Electricity Consumer Characteristics Identification

A Federated Learning Approach

Wang, Yi; Bennani, Imane Lahmam; Liu, Xiufeng; Sun, Mingyang; Zhou, Yao

Published in:
IEEE Transactions on Smart Grid

Link to article, DOI:
10.1109/TSG.2021.3066577

Publication date:
2021

Document Version
Peer reviewed version

Electricity Consumer Characteristics Identification: A Federated Learning Approach

Yi Wang, Imane Lahmam Bennani, Xiufeng Liu, Mingyang Sun, Yao Zhou,

***Abstract*—Nowadays, smart meters are deployed in millions of residential households to gain significant insights from fine-grained electricity consumption data. The information extracted from smart meter data enables utilities to identify the socio-demographic characteristics of electricity consumers and then offer them diversified services. Traditionally, this task is implemented in a centralized manner with the assumption that utilities have access to all the smart meter data. However, smart meter data are measured and owned by different retailers in the retail market who may not be willing to share their data. To this end, a distributed electricity consumer characteristics identification method is proposed based on federated learning, which can preserve the privacy of retailers. Specifically, privacy-perseverance principal component analysis (PCA) is exploited to extract features from smart meter data. On this basis, an artificial neural network is trained in a federated manner with three weighted averaging strategies to bridge between smart meter data and the socio-demographic characteristics of consumers. Case studies on the Irish Commission for Energy Regulation (CER) dataset verify that the proposed federated method has comparable performance with the centralized model on both balanced and unbalanced datasets.**

***Index Terms*—Federated learning, smart meter, data analytics, privacy-perseverance, socio-demographic characteristics**

I. Introduction

ELECTRICITY consumers and suppliers have access to an immense amount of fine-grained electricity data with the increased deployment of residential smart meters. These data are usually recorded at a regular interval, such as every 15 min, and carry valuable information on the electricity consumption behavior of consumers. Smart data analytics help in revealing hidden information and enable other uses to smart meters than billing. More specifically, the analysis of smart meter data allows for systems efficiency and energy savings. Grid management systems can have a better understanding of consumer consumption patterns, cope with peak usage [1], and coordinate consumption for an eased integration of renewable energy sources (RESs) [2]. From retailer's point of view in the retail market, smart meter data analytics can help them have more transparency on the electricity consumption behavior of consumers, thus providing diversified services to them. A substantial number of studies have been conducted on smart meter data analytics towards different applications ranging from load profiling, demand response, load forecasting, tariff design, and household characteristics identification [3].

Various machine learning techniques such as clustering, classification, and regression are applied to smart meter data to extract significant insights and hidden patterns. In addition to dynamic energy consumption data, the use of static data such as socio-demographic information of consumers and household characteristics is being studied as a driver of residential energy consumption. Studies have shown that consumer socio-demographic status has a noticeable influence on their consumption behavior [4]. Conversely, electricity consumption is also a consumer-oriented process that could be derived from consumer behavior and socio-economic information [5]. Crossing information retrieved in smart meters with key alternative information on consumers allows for a market segmentation based on the similarity of consumers' needs and behavior. With the insights on socio-economic characteristics of individual consumers, retailers can also provide personalized consumption feedback and leverage this information for effective decision making [6].

Analyzing the factors that drive energy consumption has seen an increasing interest over the past few years. This is mainly due to the availability of adequate data allowed by growing deployment of smart meters, real-time smart-home energy monitoring services, and the implementation of domestic smart metering campaigns. Hayn et al. focused on the impact of socio-demographic factors and equipment with electric appliances on residential load profiles in [7]. Along the same lines, McLoughlin et al. linked household characteristics to a series of Profile Classes (PCs) by applying Self-Organizing Maps (SOMs) and multi-nominal logistic regression on the data [8]. Tong et al. introduced the concept of energy behavior indicators to evaluate the correlation between energy behavior and household information using wavelet analysis and X-means clustering [9]. In a more recent study, Sun et al. used a concurrent k-means and spectral clustering (CKSC) method to identify the patterns of household electricity consumption, and to infer them based on dwelling, family, and household characteristics [10].

In order to characterize the electricity consumption behavior, clustering analysis is commonly applied to smart meter data as an unsupervised learning technique. The applications

This work was supported in part by the National Natural Science Foundation of China under Grant U20A20159, the CCF-Tencent Open Fund, the ERA-NET project, Flexibility for Smart Urban Energy Systems (FlexSUS) (91352), and the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform). (*Corresponding author: Yao Zhou.*)

Yi Wang is with Power Systems Laboratory, ETH Zürich, 8092 Zurich, Switzerland.

Imane Lahmam Bennani is with Department of Mechanical and Process Engineering, ETH Zürich, 8092 Zurich, Switzerland.

Xiufeng Liu is with Department of Technology, Management and Economics, Technical University of Denmark, 2800 Lyngby, Denmark.

Mingyang Sun is with College of Control Science and Engineering, Zhejiang University, 310027 Zhejiang, China.

Yao Zhou is with School of Engineering, the University of Edinburgh, EH9 3FB, Edinburgh, UK.

of such a technique to daily residential load time series in the power system's literature are summarized in [11]. In recent studies, many authors have merged electricity data with additional information obtained through questionnaires. Most of these studies are based on data collected from massive smart metering systems conducted in different countries, as implementing such a study requires a significant budget. Gouveia et al. combined daily consumption data with 110-question door-to-door surveys, with a sample of 265 households in the city of Évora, Portugal in [12]. Their analysis shows that three major groups of determinants that are physical characteristics of a dwelling, equipment, and ownership of households, and occupant profile, are characteristics of electricity consumption segmentation. Wang et al. used a deep convolutional neural network (CNN) and a support vector machine (SVM) on massive load profiles of the Irish CER dataset to identify socio-demographic information of consumers in [6]. Beckel et al. introduced a framework to automatically classify household's properties using smart meter data in [13]. With their method called CLASS, they could infer 8 out of the 18 characteristics extracted with an accuracy of more than 70% over all households. A subsequent study by Hopf et al. [14] extended the CLASS framework by extracting 88 features instead of 18 from electricity consumption data to classify household characteristics.

Although several works have been done on the identification of the socio-demographic characteristics of electricity consumers, the identification algorithms are implemented in a centralized manner, assuming that all data from smart meters can be accessed from a data lake. However, in a retail market, the smart meter data are owned by different retailers, who may not be willing to share their data. In this situation, the consumer characteristics identification model cannot directly make use of the full dataset. At the legal level, governments or organizations worldwide are also increasingly committed to data privacy protection. For example, the European Union has started to enforce the General Data Protection Regulation (GDPR) since 2017 [15]. To address the privacy concern, federated learning has been proposed to train a learning model across multiple decentralized clients with local data samples, where the original data are not shared, thus the privacy can be persevered [16]. The basic idea of federated learning is that multiple distributed clients train the machine learning model individually using the local data, and then the server summarizes the trained models to form the final global model. In general, federated learning can be divided into three categories according to the distribution of datasets, i.e., horizontal federated learning, vertical federated learning, and federated transfer learning. The concept of federated learning was originally proposed by Google in 2016 [17], where Two approaches to reducing the uplink communication costs, i.e., structured updates and sketched updates under the federated framework. Since then, it has been applied in various fields, including mobile devices, industrial engineering, healthcare, and many others [18] [19]. A sparse ternary compression (STC) framework was studied in [20] to tackle the challenges brought by the independent and identically distributed (i.i.d.) client data. Federated learning has been widely applied in the Internet of Things (IoT). An edge federated learning (EdgeFed) approach was proposed in [21] to separate the local model updating task from mobile devices. How federated learning can be applied in the autonomous IoT to make full use of various data from smart end-user devices was discussed in [22]. An activity and resource-aware federated learning model was studied in [23] for distributed mobile robots in the IoT environment. A federated learning framework without a centralized cloud server was provided in [24] over wireless networks.

To the best of our knowledge, however, federated learning has rarely been applied in power and energy systems. It is of great importance to investigate the feasibility of using this powerful tool to realize the privacy-preserving electricity consumer characteristics identification. One related work is a federated load forecasting model for the electric vehicle (EV) networks studied in [25]. Another interesting work is a federated reinforcement learning (FRL) method for the energy management of multiple smart homes considering various appliances, rooftop PV, and energy storage [26]. In this paper, we extend the current state-of-the-art by developing a federated approach to identifying the characteristics of residential consumers using smart meter data and door-to-door surveys. The feature set is first established from smart meter data. On this basis, privacy-persevering principal component analysis (PCA) will be carried out to further extract the features. Classifiers based on artificial neural network (ANN) are then trained in a federated manner to predict the class labels for the selected relevant characteristics, which include occupant information (e.g., social class, number of residents), dwelling properties (e.g., type of home), and domestic appliances (e.g., light bulbs proportion).

To summarize, this paper makes the following contributions:

1) Proposing a federated framework for electricity consumer characteristics identification for the first time that can protect the privacy of the retailers without sharing smart meter data;
2) Providing a federated PCA approach for further feature extraction which can be implemented with a simple communication structure;
3) Investigating two performance-based weighted combination methods for the federated ANN model training and compared them with traditional data size-based weighted combination method;
4) Verifying the effectiveness of the proposed method by conducting comprehensive case studies on the Irish Commission for Energy Regulation (CER) open dataset.

The remainder of this paper is organized as follows: Section II introduces the Irish CER dataset and defines the federated consumer characteristics identification problem. Section III provides the framework and details the algorithms for the identification task. Section IV conducts case studies and makes comparisons to verify the effectiveness of our proposed method. Section V draws the conclusions.

II. Problem Statement

This section first introduces the dataset to be studied and then defines the problem to be addressed in this paper.

A. Irish CER Dataset

Our study is based on the Irish CER dataset [27], which was collected by CER as part of the Smart Metering Project (2007) to assess the performance of smart meters and their impact on consumers' behavior in Ireland. The dataset contains raw smart meter data gathered with a 30-min granularity from private households for 536 days from July 2009 to December 2010. Among the households that took part in this experiment, a total of 4,232 are residential consumers. Fig. 1 shows the weekly averaged load profiles of four consumers with distinct patterns in terms of both shape and magnitude.

In connection with the measurement of household electricity consumption, participants were also required to fill out a questionnaire containing questions on occupant socio-demographic status (e.g., employment status, social class), their consumption behavior (e.g., interest in reducing bill), household properties (e.g., floor area, number of bedrooms), and home appliances (e.g., number of washing machines). In this paper, 15 socio-demographic characteristics are selected from the 140 questions of the pre-trial survey. For better visualization and interpretability of the results, a restricted number of class labels have been assigned to each characteristic. Table I lists the selected electricity consumer characteristics to be identified. The chosen class labels and the number of consumers associated with each characteristic are also given. These characteristics are classified into three main categories: (1) occupant information, (2) dwelling properties, and (3) domestic appliance properties. In the survey sheet, the rows with missing data and properties' specific classes involving insignificant proportions of consumers were removed. From Table I, it can be seen that the class labels are very unbalanced for some of the characteristics. For example, 3,424 consumers do not live alone, while only 808 consumers live alone. This aspect of imbalance in the data will be taken into account in the algorithm design and performance evaluation of the characteristics identification model.

B. Federated Consumer Characteristics Identification

In a competitive retail market, consumers are served by N electricity retailers. The set and number of consumers served by the n-th retailer are denoted as C_n and c_n, respectively. The retailer has full access to the electricity consumption data $\mathbf{D}_n$ and the socio-demographic characteristics Y_n of these c_n consumers. The size of the matrix Y_n is $c_n \times K$, where $K = 15$ is the number of characteristics to be identified, and the vector $Y_{n,k}$ denotes the k-th characteristics of these c_n consumers.

The electricity consumption data and socio-demographic characteristics of a total of $M = \sum_{n=1}^{N} c_n$ consumers are $\mathbf{D} = [\mathbf{D}_1^T, \mathbf{D}_2^T, \ldots, \mathbf{D}_N^T]^T$ and $\mathbf{Y} = [Y_1^T, Y_2^T, \ldots, Y_N^T]^T$, respectively. Let's $\mathbf{Y}_k$ denote the k-th characteristics of all consumers. The electricity consumer characteristics identification problem is essentially a classification model that makes a bridge between the load profile $\mathbf{D}$ shown in Fig. 1 and the socio-demographic characteristics $\mathbf{Y}$ shown in Table I. Mathematically, a classification model f_k should be trained based on the load profiles $\mathbf{D}$ and the label $\mathbf{Y}_k$ for the k-th characteristics:

$$y_k = f_k(\omega_k, \mathbf{D}) \tag{1}$$

where ω_k and y_k denotes the parameters and label of the classification model f_k for the k-th characteristics, respectively.

Previous work proposed in [13] and [6] trained the classification or regression model f_k in a centralized manner, with both $\mathbf{D}$ and $\mathbf{Y}$ directly accessed from a data lake. However, in a competitive retail market, electricity consumption data and socio-demographic characteristics of consumers are the important digital assets of retailers. They are often unwilling to share these data. Therefore the characteristics identification model f_k has to be trained in a decentralized manner. The federated learning framework can be applied in this situation where only very limited information, i.e., $\omega_{k,n}$ and ω_k are exchanged between a computational center and retailers. The federated model for identifying consumer characteristics should address two main issues in a decentralized manner:

1) Since the original smart meter data $\mathbf{D}$ is of high dimensions, it cannot be directly fed into a classification model, and its important features should be first extracted before training the classification model;
2) An effective and highly accurate classification model f_k should be trained for each consumer characteristic.

III. Proposed Methodology

This section provides a framework and technical details to address the abovementioned two issues.

A. Framework

Feature extraction and model training are the two basic tasks for a classification model. For the federated consumer characteristics identification problem, extracting features from smart meter data and training classification models for different characteristics should also be implemented, but in a distributed way.

As shown in Fig. 2, the procedure of the proposed method consists of three main stages: 1) feature set formulation, where a set of features is manually calculated according to the basic understanding of electricity consumption behavior; 2) privacy-persevering feature extraction, where PCA is implemented in a distributed way to further extract the features as input to the regression model; 3) federated characteristics identification, where the federated ANN model is trained for this task. The details of the three stages are presented below.

B. Feature Set Formulation

The formulation of the feature set is a crucial step for time series classification problems. A well-performed feature extraction allows better accuracy and interpretability of the classification results [28]. The daily load profile of an individual electricity consumer is highly dynamic. The selected features of consumers should be stable and can reflect daily and weekly patterns. Therefore, the average weekly load profile is first calculated for each consumer. On this basis, the feature set is formulated manually based on the work carried out in [13] and

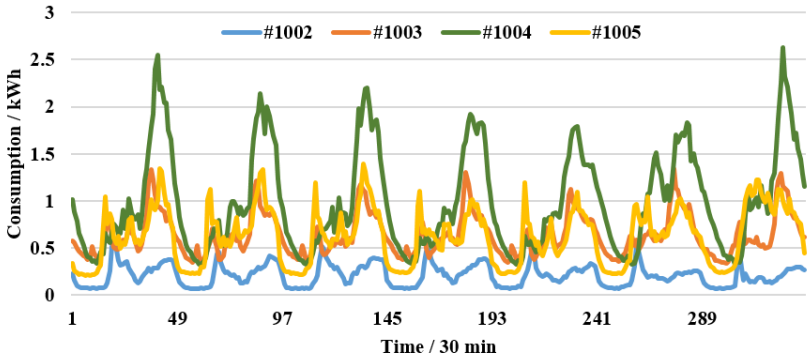

Fig. 1: Weekly averaged load profiles of four consumers.

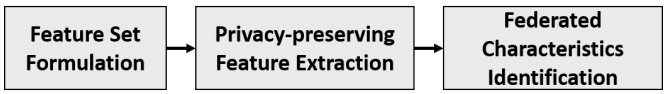

Fig. 2: Procedure of the proposed method.

[29]. The feature set contains three categories: statistics (e.g., total, mean, maximum, and minimum consumption), ratios (e.g., load factors of the whole week), and statistics (e.g., quantiles, entropy). The descriptive characteristics specific to time series data have been considered (e.g., kurtosis, skewness). The contrast in the energy consumption behavior of consumers on weekdays (Mon-Fri) and weekends are also taken into account separately, which are included in the feature set. Finally, the feature at different time intervals of the day and these are identified as baseload (2 am $-$ 5 am), morning (6 am $-$ 10 am), midday (11 am $-$ 2 pm), and evening (6 pm $-$ 10 pm).

An exhaustive list of features is provided in Table II. The listed features are computed for each consumer from weekly aggregated consumption data on weekdays and weekends over the whole trial period. A total of 73 features were calculated to form the final feature set. Each feature has a self-explanatory name. For instance, "mean_morning_wd" refers to the mean of electricity consumption on a weekday in the morning. Similarly, "max_base_wee" refers to the maximum consumption during baseload hours on a weekend day. "GE_0.5kW_proportion" denotes the proportion of electricity consumption values greater than or equal to 0.5kW over the week, while "GT_mean_proportion" means the proportion of consumption greater than the weekly mean. Last, "LF" refers to the Load Factor ratio.

If the retailers have consensus on how to formulate the feature set, this task can be performed by the retailers themselves without the need to share data. For simplicity, $\mathbf{D}_n$ denotes the feature set formulated by the n-th retailer instead of the original smart meter data in the rest of the paper.

C. Privacy-preserving Feature Extraction

Principle component analysis (PCA) is a commonly used dimension reduction method for extracting important features from the original data. The m-th line of the dataset $\mathbf{D}$, denoted as $\mathbf{d}_m$, corresponds to the characteristics of the m-th consumer and should be normalized to zero by subtracting the mean value of these characteristics $\mu = \frac{1}{M}\sum_{m=1}^{M}\mathbf{D}_m$. On this basis, the covariance matrix can be calculated as [30]:

$$\text{Cov} = \sum_{m=1}^{M}(\mathbf{d}_m - \mu)(\mathbf{d}_m - \mu)^T = \sum_{m=1}^{M}\mathbf{d}_m\mathbf{d}_m^T - M\mu\mu^T \tag{2}$$

If we denote $\mathbf{A}_n = \sum_{m\in C_n}\mathbf{d}_m\mathbf{d}_m^T$, $\mathbf{B}_n = \sum_{m\in C_n}\mathbf{d}_m$, Eq. (2) can be written as:

$$\text{Cov} = \sum_{n=1}^{N}\mathbf{A}_n - \frac{1}{\sum_{n=1}^{N}c_n}(\sum_{n=1}^{N}\mathbf{B}_n)(\sum_{n=1}^{N}\mathbf{B}_n)^T \tag{3}$$

The next step is to calculate the eigenvectors and eigenvalues of the covariance matrix Cov using singular value decomposition (SVD). A subset of the eigenvectors with top eigenvalues is formulated into a projection matrix $\mathbf{T}$ and the dimension reduced feature vector for the m-th consumer and feature matrix for the consumers served by the n-th retailer are calculated as:

$$\mathbf{d}'_m = \mathbf{d}_m\mathbf{T}, \quad \mathbf{D}'_n = \mathbf{D}_n\mathbf{T} \tag{4}$$

It can be seen from Eq. (3) that only $\sum_{n=1}^{N}\mathbf{A}_n$, $\sum_{n=1}^{N}\mathbf{B}_n$, and $\sum_{n=1}^{N}c_n$ are needed for every retailer to calculate the covariance matrix using additive operations. Thus additive homomorphic encryption can be used to implement the privacy-preserving feature extraction. The Paillier algorithm is an effective approach for the partial additive homomorphic encryption scheme [31]. It supports the addition of two encrypted integers and the multiplication of an encrypted integer by an unencrypted integer. However, sometimes the data to be encrypted are not an integer in the real world. For example, the matrix $\mathbf{A}_n$ and $\mathbf{B}_n$ are floating-point numbers. Fortunately, the Paillier cryptosystem can also be extended to floating-point numbers and can be realized in Python [32] [33].

The basic idea of the proposed federated PCA is shown in Fig. 3. In our proposed framework, there are two communication schemes marked by red and black arrows.

The red communication scheme is a sequential link starting from the first retailer and ending at the server. Specifically, a server is first established in the cloud, and the server generates a key pair, i.e., the public key and private key, where the

TABLE I: The socio-demographic and household characteristics to be studied

Socio-demographic characteristics	Class labels	Number
(1) Occupant information		
Age of income earner	1: Young <=35 2: Middle age 36-65 3: Old>65	436 2819 953
Employment status	1: Employed 2: Unemployed	2536 1696
Retirement status	1: Retired 2: Not retired	1285 2947
Social class	1: AB 2: C1-C2 3: DE	642 1840 1593
Have children	1: Yes 2: No	1229 3003
Live alone	1: Yes 2: No	808 3424
Number of residents	1: Few <=2 2: Many >2	2199 2033
(2) Dwelling properties		
Type of home	1: Detached house or bungalow 2: Semi-detached or terraced	2189 1964
Rented or owned	1: Rented 2: Owned	299 3921
Number of bedrooms	1: Low <=2 2: Middle =3 3: High >3	404 1884 1944
Age of house	1: Very new <=15 2: New >15 and <=30 3: Old >30	1667 790 1771
(3) Domestic appliance properties		
Cooking	1: Electric 2: Not electric	2960 1272
Number of home appliances	1: Low <=5 2: Middle >5 and <=8 3: High >8	1585 2251 396
Number of entertainment appliances	1: Low <=3 2: Middle (4,5) 3: High >5	2050 1290 892
Energy efficient light-bulb proportion	1: Up to half 2: Three quarters or more	2746 1486

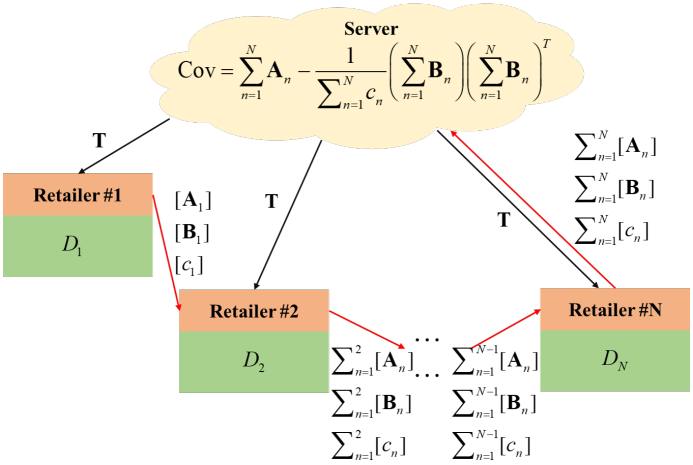

Fig. 3: Communication scheme of federated PCA.

public key is shared with all retailers for encryption, and the private key is held by the server for decryption. Then, each retailer computes and encrypts $\mathbf{A}_n$, $\mathbf{B}_n$, and c_n locally as $[\mathbf{A}_n]$, $[\mathbf{B}_n]$, and $[c_n]$, respectively. Since the server has a private key to decrypt these values, these encrypted values cannot be sent directly to the server for summation. Instead, these encrypted values are summed sequentially by passing through all the N retailers. The last retailer sends the final summations $\sum_{n=1}^{N}[\mathbf{A}_n]$, $\sum_{n=1}^{N}[\mathbf{B}_n]$, and $\sum_{n=1}^{N}[c_n]$ to the server.

On this basis, the server only decrypts these summations, calculates the covariance matrix Cov using Eq. (3), and obtains the projection matrix $\mathbf{T}$ by SVD. The black communication scheme is a distribution link starting from the server and ending at different retailers. Based on the black communication scheme, Finally, the server distributes $\mathbf{T}$ to all retailers so that the retailers can calculate the reduced matrix using Eq. (4).

The details of the privacy-perserving PCA are summarized in Algorithm 1.

Note that the federated PCA produces accurate instead of approximated results because there is no information loss in the two communication schemes with an additive homomorphic encryption approach. The summations, $\sum_{n=1}^{N}[\mathbf{A}_n]$, $\sum_{n=1}^{N}[\mathbf{B}_n]$, and $\sum_{n=1}^{N}[c_n]$ are calculated sequentially by the retailers. Since only the summations are sent across retailers, and the server does not have access to the important statistical data $\mathbf{A}_n$, $\mathbf{B}_n$ and c_n of each retailer, privacy is preserved. Actually, there are several works that have been done about distributed PCA to protect the privacy of individual participants [34]. These methods are based on one communication

TABLE II: List of the 73 extracted consumption features

Statistics	Ratios	Distribution
mean_weekdays	Load_factor_week	GE_0.5kW_proportion
mean_weekend	LF_base_wd	GE_1kW_proportion
mean_week	LF_base_wee	GE_2kW_proportion
max_week	LF_morning_wd	GT_mean_proportion
min_week	LF_morning_wee	first_location_of_max
total_week	LF_midday_wd	last_location_of_maxi
mean_base_wd	LF_midday_wee	first_location_of_min
mean_base_wee	LF_evening_wd	last_location_of_min
mean_morning_wd	LF_evening_wee	variance_week
mean_morning_wee	min/mean_wd	quantile_25
mean_midday_wd	min/mean_wee	quantile_75
mean_midday_wee	mean_evening/midday_wd	median_week
mean_evening_wd	mean_evening/midday_wee	skewness_
mean_evening_wee	mean_morning/midday_wee	kurtosis_
max_weekday	mean_morning/midday_wd	entropy_
max_weekend	mean_base/midday_wd	autocorrelation_
max_base_wd	mean_base/midday_wee	
max_base_wee	mean_weekday/week	
max_morning_wd	mean_weekend/week	
max_morning_wee	total_weekday/week	
max_midday_wd	total_weekend/week	
max_midday_wee		
max_evening_wd		
max_evening_wee		
min_weekday		
min_weekend		
min_base_wd		
min_base_wee		
min_midday_wd		
min_midday_wee		
min_morning_wd		
min_morning_wee		
total_weekday		
total_weekend		

Algorithm 1: Privacy-preserving PCA

Input: $\mathbf{D} = [\mathbf{D}_1^T, \mathbf{D}_2^T, \ldots, \mathbf{D}_N^T]^T$, $tp_A = tp_B = tp_c = 0$.
Server generates a key pair and shares the public key.
for $n = 1, \ldots N-1$
  Retailer n computes and encrypts $\mathbf{A}_n$, $\mathbf{B}_n$, and c_n as $[\mathbf{A}_n]$, $[\mathbf{B}_n]$, and $[c_n]$.
  Retailer n computes the summations $tp_A = tp_A + [\mathbf{A}_n]$, $tp_B = tp_B + [\mathbf{B}_n]$, and $tp_c = tp_c + [c_n]$, which are then sent to Retailer $n+1$;
end
Retailer N computes $tp_A = tp_A + [\mathbf{A}_N]$, $tp_B = tp_B + [\mathbf{B}_N]$, and $tp_c = tp_c + [c_N]$, and send them to Server.
Server decrypts these summations, calculates $\mathbf{T}$ and distributes it to retailers.
Each **retailer** calculated $\mathbf{D}'_n$ using Eq. (4).
Output: Reduced matrix $\mathbf{D}'_n$ for each retailer.

scheme, and some of them have no encryption systems. Thus, these methods need to make some approximations or add noises to protect privacy, which are different from the proposed federated PCA method.

D. Federated Characteristics Identification

Based on the feature extracted by PCA, an feed-forward ANN model with dense connections is trained for each characteristics. Assuming that there are a certain number of hidden layers in ANN and for each hidden layer we have

$$\boldsymbol{z}^{(l)} = W^{(l)}\boldsymbol{v}^{(l)} + \boldsymbol{b}^{(l)} \tag{5}$$

where $W^{(l)}$ and $\boldsymbol{b}^{(l)}$ denote the weight vector and bias of the l-th hidden layer; $\boldsymbol{v}^{(l)} = \phi(\boldsymbol{z}^{(l-1)})$ is the input of the l-th layer as well as the output of the $(l-1)$-th layer, and $\phi(\cdot)$ is the *tanh* activation function. *Softmax* is used in the last layer for the identification (classification) task.

If we denote $\boldsymbol{\omega}_k$ as the collection of all the weights and bias of the ANN for the k-th characteristics, the whole identification model can be represented as:

$$\hat{y}_{k,m} = f_k(\boldsymbol{\omega}_k, \mathbf{d}_m) \tag{6}$$

A smooth function categorical cross-entropy is used to guide the model training:

$$L_k = \frac{1}{M}\sum_{m=1}^{M}[y_{k,m}\log\hat{y}_{k,m} - (1-y_{k,m})\log(1-\hat{y}_{k,m})] \tag{7}$$

Thus, the global gradient can be calculated using the back-propagation algorithm

$$\nabla\boldsymbol{\omega}_k = \frac{\partial L_k(\boldsymbol{\omega}_k)}{\partial\boldsymbol{\omega}_k} \tag{8}$$

As shown in Fig. 4, in the federated framework the ANN model $f_k(\boldsymbol{\omega}_k, \mathbf{d}_m)$ is not directly trained on the whole dataset. Each retailer trains the ANN using its own datasets $\mathbf{D}_n$ and $\mathbf{Y}_n$ and obtains the n-th gradient:

$$\nabla\boldsymbol{\omega}_{n,k} = \frac{\partial L_k(\boldsymbol{\omega}_{n,k})}{\partial\boldsymbol{\omega}_{n,k}} \tag{9}$$

These local gradients are then sent to and combined in the server to calculate the global gradient [35]:

$$\nabla\boldsymbol{\omega}_k = \sum_{n=1}^{N} a_{n,k}\nabla\boldsymbol{\omega}_{n,k} \tag{10}$$

where $a_{n,k}$ denotes the weight of the gradient of the n-th retailer for the k-th characteristics.

Finally, the server updates the parameter $\boldsymbol{\omega}_k$ using Adam [36] which has an adaptive learning rate by introducing the

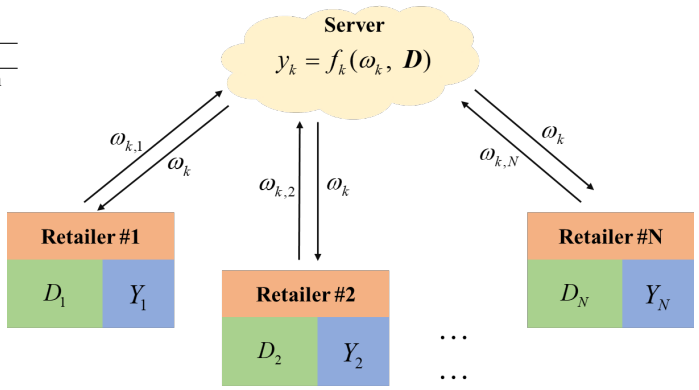

Fig. 4: Federated consumer characteristics identification in a competitive retail market

first and second moments. Specifically, at the t-th iteration, the biased first moment m_t and second moment v_t are first calculated as [36]:

$$\begin{aligned} m^t &= \beta_1^t m^{t-1} + (1-\beta_1^t)\nabla\boldsymbol{\omega}_k^t \\ v^t &= \beta_2^t v^{t-1} + (1-\beta_2^t)\nabla{\boldsymbol{\omega}_k^t}^2 \end{aligned} \tag{11}$$

where β_1^t and β_2^t denote the exponential moving average of $\nabla\boldsymbol{\omega}_k^t$ and $\nabla{\boldsymbol{\omega}_k^t}^2$, respectively. Then the unbiased first moment $\hat{m_t}$ and unbiased second moment $\hat{v_t}$ are calculated as:

$$\hat{m_t} = \frac{m_t}{1-\beta_1^t},\ \hat{v_t} = \frac{v_t}{1-\beta_2^t} \tag{12}$$

Thus, the global parameter $\boldsymbol{\omega}_k^t$ is updated as:

$$\boldsymbol{\omega}_k^{t+1} = \boldsymbol{\omega}_k^t - \frac{\lambda\hat{m_t}}{\sqrt{\hat{v_t}+\epsilon}} \tag{13}$$

where ϵ is a very small constant to avoid zero division; λ denotes the learning rate.

Three types of weights are defined to combine the local gradients in the server:

1) Normal: the weights are proportional to the sizes of the training datasets in different retailers $a_{n,k} = c_n/\sum_{n=1}^{N} c_n$. Since each consumer corresponds to one data sample, c_n, the number of consumers served by the n-th retailer, also means the number of data sample owned by the n-th retailer in the federated learning framework.
2) LA: the weights are proportional to the average losses of the training datasets in different retailers $a_{n,k} = L_{n,k}/\sum_{n=1}^{N} L_{n,k}$.
3) LS: the weights are proportional to the total losses of the training datasets in different retailers $a_{n,k} = c_n L_{n,k}/\sum_{n=1}^{N} c_n L_{n,k}$.

The full algorithm for federated ANN for the k-th characteristics identification is summarized in Algorithm 2.

To further enhance the communication security of the links between the retailers and the server, the Paillier cryptosystem is also applied to encrypt the gradients $\nabla\boldsymbol{\omega}_{n,k}$ learned by different retailers and corresponding weights $a_{n,k}$. Then, the gradients can be combined in the server. Since the gradients and corresponding weights are floating-point numbers, these numbers should also be encrypted by extended Paillier cryptosystem [32].

Algorithm 2: Federated ANN for the k-th Characteristics Identification

Input: $\mathbf{D}' = [{\mathbf{D}'_1}^T, {\mathbf{D}'_2}^T, \ldots, {\mathbf{D}'_N}^T]^T$, maximum iteration number t^{max}, random initialization parameters $\boldsymbol{\omega}_n^0$.
Server shares $\boldsymbol{\omega}_n^0$ with all retailers.;
while $t \leq t^{max}$ **do**
  for $n = 1$ *to* N **do**
    Retailer n computes local gradient $\nabla\boldsymbol{\omega}_{n,k}$ of ANN using Backpropagation with 3 iterations and records the average training loss $L_{n,k}^t$;
    $n = n + 1$;
  end
  Server computes the weight $a_{n,k}^t$ and the global gradient: $\nabla\boldsymbol{\omega}_k^t = \sum_{n=1}^{N} a_{n,k}^t \nabla\boldsymbol{\omega}_{n,k}^t$;
  Server updates the parameter $\boldsymbol{\omega}_k^{t+1}$ using Adam and shares it with all retailers;
  $t = t + 1$
end
Output: The k-th characteristics identification model $f_k(\boldsymbol{\omega}_k, \mathbf{d}_m)$.

IV. Case Studies

This section conducts comprehensive case studies on the Irish CER dataset to verify the effectiveness of the federated identification model.

A. Experimental Setups

The whole dataset is first divided into two parts: the first 80% is used to train the federated characteristics identification model; the rest 20% is used to test the performance of the model. Furthermore, the training dataset is partitioned into five parts for $N = 5$ retailers, each of which has approximately 16% consumers. Since the training dataset is not colossal, the ANN model has only one hidden layer with a dropout in the connections between the hidden layer and the output layer. The number of principal components to retain in PCA and the hyperparameters of the ANN model are tuned by cross-validation.

B. Evaluation Metrics

The predicted characteristics on the test dataset can be compared with the actual characteristics associated with the same consumers to obtain a confusion matrix for the classification problem [37]. Taking the binary classification problem as an example, the confusion matrix is:

$$\mathrm{CM} = \begin{bmatrix} \mathrm{TP} & \mathrm{FN} \\ \mathrm{FP} & \mathrm{TN} \end{bmatrix} \tag{14}$$

where TP (True positive) and TN (True negative) denote the number of positive and negative instances that are correctly classified, respectively, while FN (False negative) and FP (False positive) denote the number of positive instances and negative instances that are misclassified, respectively.

Based on the confusion matrix, the accuracy is defined in order to evaluate the quality of the prediction by computing the total number of correct predictions across all instances:

$$\mathrm{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \tag{15}$$

An accuracy value of 1 (or 100%) corresponds to a perfect classification, while the accuracy of 0 means a perfect misclassification. Accuracy shows significant limitations when working with an unbalanced dataset, as it does not discriminate between the number of correctly classified instances of different classes. This means that high accuracy can unfairly be achieved by a classifier that only predicts the majority class.

A well-known metric used in the framework of an unbalanced dataset is the Matthews Correlation Coefficient (MCC).

In parallel to the accuracy score, the MCC is also used to assess the performance of the identification model, as it is a metric better suited to overcome the bias of accuracy due to class imbalance. The MCC coefficient is defined as:

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\cdot(\mathrm{TP}+\mathrm{FN})\cdot(\mathrm{TN}+\mathrm{FP})\cdot(\mathrm{TN}+\mathrm{FN})}} \tag{16}$$

MCC takes a value in the interval $[-1, 1]$. An MCC of 1 corresponds to a perfect classification, while an MCC of -1 corresponds to a perfect misclassification. Unlike Accuracy, the proportion of each class of the confusion matrix is considered in MCC. In addition, a high score can only be achieved if the classifier has been able to correctly predict the majority of positive data instances and the majority of negative data instances while reducing FP and FN [38].

C. Basic Results

Table III gives the performance of the different identification methods for 15 consumer socio-demographic characteristics in terms of Accuracy and MCC, where "Centralized" and "C-PCA" denote the centralized ANN models without and with PCA-based feature extraction, respectively. Comparing the Centralized method and C-PCA method, it can be observed that, overall, the performance is improved through feature extraction with the averaged Accuracy and MCC score of 68.53%, and 0.2989, respectively. There is an increase in Accuracy up to 1.69% and up to 12.5% for the MCC. This means that PCA-based feature extraction is an effective means of improving the performance of these identification models.

The federated learning models with differently weighted combination strategies (LA, LS, and Normal) exhibit similar performance due to the fact that the datasets of all retailers serve a similar number of consumers who are i.i.d. It is also interesting to observe that even though the federated learning model is trained locally by each retailer and combined in the server, the performance is not always worse than the centralized approach. This is because that the neural network training problem is nonconvex and thus, training the model in a distributed manner with their local datasets may result in a higher accuracy if better suboptimal solutions or the global optimal solution can be identified.

In general, the results show that household characteristics can be revealed from electricity consumption data by the identification models with fair Accuracy and MCC scores. More specifically, the characteristics identification model achieves an accuracy of approximately 84% for "#4 Live alone", and an MCC of 0.48 for predicting "#3 Number of residents". In the identification of "#1 Employment status" and "#2 Retirement status", "#6 Number of bedrooms", "#10 Type of home" and "#12 Number of entertainment appliances", as well as of "#13 cooking type", the proposed model has been able to predict the characteristic with an MCC of about 0.25 to 0.45. Conversely, the most difficult characteristics identified by our model are "#14 Age of house" and "#15 Rent or owned", with MCC scores of only about 0.18 and 0.11, respectively. A low identification performance may reflect a low correlation between the smart meter data and the age of the house.

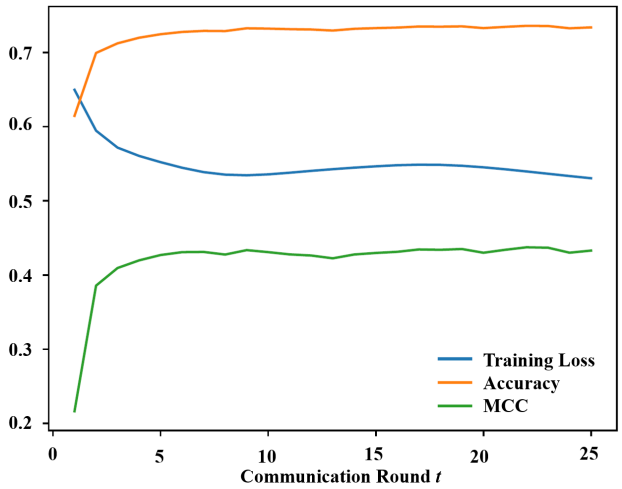

Fig. 5: Convergence of the federated model for "#1 Employment status".

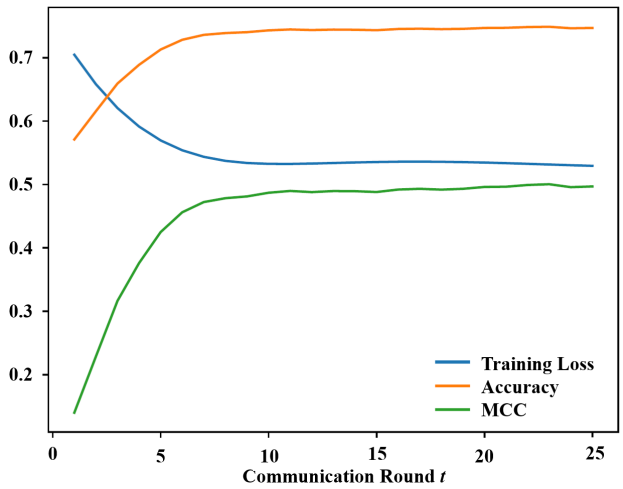

Fig. 6: Convergence of the federated model for "#5 Number of residents".

It is, however, important to acknowledge the imbalance of the Irish CER dataset used in this study. The dataset contains a disproportionate ratio of observations in each class for most characteristics, which may lead to biased results in terms of Accuracy. In this sense, more attention should be paid to the MCC score when interpreting the results.

Fig. 5 and Fig. 6 present the changes in training loss over all training datasets, the Accuracy and MCC of the federated model on the test dataset in different communication rounds to identify the "#1 Employment status" and "#5 Number of residents". Note that the convergences curves are the averages of five independent simulations. Both figures show fast convergences of the federated model. The federated model is combined according to the total losses of the training datasets in different retailers, i.e., the LS method. Similar trends and fast convergences can also be observed in the LA and Normal methods for different characteristics identification.

Table IV gives the performance of three commonly used classification models, i.e., K-Nearest Neighbors (KNN), Ran-

TABLE III: Performance of different identification models for 15 consumer characteristics.

Characteristics	Accuracy					MCC				
	Centralized	C-PCA	LA	LS	Normal	Centralized	C-PCA	LA	LS	Normal
#1 Employment status	0.7190	0.7332	0.7367	0.7432	0.7420	0.4099	0.4426	0.4480	0.4609	0.4586
#2 Retirement status	0.7591	0.7698	0.7577	0.7594	0.7651	0.4265	0.4433	0.4057	0.4065	0.4255
#3 Have children	0.7686	0.7769	0.7727	0.7739	0.7739	0.4144	0.4205	0.4084	0.4105	0.4104
#4 Live alone	0.8442	0.8288	0.8412	0.8453	0.8465	0.4415	0.4820	0.4801	0.4703	0.4894
#5 Number of residents	0.7485	0.7580	0.7503	0.7500	0.7527	0.4959	0.5155	0.4998	0.4994	0.5044
#6 Number of bedroom	0.5302	0.5882	0.5911	0.5920	0.5941	0.1961	0.2551	0.2682	0.2655	0.2679
#7 Age	0.6770	0.6805	0.6728	0.6728	0.6743	0.2804	0.2082	0.2261	0.2351	0.2381
#8 Number home appliances	0.6423	0.6572	0.6610	0.6665	0.6706	0.3218	0.3442	0.3512	0.3597	0.3676
#9 Light bulb proportion	0.6021	0.5762	0.6308	0.6322	0.6284	0.0080	0.0328	0.0134	0.0261	0.0161
#10 Type of home	0.6017	0.6173	0.6191	0.6191	0.6176	0.2007	0.2291	0.2336	0.2335	0.2303
#11 Social class	0.5583	0.5656	0.5644	0.5660	0.5647	0.2454	0.2638	0.2617	0.2624	0.2606
#12 Number of entertainment appliances	0.5396	0.5714	0.5517	0.5466	0.5487	0.2273	0.2914	0.2636	0.2517	0.2575
#13 Cooking	0.7155	0.7332	0.7267	0.7264	0.7311	0.2029	0.2790	0.2740	0.2700	0.2829
#14 Age of house	0.4782	0.4947	0.5068	0.5162	0.5121	0.1280	0.1603	0.1791	0.1951	0.1878
#15 Rent or owned	0.9265	0.9289	0.9295	0.9301	0.9298	-0.0135	0.1152	0.1029	0.1245	0.1137
Average	**0.6740**	**0.6853**	**0.6875**	**0.6893**	**0.6901**	**0.2657**	**0.2989**	**0.2944**	**0.2981**	**0.3007**

dom Forest (RF) [39], and XGBoost [40] for consumer characteristic identification. It can be seen that among these three methods, XGBoost has the best performance in terms of average Accuracy and MCC. However, the applied ANN model with PCA has better performance than XGBoost, especially for MCC.

D. Tests on Unbalanced Datasets

The above case studies are conducted on evenly distributed datasets, i.e., the number of consumers and distributions of characteristics are similar among different retailers. However, the consumers served by different retailers are probably unbalanced. Thus, two unbalanced cases are studied here.

The first one is the number-unbalanced case, where five retailers serve 5%, 10%, 10%, 15%, and 40% consumers, respectively. Table V presents the performance of the federated methods with different combination strategies. It can be observed that even though the LA method has not taken the number of consumers into account, it can achieve comparable performance with the LS and Normal methods. The main reason is that the LA method has a regularization effect over local updates that are too aggressive. In a round of federated training, local updates of retailers with a higher amount of data may diverge from the initial model weights. Averaging based on the amount of data gives higher weights to these updates that diverge and thus may result in an aggregated update that is far from the initial weights of the model. In contrast, averaging based on the average loss gives high weights to local updates from the clients with worse performance, but not to local updates from the clients with a higher amount of data. This approach might be too aggressive and thus introduces an additional regularization effect.

The second one is the characteristics-unbalanced case where the characteristics of the consumers served by the same retailer are highly unbalanced. Table VI gives the performance of the federated methods with different combination strategies. It can be seen that the LA and LS methods perform slightly worse but comparable to the centralized and balanced cases. While the Normal method, as a commonly used combination method, shows much worse performance in this situation. The MCC scores for different characteristics are close to zero, which means that the trained models only guess the characteristics at random. This characteristics-unbalanced case highlights the importance of considering training losses in the server combination process.

Based on the above experiments, several meaningful observations can be summarized as follows:

1) PCA-based feature extraction helps to enhance the performance of the ANN model;
2) The ANN model has comparable performance with XGBoost in terms of Accuracy but shows large improvements in terms of MCC;
3) The federated ANN model have similar performance on the balanced dataset or number-unbalanced dataset with three different weighted averaging strategies.
4) LA and LS have better performance than Normal on the characteristics-unbalanced dataset.

V. Conclusions and Future Works

This paper presents a federated learning approach for electricity consumer characteristics identification, which consists of privacy-persevering PCA and federated ANN model. The case studies show that by extracting features using PCA, the identification models gain better performance. The federated identification models with LA, LS, and Normal combination strategies have comparable performance to the centralized method. However, when the characteristics of the consumers are highly unbalanced among different retailers, the Normal method performs much worse. The LA and LS methods still have comparable performance since they take into account the training losses of different retailers in the server combination process.

Extra benefits (such as the benefits from a better understanding of the consumers and implementation of demand response) can be gained by applying the federated learning approach. How to share the benefits among retailers who contribute their own data is an interesting topic and will be studied in the future. In addition, clustering is an effective approach for electricity consumer behavior analysis. Thus, another of our future works is to develop a federated clustering algorithm on smart meter data.

TABLE IV: Comparisons with other machine learning models.

Characteristics	Accuracy			MCC		
	KNN	RF	XGBoost	KNN	RF	XGBoost
#1 Employment status	0.6217	0.6929	0.7416	0.1762	0.3430	0.4469
#2 Retirement status	0.6517	0.7191	0.7116	0.4287	0.4446	0.4146
#3 Have children	0.6667	0.6891	0.7191	0.0923	0.0850	0.2208
#4 Live alone	0.7940	0.7903	0.8015	0.2656	0.2228	0.2607
#5 Number of residents	0.6667	0.6816	0.7116	0.3386	0.3689	0.4284
#6 Number of bedroom	0.5281	0.6330	0.6367	0.2049	0.3514	0.3486
#7 Age	0.6330	0.6404	0.6292	0.1628	0.1133	0.1042
#8 Number home appliances	0.5843	0.5730	0.6142	0.2311	0.1873	0.2680
#9 Light bulb proportion	0.5281	0.6217	0.6554	-0.1199	-0.0027	0.1178
#10 Type of home	0.5056	0.5918	0.6918	0.0229	0.1939	0.1982
#11 Social class	0.4045	0.4757	0.5131	0.0462	0.1414	0.2034
#12 Number of entertainment appliances	0.5581	0.5893	0.5955	0.2630	0.2878	0.3350
#13 Cooking	0.6667	0.7341	0.7491	0.0216	0.2412	0.2900
#14 Age of house	0.3570	0.4644	0.4681	0.0004	0.1173	0.1205
#15 Rent or owned	0.9272	0.9287	0.9299	-0.0008	0.1005	0.1237
Average	**0.6069**	**0.6549**	**0.6712**	**0.1201**	**0.1984**	**0.2418**

TABLE V: Performances of number-unbalanced case.

Charact-eristics	Accuracy			MCC		
	LA	LS	Normal	LA	LS	Normal
#1	0.7426	0.7414	0.7296	0.4609	0.4578	0.434
#2	0.7662	0.771	0.7591	0.4287	0.4446	0.4146
#3	0.7721	0.771	0.7698	0.402	0.4041	0.3984
#4	0.8536	0.8512	0.8383	0.5093	0.5169	0.4419
#5	0.7556	0.7591	0.7568	0.5103	0.5176	0.5125
#6	0.5929	0.5953	0.5905	0.2651	0.2683	0.2634
#7	0.6853	0.6865	0.6793	0.2443	0.2534	0.2334
#8	0.6632	0.6692	0.6736	0.3523	0.366	0.3738
#9	0.6293	0.6399	0.6281	0.0281	0.0636	0.0079
#10	0.6197	0.6366	0.627	0.2351	0.2693	0.2508
#11	0.5632	0.573	0.5558	0.2557	0.2736	0.244
#12	0.5419	0.5514	0.5478	0.2457	0.2585	0.2547
#13	0.7178	0.7285	0.7131	0.2471	0.2734	0.2659
#14	0.5077	0.497	0.51	0.1781	0.1613	0.1836
#15	0.9289	0.9301	0.9301	0.0813	0.1245	0.1245
Average	**0.6893**	**0.6934**	**0.6873**	**0.2963**	**0.3102**	**0.2936**

TABLE VI: Performance of characteristics-unbalanced case

Charact-eristics	Accuracy			MCC		
	LA	LS	Normal	LA	LS	Normal
#1	0.7438	0.7349	0.5998	0.4804	0.4352	0
#2	0.7642	0.7606	0.6963	0.4504	0.4165	-0.0057
#3	0.7659	0.7772	0.7172	0.4276	0.4354	0.1238
#4	0.8247	0.8388	0.8084	0.5283	0.5058	0.0093
#5	0.7541	0.745	0.7364	0.5076	0.4935	0.4902
#6	0.5908	0.5876	0.5607	0.2657	0.2541	0.2234
#7	0.6746	0.6823	0.6683	0.2175	0.2402	-0.0214
#8	0.671	0.6785	0.6226	0.3717	0.3832	0.3204
#9	0.6009	0.6092	0.6494	0.0198	0.0416	0
#10	0.6137	0.6035	0.5502	0.2257	0.2051	0.1152
#11	0.5598	0.5586	0.4595	0.2533	0.2439	0.0626
#12	0.546	0.5204	0.5133	0.2697	0.2144	0.1413
#13	0.7338	0.74	0.696	0.317	0.3129	-0.0355
#14	0.4782	0.5047	0.3943	0.1481	0.1844	0
#15	0.8789	0.9165	0.9289	0.187	0.2081	0
Average	**0.68**	**0.6838**	**0.6401**	**0.3113**	**0.305**	**0.0949**

REFERENCES

[1] X. Liu and P. S. Nielsen, "A hybrid ICT-solution for smart meter data analytics," *Energy*, vol. 115, pp. 1710–1722, 2016.

[2] G. Giaconi, D. Gündüz, and H. V. Poor, "Smart meter privacy with renewable energy and an energy storage device," *IEEE Trans. Information Forensics and Security*, vol. 13, no. 1, pp. 129–142, 2017.

[3] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2018.

[4] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, "Clustering analysis of residential electricity demand profiles," *Applied Energy*, vol. 135, pp. 461–471, 2014.

[5] Y. Han, X. Sha, E. Grover-Silva, and P. Michiardi, "On the impact of socio-economic factors on power load forecasting," in *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, 2014, pp. 742–747.

[6] Y. Wang, Q. Chen, D. Gan, J. Yang, D. S. Kirschen, and C. Kang, "Deep learning-based socio-demographic information identification from smart meter data," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 2593–2602, 2018.

[7] M. Hayn, V. Bertsch, and W. Fichtner, "Electricity load profiles in europe: The importance of household segmentation," *Energy Research & Social Science*, vol. 3, pp. 30–45, 2014.

[8] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Applied Energy*, vol. 141, pp. 190–199, 2015.

[9] X. Tong, R. Li, F. Li, and C. Kang, "Cross-domain feature selection and coding for household energy behavior," *Energy*, vol. 107, pp. 9–16, 2016.

[10] L. Sun, K. Zhou, and S. Yang, "An ensemble clustering based framework for household load profiling and driven factors identification," *Sustainable Cities and Society*, vol. 53, p. 101958, 2020.

[11] L. Jin, D. Lee, A. Sim, S. Borgeson, K. Wu, C. A. Spurlock, and A. Todd, "Comparison of clustering techniques for residential energy behavior using smart meter data," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), Tech. Rep., 2017.

[12] J. P. Gouveia and J. Seixas, "Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys," *Energy and Buildings*, vol. 116, pp. 666–676, 2016.

[13] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397–410, 2014.

[14] K. Hopf, M. Sodenkamp, I. Kozlovkiy, and T. Staake, "Feature extraction and filtering for household classification based on smart electricity meter data," *Computer Science-Research and Development*, vol. 31, no. 3, pp. 141–148, 2016.

[15] P. Voigt and A. Von dem Bussche, "The EU general data protection regulation (GDPR)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[16] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[17] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[18] L. Li, Y. Fan, M. Tse, and K.-Y. Lin, "A review of applications in federated learning," *Computers & Industrial Engineering*, vol. 149, p. 106854, 2020.

[19] S. A. Rahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi, and M. Guizani, "A survey on federated learning: The journey from

centralized to distributed on-site learning and beyond," *IEEE Internet of Things Journal*, pp. 1–1, 2020.

[20] F. Sattler, S. Wiedemann, K. R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-i.i.d. data," *IEEE Trans. Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3400–3413, 2020.

[21] Y. Ye, S. Li, F. Liu, Y. Tang, and W. Hu, "Edgefed: Optimized federated learning based on edge computing," *IEEE Access*, vol. 8, pp. 209 191–209 198, 2020.

[22] A. Imteaj and M. H. Amini, "Distributed Sensing Using Smart End-User Devices: Pathway to Federated Learning for Autonomous IoT," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, pp. 1156–1161.

[23] A. Imteaj and M. H. Amini, "Fedar: Activity and resource-aware federated learning model for distributed mobile robots," *arXiv preprint arXiv:2101.03705*, 2021.

[24] L. U. Khan, M. Alsenwi, Z. Han, and C. S. Hong, "Self organizing federated learning over wireless networks: A socially aware clustering approach," in *2020 International Conference on Information Networking (ICOIN)*, 2020, pp. 453–458.

[25] Y. M. Saputra, D. T. Hoang, D. N. Nguyen, E. Dutkiewicz, M. D. Mueck, and S. Srikanteswara, "Energy demand prediction with federated learning for electric vehicle networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[26] S. Lee and D. H. Choi, "Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources," *IEEE Trans. Industrial Informatics*, pp. 1–1, 2020.

[27] "Irish Social Science Data Archive.(2012). Commission for Energy Regulation (CER) Smart Metering Project." *[Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/*.

[28] K. Hopf, M. Sodenkamp, I. Kozlovkiy, and T. Staake, "Feature extraction and filtering for household classification based on smart electricity meter data," *Computer Science-Research and Development*, vol. 31, no. 3, pp. 141–148, 2016.

[29] K. Gajowniczek, T. Zabkowski, and M. Sodenkamp, "Revealing household characteristics from electricity meter data with grade analysis and machine learning algorithms," *Applied Sciences*, vol. 8, no. 9, p. 1654, 2018.

[30] M. Al-Rubaie, P.-y. Wu, J. M. Chang, and S.-Y. Kung, "Privacy-preserving pca on horizontally-partitioned data," in *2017 IEEE Conference on Dependable and Secure Computing*. IEEE, 2017, pp. 280–287.

[31] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 1999, pp. 223–238.

[32] CSIRO's Data61, "Python paillier library," https://github.com/data61/python-paillier, 2013.

[33] S. Mazza, D. Patel, and I. Viola, "Homomorphic-encrypted volume rendering," *arXiv preprint arXiv:2009.02122*, 2020.

[34] S. X. Wu, H. Wai, L. Li, and A. Scaglione, "A review of distributed algorithms for principal component analysis," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1321–1340, 2018.

[35] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *arXiv preprint arXiv:1602.05629*, 2016.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019.

[38] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, December 2020.

[39] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[40] 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*, 2016, pp. 785–794.