy poverty outcomes [20, 83]. While their impact is less pronounced than educational or nutritional indicators, their presence in the model's decision-making process reveals the multifaceted nature of energy poverty, which transcends a single domain.

In contrast, the variables "Employed" and "Married" occupy lower rungs on the feature importance ladder, yet they eclipse the very foundational demographic attributes of "Sex" and "Age." The relatively modest yet nonnegligible weight of employment and marital status may hint at the subtler socioeconomic currents influencing energy poverty. These factors, often reflective of household dynamics and earning potential, evidently play supporting roles in the overarching structure of energy poverty [84, 85].

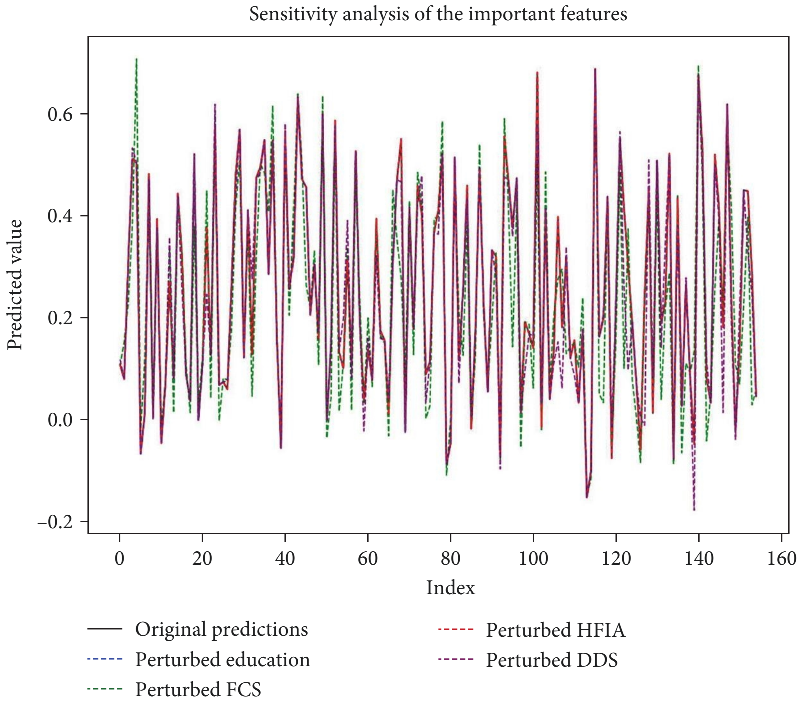


Figure 9: Sensitivity analysis for selected predictors.

At the base of the chart, "Sex" and "Age" are depicted as peripheral players, challenging common assumptions about the universality of demographic influence. Their minimal impact within the model's architecture signals a potential departure from traditional narratives that often emphasize demographic characteristics [86, 87]. This revelation prompts a re-evaluation of target demographics in energy poverty interventions and encourages a pivot toward more influential factors identified by the model.

4.4. Sensitivity Analysis. Figure 9 presents the impact that variations in key features have on the predicted values of MEPI as modeled by an ML algorithm. It is a testament to the intricate and nuanced relationship between input variables and their consequent predictions. At the forefront of this analysis is the perturbation of the "Education," depicted by the blue dashed line, which noticeably diverges from the original predictions shown by the solid purple line. This divergence is not just a mere fluctuation but a pronounced shift, underscoring the pivotal role of education in determining energy poverty outcomes, as previously revealed. This is in concordance with the feature importance analysis, where "Education" was identified as the most significant predictor, highlighting the potential for educational attainment to serve as a lever in addressing energy poverty.

Similarly, the sensitivity of the model to changes in "FCS" and "HFIA," represented by the green and red dashed lines, respectively, reveals the strong link between food security measures and energy poverty. The variations in predictions upon perturbing these features signal the model's recognition of food security as a vital component of energy poverty. This relationship is well-documented in the literature, where the security of food access has been repeatedly correlated with socioeconomic well-being, which encompasses energy access [88, 89].

The "DDS" feature, illustrated by the magenta dashed line, also shows significant variability in model predictions, akin to the "FCS" and "HFIA" features. This underscores the importance of dietary diversity, further cementing the interplay between food-related metrics and energy poverty. The overlapping nature of the perturbed lines for "FCS," "HFIA," and "DDS" suggests a potential interdependence between these features, indicative of the multifaceted and interconnected factors that contribute to energy poverty. Moreover, the unique pattern of deviation observed for the "Education" feature implies a potential nonlinear relationship with MEPI, suggesting that educational interventions might yield complex but potentially profound impacts on energy poverty.

In essence, this sensitivity analysis serves as a crucial step in model validation, providing confidence not just in the model's current predictive power but also in its potential reliability under varying conditions. It reveals the depth of the model's internal mechanics, where each input feature does not operate in isolation but in concert with others, contributing to the rich tapestry of predictors that delineate the landscape of energy poverty.

An in-depth sensitivity analysis of additional parameters can be found in the Supplementary Materials (refer to Figures S1, S2, and S3), offering an extensive assessment of the stability of the predictive modeling techniques employed in this study.

5. Conclusion and Policy Implications

5.1. Conclusion. This study explores the predictive modeling of the MEPI employing a variety of ML ensemble techniques.

Through this study, the critical socioeconomic issue of energy poverty, which impacts numerous communities worldwide, has been examined, highlighting the importance of several key parameters and their interrelations with energy poverty. This indicates that the interactions between energy poverty and the socioeconomic variables examined in the study are inherently complex, a conclusion also supported by Baker et al. [90] and Longa et al. [71].

The study identified the XGBoost-RF ensemble as the standout model during the training phase due to its superior predictive capabilities. This model's proficiency in deciphering complex data patterns underscores its potential in informing policies and interventions aimed at mitigating energy poverty. However, the subsequent testing phase shifted the performance hierarchy slightly, with the XGBoost-MLR ensemble showcasing better generalizability, as reflected in marginally superior R^2 and WI scores. The XGBoost-ANN ensemble found a balance between these extremes, illustrating the strengths and limitations of neural networks in predicting energy poverty.

A key finding from the feature importance analysis was the significant role of "Education" in determining MEPI, corroborating existing literature that highlights education's impact on socioeconomic outcomes. Furthermore, nutritional indicators such as the FCS, HFIA, and DDS emerged as critical, highlighting the link between food security and energy access. Sensitivity analysis reaffirmed the model's dependency on these features, offering insights into the reliability and resilience of the predictions against changes in input data.

This analysis emphasizes the complex nature of energy poverty and the necessity of employing diverse methodological approaches for its understanding and prediction. The successful integration of XGBoost with RF, MLR, and ANN techniques presents a promising path forward for the creation of sophisticated predictive tools. These tools can significantly aid in the strategic distribution of resources and the crafting of targeted interventions, thereby enriching the academic discourse on energy poverty and showcasing the potential of ML in policy formulation.

Nevertheless, the cross-sectional design of this study limits the ability to establish causality or track the evolution of energy poverty over time. Longitudinal data could provide a clearer picture of causality and trend analysis, offering a deeper understanding of how shifts in predictor variables affect energy poverty levels. Moreover, the study's scope does not fully extend to the broader social and environmental ramifications of energy poverty, which encompasses significant health, educational, and environmental outcomes. Future research should embrace these wider impacts to furnish a comprehensive view of energy poverty and its extensive effects.

Additionally, while the study lays a statistical groundwork for predicting energy poverty, it stops short of assessing the effectiveness of specific intervention strategies. Future research should aim to bridge this gap by linking predictive modeling with the empirical evaluation of policy measures to identify the most efficacious strategies for combating energy poverty.

Future research should adopt a multidimensional approach, incorporating diverse contexts, objective data, and longitudinal studies. Such efforts should focus on highlighting the causal mechanisms of energy poverty and evaluating intervention efficacy, ultimately enhancing the practical application of predictive modeling in this vital domain of social policy.

5.2. Policy Implications. The study carries profound policy implications, particularly for stakeholders seeking to address the multifaceted challenges of energy poverty. The key findings highlight the importance of leveraging predictive modeling to inform and prioritize policy interventions. The prominence of "Education" as a determinant of energy poverty suggests that educational policies could be a critical lever in mitigating energy poverty. Investments in educational infrastructure, access to quality education, and adult literacy programs could empower individuals with the knowledge and skills necessary to improve their energy access and usage.

The significance of "FCS," "HFIA," and "DDS" in predicting energy poverty indicates that food security is inextricably linked to energy access. Policies that aim to secure reliable and diverse food supplies could inadvertently contribute to reducing energy poverty. Initiatives could include agricultural support programs, subsidies for nutritionally rich foods, and the development of sustainable food systems that also incorporate energy-efficient practices.

The moderate influence of "Monthly Income" and "Household Size" on energy poverty points to the potential benefits of economic and social support measures [7, 91]. Financial assistance programs, targeted subsidies for energy-efficient appliances, and support for energy-efficient housing could directly impact households' energy poverty levels. Additionally, social programs that support employment and stable family structures could indirectly contribute to energy security.

The policy implications arising from this study are as multidimensional as the issue of energy poverty itself. The adoption of data-driven approaches, as demonstrated in this study, has the potential to revolutionize policy formulation and implementation, leading to more precise, impactful, and sustainable energy poverty alleviation efforts.

Abbreviations

ANN:	Artificial neural network
DDS:	Household dietary diversity score
FCS:	Food consumption score
HFIA:	Household fuel insecurity access
MAE:	Mean average error
MEPI:	Multidimensional energy poverty index
ML:	Machine learning
MLR:	Multiple linear regression
MSE:	Mean squared error
NSE:	Nash–Sutcliffe coefficient
PCC:	Pearson correlation coefficient
RF:	Random forests
RMSE:	Root mean squared error
WI:	Willmott's index
XGBoost:	Extreme gradient boosting.

Data Availability

Data will be available upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Sidique Gawusu was responsible for conceptualization, writing—original draft, software, methodology, visualization, data curation, investigation, and project administration. Abubakari Ahmed was responsible for writing—review and editing, and investigation, project administration. Seidu Abdulai Jamatutu was responsible for writing—review and editing, and investigation, project administration.

Supplementary Materials

The supplementary material further substantiates the model's validity by illustrating its performance dynamics when subjected to various parameter changes, ensuring that stakeholders can trust the model's predictions in diverse scenarios. This type of analysis is crucial for confirming the model's adaptability and fairness, particularly in applications with significant socioeconomic impacts. (*Supplementary Materials*)

References

[1] M. A. Okyere and B. Lin, "Invisible among the vulnerable: a nuanced perspective of energy poverty at the intersection of gender and disability in South Africa," *Humanities and Social Sciences Communications*, vol. 10, no. 1, Article ID 227, 2023.

[2] S. Ngarava, L. Zhou, T. Ningi, M. M. Chari, and L. Mdiya, "Gender and ethnic disparities in energy poverty: the case of South Africa," *Energy Policy*, vol. 161, Article ID 112755, 2022.

[3] R. Sonnino, "The new geography of food security: exploring the potential of urban food strategies," *The Geographical Journal*, vol. 182, no. 2, pp. 190–200, 2016.

[4] M. K. Mahalik, H. Mallick, and H. Padhan, "Do educational levels influence the environmental quality? The role of renewable and non-renewable energy demand in selected BRICS countries with a new policy perspective," *Renewable Energy*, vol. 164, pp. 419–432, 2021.

[5] M. Igawa and S. Managi, "Energy poverty and income inequality: an economic analysis of 37 countries," *Applied Energy*, vol. 306, Article ID 118076, 2022.

[6] F. Rao, Y. M. Tang, K. Y. Chau, W. Iqbal, and M. Abbas, "Assessment of energy poverty and key influencing factors in N11 countries," *Sustainable Production and Consumption*, vol. 30, pp. 1–15, 2022.

[7] A.-R. Qurat-ul-Ann and F. M. Mirza, "Determinants of multidimensional energy poverty in Pakistan: a household level analysis," *Environment, Development and Sustainability*, vol. 23, no. 8, pp. 12366–12410, 2021.

[8] K. Abbas, K. M. Butt, D. Xu et al., "Measurements and determinants of extreme multidimensional energy poverty using machine learning," *Energy*, vol. 251, Article ID 123977, 2022.

[9] K. Wang, Y.-X. Wang, K. Li, and Y.-M. Wei, "Energy poverty in China: an index based comprehensive evaluation," *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 308–323, 2015.

[10] G. Y. Obeng, H.-D. Evers, F. O. Akuffo, I. Braimah, and A. Brew-Hammond, "Solar photovoltaic electrification and rural energy-poverty in Ghana," *Energy for Sustainable Development*, vol. 12, no. 1, pp. 43–54, 2008.

[11] Z. Hong and I. K. Park, "Comparative analysis of energy poverty prediction models using machine learning algorithms," *Journal of Korea Planning Association*, vol. 56, no. 5, pp. 239–255, 2021.

[12] B. Tundys and A. Bretyn, "Energy transition scenarios for energy poverty alleviation: analysis of the delphi study," *Energies*, vol. 16, no. 4, Article ID 1870, 2023.

[13] U. Ruiz-Rivas, Y. Tahri, M. M. Arjona, M. Chinchilla, R. Castaño-Rosa, and J. Martínez-Crespo, "Energy poverty in developing regions: strategies, indicators, needs, and technological solutions," in *Energy Poverty Alleviation*, pp. 17–39, Springer International Publishing, Cham, 2022.

[14] S. Gawusu, S. A. Jamatutu, X. Zhang et al., "Spatial analysis and predictive modeling of energy poverty: insights for policy implementation," *Environment, Development and Sustainability*, 2024.

[15] W. van Hove, F. D. Longa, and B. van der Zwaan, "Identifying predictors for energy poverty in Europe using machine learning," *Energy and Buildings*, vol. 264, Article ID 112064, 2022.

[16] S. Gawusu, M. S. Tando, A. Ahmed et al., "Decentralized energy systems and blockchain technology: implications for alleviating energy poverty," *Sustainable Energy Technologies and Assessments*, vol. 65, Article ID 103795, 2024.

[17] B. Legendre and O. Ricci, "Measuring fuel poverty in France: which households are the most fuel vulnerable?" *Energy Economics*, vol. 49, pp. 620–628, 2015.

[18] M. N. Rajić, M. B. Milovanović, D. S. Antić, R. M. Maksimović, P. M. Milosavljević, and D. L. Pavlović, "Analyzing energy poverty using intelligent approach," *Energy & Environment*, vol. 31, no. 8, pp. 1448–1472, 2020.

[19] J. C. Romero, P. Linares, and X. López, "The policy implications of energy poverty indicators," *Energy Policy*, vol. 115, pp. 98–108, 2018.

[20] K. Abbas, S. Li, D. Xu, K. Baz, and A. Rakhmetova, "Do socioeconomic factors determine household multidimensional energy poverty? Empirical evidence from South Asia," *Energy Policy*, vol. 146, Article ID 111754, 2020.

[21] W. Hurst, C. A. C. Montanez, and N. Shone, "Towards an approach for fuel poverty detection from gas smart meter data using decision tree learning," in *IMMS '20: Proceedings of the 3rd International Conference on Information Management and Management Science*, pp. 23–28, ACM, New York, NY, USA, 2020.

[22] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, Article ID 105524, 2020.

[23] S. Gawusu, "Impact of renewable energy integration on commodity markets," *SSRN Electronic Journal*, 2024.

[24] T. Jayalakshmi and A. Santhakumaran, "Statistical normalization and back propagationfor classification," *International Journal of Computer Theory and Engineering*, pp. 89–93, 2011.

[25] P. I. Korah, A. M. Nunbogu, and B. A. A. Akanbang, "Spatio-temporal dynamics and livelihoods transformation in Wa, Ghana," *Land Use Policy*, vol. 77, pp. 174–185, 2018.

[26] I. Wilks, *Wa and the Wala: Islam and Polity in Northwestern Ghana*, Cambridge University Press, Cambridge, 1989.

[27] Ghana Statistical Service, "2010 population & housing census: district analytical report," 2014, Standfors Libraries (accessed February 26, 2024). https://searchworks.stanford.edu/view/11551814.

[28] A. Ahmed, "Urban water-energy-food nexus in the kitchen and social practices of diet and cooking: implications for household sustainability," *Environment, Development and Sustainability*, 2024.

[29] M. I. Dzudzor and N. Gerber, "Urban households' food safety knowledge and behaviour: choice of food markets and cooking practices," *Journal of Agriculture and Food Research*, vol. 14, Article ID 100728, 2023.

[30] D. A. Azorliade, D. K. Twerefou, and D. B. K. Dovie, "The impact of household cooking fuel choice on healthcare expenditure in Ghana," *Frontiers in Environmental Science*, vol. 10, Article ID 861204, 2022.

[31] A. Ahmed, S. B. Asabere, E. A. Adams, and Z. Abubakari, "Patterns and determinants of multidimensional poverty in secondary cities: implications for urban sustainability in African cities," *Habitat International*, vol. 134, Article ID 102775, 2023.

[32] P. Nussbaumer, M. Bazilian, and V. Modi, "Measuring energy poverty: focusing on what matters," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 1, pp. 231–243, 2012.

[33] K. P. Singh, N. Basant, and S. Gupta, "Support vector machines in water quality management," *Analytica Chimica Acta*, vol. 703, no. 2, pp. 152–162, 2011.

[34] J. Heaton, S. McElwee, J. Fraley, and J. Cannady, "Early stabilizing feature importance for TensorFlow deep neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 4618–4624, IEEE, Anchorage, AK, USA, 2017.

[35] M. Ali, R. Jiang, H. Ma et al., "Machine learning—a novel approach of well logs similarity based on synchronization measures to predict shear sonic logs," *Journal of Petroleum Science and Engineering*, vol. 203, Article ID 108602, 2021.

[36] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer New York, New York, NY, 2013.

[37] J. G. Dy and C. E. Brodley, "Feature selection (unsupervised learning)," *Journal of Machine Learning Research*, vol. 5, pp. 845–889, 2004.

[38] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature selection for text classification: a review," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3797–3816, 2019.

[39] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[40] A. Alsahaf, N. Petkov, V. Shenoy, and G. Azzopardi, "A framework for feature selection through boosting," *Expert Systems with Applications*, vol. 187, Article ID 115895, 2022.

[41] V. Allocca, M. Di Napoli, S. Coda et al., "A novel methodology for groundwater flooding susceptibility assessment through machine learning techniques in a mixed-land use aquifer," *Science of The Total Environment*, vol. 790, Article ID 148067, 2021.

[42] W. Chin, J.-H. Cheah, Y. Liu, H. Ting, X.-J. Lim, and T. H. Cham, "Demystifying the role of causal-predictive modeling using partial least squares structural equation modeling in information systems research," *Industrial Management & Data Systems*, vol. 120, no. 12, pp. 2161–2209, 2020.

[43] X. Wang, Y. Zhang, B. Yu et al., "Prediction of protein-protein interaction sites through eXtreme gradient boosting with kernel principal component analysis," *Computers in Biology and Medicine*, vol. 134, Article ID 104516, 2021.

[44] X. Xu and Y. Zhang, "Edible oil wholesale price forecasts via the neural network," *Energy Nexus*, vol. 12, Article ID 100250, 2023.

[45] K. A. Marill, "*Advanced Statistics*: linear regression, part II: multiple linear regression," *Academic Emergency Medicine*, vol. 11, no. 1, pp. 94–102, 2004.

[46] K. F. Nimon and F. L. Oswald, "Understanding the results of multiple linear regression," *Organizational Research Methods*, vol. 16, no. 4, pp. 650–674, 2013.

[47] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[48] E. Wang, "Decomposing core energy factor structure of U.S. commercial buildings through clustering around latent variables with random forest on large-scale mixed data," *Energy Conversion and Management*, vol. 153, pp. 346–361, 2017.

[49] F. W. Yu, W. T. Ho, K. T. Chan, and R. K. Y. Sit, "Critique of operating variables importance on chiller energy performance using random forest," *Energy and Buildings*, vol. 139, pp. 653–664, 2017.

[50] S. Hegelich, "Decision trees and random forests: machine learning techniques to classify rare events," *European Policy Analysis*, vol. 2, no. 1, pp. 98–120, 2016.

[51] J. Wu, L. Kong, M. Yi et al., "Prediction and screening model for products based on fusion regression and XGBoost classification," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 4987639, 14 pages, 2022.

[52] S. Abirami and P. Chitra, "Energy-efficient edge based real-time healthcare support system," in *Advances in Computers*, vol. 117, pp. 339–368, Elsevier, 2020.

[53] N. Baig, J. Usman, S. I. Abba, M. Benaafi, and I. H. Aljundi, "Fractionation of dyes/salts using loose nanofiltration membranes: insight from machine learning prediction," *Journal of Cleaner Production*, vol. 418, Article ID 138193, 2023.

[54] H. Tao, S. I. Abba, A. M. Al-Areeq et al., "Hybridized artificial intelligence models with nature-inspired algorithms for river flow modeling: a comprehensive review, assessment, and possible future research directions," *Engineering Applications of Artificial Intelligence*, vol. 129, Article ID 107559, 2024.

[55] L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni, "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors," *Nature Biotechnology*, vol. 36, no. 5, pp. 421–427, 2018.

[56] M. D. Luecken, M. Büttner, K. Chaichoompu et al., "Benchmarking atlas-level data integration in single-cell genomics," *Nature Methods*, vol. 19, no. 1, pp. 41–50, 2022.

[57] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, "A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting," *Annals of Data Science*, vol. 10, no. 1, pp. 183–208, 2023.

[58] E. M. Dos Santos, R. Sabourin, and P. Maupin, "Overfitting cautious selection of classifier ensembles with genetic algorithms," *Information Fusion*, vol. 10, no. 2, pp. 150–162, 2009.

[59] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: a review with examples from ecology," *Ecological Monographs*, vol. 93, no. 1, Article ID e1557, 2023.

[60] B. Lin and M. A. Okyere, "Does energy poverty affect the well-being of people: evidence from Ghana," *Sustainable Production and Consumption*, vol. 28, pp. 675–685, 2021.

[61] K. Abbas, X. Xie, D. Xu, and K. M. Butt, "Assessing an empirical relationship between energy poverty and domestic

health issues: a multidimensional approach,” *Energy*, vol. 221, Article ID 119774, 2021.

[62] T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, New York, NY, USA, 2016.

[63] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, New York, NY: US, 2021.

[64] R. Xiao, T. Zayed, M. A. Meguid, and L. Sushama, “Improving failure modeling for gas transmission pipelines: a survival analysis and machine learning integrated approach,” *Reliability Engineering & System Safety*, vol. 241, Article ID 109672, 2024.

[65] M. Abed, M. A. Imteaz, A. N. Ahmed, and Y. F. Huang, “Modelling monthly pan evaporation utilising random forest and deep learning algorithms,” *Scientific Reports*, vol. 12, no. 1, Article ID 13132, 2022.

[66] A. Gbadamosi, H. Adamu, J. Usman et al., “New-generation machine learning models as prediction tools for modeling interfacial tension of hydrogen-brine system,” *International Journal of Hydrogen Energy*, vol. 50, pp. 1326–1337, 2024.

[67] D. M. Hawkins, “The problem of overfitting,” *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[68] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[69] V. Asghari, Y. F. Leung, and S.-C. Hsu, “Deep neural network based framework for complex correlations in engineering metrics,” *Advanced Engineering Informatics*, vol. 44, Article ID 101058, 2020.

[70] S. Bedi, A. Samal, C. Ray, and D. Snow, “Comparative evaluation of machine learning models for groundwater quality assessment,” *Environmental Monitoring and Assessment*, vol. 192, no. 12, Article ID 776, 2020.

[71] F. D. Longa, B. Sweerts, and B. van der Zwaan, “Exploring the complex origins of energy poverty in The Netherlands with machine learning,” *Energy Policy*, vol. 156, Article ID 112373, 2021.

[72] S. González, S. García, J. Del Ser, L. Rokach, and F. Herrera, “A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities,” *Information Fusion*, vol. 64, pp. 205–237, 2020.

[73] M. Hamza and D. Larocque, “An empirical comparison of ensemble methods based on classification trees,” *Journal of Statistical Computation and Simulation*, vol. 75, no. 8, pp. 629–643, 2005.

[74] J. Xiao, Y. Li, L. Xie, D. Liu, and J. Huang, “A hybrid model based on selective ensemble for energy consumption forecasting in China,” *Energy*, vol. 159, pp. 534–546, 2018.

[75] S. Kotsiantis, “Combining bagging, boosting, rotation forest and random subspace methods,” *Artificial Intelligence Review*, vol. 35, no. 3, pp. 223–240, 2011.

[76] J. Zhang, E. B. Martin, and A. J. Morris, “Process monitoring using non-linear statistical techniques,” *Chemical Engineering Journal*, vol. 67, no. 3, pp. 181–189, 1997.

[77] D. Perezmarin, A. Garridovaro, and J. Guerrero, “Non-linear regression methods in NIRS quantitative analysis,” *Talanta*, vol. 72, no. 1, pp. 28–42, 2007.

[78] H. J. Motulsky and L. A. Ransnas, “Fitting curves to data using nonlinear regression: a practical and nonmathematical review,” *The FASEB Journal*, vol. 1, no. 5, pp. 365–374, 1987.

[79] A. Krenker, J. Bešter, and A. Kos, “Introduction to the artificial neural networks,” in *Artificial Neural Networks: Methodological Advances and Biomedical Applications*, K. Suzuki, Ed., pp. 1–18, InTech, 2011.

[80] Y.-S. Park and S. Lek, “Artificial neural networks: multilayer perceptron for ecological modeling,” in *Developments in Environmental Modelling*, vol. 28, pp. 123–140, Elsevier, 2016.

[81] J. Rabault, M. Kuchta, A. Jensen, U. Réglade, and N. Cerardi, “Artificial neural networks trained through deep reinforcement learning discover control strategies for active flow control,” *Journal of Fluid Mechanics*, vol. 865, pp. 281–302, 2019.

[82] R. Pino-Mejías, A. Pérez-Fargallo, C. Rubio-Bellido, and J. A. Pulido-Arcas, “Artificial neural networks and linear regression prediction models for social housing allocation: fuel poverty potential risk index,” *Energy*, vol. 164, pp. 627–641, 2018.

[83] R. Banerjee, V. Mishra, and A. A. Maruta, “Energy poverty, health and education outcomes: evidence from the developing world,” *Energy Economics*, vol. 101, Article ID 105447, 2021.

[84] N. Longhurst and T. Hargreaves, “Emotions and fuel poverty: the lived experience of social housing tenants in the United Kingdom,” *Energy Research & Social Science*, vol. 56, Article ID 101207, 2019.

[85] N. Simcock, J. Frankowski, and S. Bouzarovski, “Rendered invisible: institutional misrecognition and the reproduction of energy poverty,” *Geoforum*, vol. 124, pp. 1–9, 2021.

[86] M. Jayasinghe, E. A. Selvanathan, and S. Selvanathan, “Energy poverty in Sri Lanka,” *Energy Economics*, vol. 101, Article ID 105450, 2021.

[87] M. Hasanujzaman and M. A. Omar, “Household and non-household factors influencing multidimensional energy poverty in Bangladesh: demographics, urbanization and regional differentiation via a multilevel modeling approach,” *Energy Research & Social Science*, vol. 92, Article ID 102803, 2022.

[88] B. Wu, S. Liu, J. Wang, S. Tahir, and A. K. Patwary, “Assessing the mechanism of energy efficiency and energy poverty alleviation based on environmental regulation policy measures,” *Environmental Science and Pollution Research*, vol. 28, no. 30, pp. 40858–40870, 2021.

[89] J. M. Fry, L. Farrell, and J. B. Temple, “Energy poverty and food insecurity: is there an energy or food trade-off among low-income Australians?” *Energy Economics*, vol. 123, Article ID 106731, 2023.

[90] K. J. Baker, R. Mould, and S. Restrick, “Rethink fuel poverty as a complex problem,” *Nature Energy*, vol. 3, no. 8, pp. 610–612, 2018.

[91] Y. Li, K. Chen, R. Ding, J. Zhang, and Y. Hao, “How do photovoltaic poverty alleviation projects relieve household energy poverty? Evidence from China,” *Energy Economics*, vol. 118, Article ID 106514, 2023.