Data-Driven Outage Restoration Time Prediction via Transfer Learning With Cluster Ensembles

Dingwei Wang, *Graduate Student Member, IEEE*, Yuxuan Yuan, *Graduate Student Member, IEEE*, Rui Cheng, *Graduate Student Member, IEEE*, and Zhaoyu Wang, *Senior Member, IEEE*

***Abstract*—This article develops a data-driven approach to accurately predict the restoration time of outages under different scales and factors. To achieve the goal, the proposed method consists of three stages. First, given the unprecedented amount of data collected by utilities, a SDESC method is proposed to decompose historical outage datasets, which enjoys good computational efficiency and scalability. Specifically, each outage sample is represented by a linear combination of a small number of selected dictionary samples using a density-based method. Then, the dictionary-based representation is utilized to perform the spectral analysis to group the data samples with similar features into the same subsets. In the second stage, a knowledge-transfer-added restoration time prediction model is trained for each subset by combining weather information and outage-related features. The transfer learning technology is introduced to deal with the underestimation problem caused by data imbalance in different subsets, thus improving the model performance. Furthermore, to connect unseen outages with the learned outage subsets, a t-distributed stochastic neighbor embedding-based strategy is applied. The proposed method fully builds on and is also tested on a large real-world outage dataset from a utility provider with a time span of six consecutive years. The numerical results validate that our method has high prediction accuracy while showing good stability against real-world data limitations.**

***Index Terms*—Distribution network, outage restoration time prediction, sparse dictionary-based ensemble spectral clustering, transfer learning.**

NOMENCLATURE

ANN	Artificial neural networks.
CI	Customers interrupted.
CNN	Convolutional neural networks.
DBI	Davies-Bouldin validation index.
GBM	Gradient boosting machine.
LASSO	Least absolute shrinkage and selection operator.
LM	Levenberg-Marquardt.
NOAA	National Oceanic and Atmospheric Administration.
RF	Random forest.
RT	Restoration time.
SC	Spectral clustering.
SDESC	Sparse dictionary-based ensemble spectral clustering.
SVR	Support vector regressions.
t-SNE	t-distributed stochastic neighbor embedding.
$\mathbb{A}_i$	Subset of data representation set $\mathbb{V}$
$\overline{\mathbb{A}_i}$	complement of $\mathbb{A}_i$
$b_j^{(0)}$	Bias of jth hidden neuron.
$b_l^{(1)}$	Bias of lth putput neuron.
$\mathbf{C}$	Cluster subset.
$c_o^{t_i}$	Cumulative total outages at time t_i
$c_r^{t_i}$	Cumulative total restorations at time t_i
$c_{outages}^{t_i}$	Number of simultaneous outages at time t_i
$cov(\cdot)$	Covariance.
$\mathbf{D}$	Dictionary matrix.
$\mathbf{D}_{\langle i \rangle}$	Sub-matrix of $\mathbf{D}$
$\mathbb{D}_{\langle i \rangle}$	Set consisting of all column vectors in $\mathbf{D}_{\langle i \rangle}$
$\mathbf{d}_j$	j^{th} Column vector of the dictionary matrix.
$d(\mathbb{A}_i)$	Sum of the weights of vertices in $\mathbf{A}_i$
$\mathbb{E}$	Set of Edge of the undirected graph.
$\mathbf{f}$	Input feature vector in correlation criteria.
$\mathbb{G}$	Undirected similarity graph.
h	Bandwidths of the kernel function.
$\mathbf{I}$	Diagonal matrix corresponding to the adjacency matrix.
$K_h(\cdot)$	Kernel function with bandwidths h
k	Optimal number of clusters.
$\mathbf{L}$	Graph Laplacian matrix.
M	Uncertainty of random variable.
m	Total number of the outage events.
$\mathbf{N}$	New matrix contains reconstructed data points.
n	Number of features.
$\mathbf{O}$	Historical outage dataset.
$\mathbf{o}_i$	Outage related information in the ith outage event.
o_{ij}	The ith row and j^{th} column entry in the outage dataset.
$P(\cdot)$	Probability.
$P(\cdot\|\cdot)$	Conditional probability.
$P(\cdot,\cdot)$	Joint probability.
p	Number of dictionaries.
$\mathbf{R}$	Representation matrix.
$\hat{r}$	Number of nearest dictionaries of $\mathbf{o}_i$
$\mathbf{r}_i$	i^{th} column vector of the representation matrix.

Manuscript received 22 January 2022; revised 19 June 2022 and 16 November 2022; accepted 9 January 2023. Date of publication 12 January 2023; date of current version 26 December 2023. This work was supported by the National Science Foundation under Grants EPCN 2042314 and 1929975. Paper no. TPWRS-00124-2022. *(Corresponding author: Zhaoyu Wang.)*

The authors are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: dingwei@iastate.edu; yuanyx@iastate.edu; ruicheng@iastate.edu; wzy@iastate.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPWRS.2023.3236513.

Digital Object Identifier 10.1109/TPWRS.2023.3236513

r_{ji}	The jth row and i^{th} column entry of the representation matrix.
S	Objective function of graph partitioning.
$\tilde{s}(\cdot,\cdot)$	Sum of the weights between vertices.
U	Decrease in uncertainty.
$\mathbf{u}$	Output vector in correlation criteria.
$\mathbb{V}$	Set of data point representation.
$var(\cdot)$	Variance.
$\mathbf{v}_i$	Vertex in $\mathbb{V}$
$\mathbf{W}$	Adjacency matrix.
w_{ij}	Weight between two selected vertices.
$\tilde{X}, \tilde{Y}$	Random feature variables.
$\tilde{x}, \tilde{y}$	Realizations of $\tilde{X}, \tilde{Y}$
$\mathbf{z}_i$	Simulated data in low-dimensional.
α	Local scaling parameter.
Γ	Pearson correlation coefficient.
γ	Minimum number of neighbors.
ϵ	Maximum value of the Frobenius norm of the representation matrix $\mathbf{R}$
θ	Number of vertices in graphs.
κ	Number of neighbor points.
λ	Vector Variance of the Gaussian function.
ξ	Radius of the minimum number of neighbors.
ϕ	Kullback divergence.

I. Introduction

RECENT Texas blackout has demonstrated the danger and inconveniences people face during a severe power outage event. In general, power outages have significant impacts on production, transportation, communication, and health supply service, resulting in significant economic losses. When an outage occurs, utilities need to make a series of decisions quickly, including detecting and locating the fault, estimating costs and the number of customers affected, predicting outage restoration time, planning repair strategies, and dispatching crews [1]. From the customer's perspective, the most important and concerned information is timely and accurate outage recovery time prediction, which will greatly help them plan for subsequent arrangements in advance. However, it is challenging to estimate outage recovery time as power outages are typically unplanned with limited information. Moreover, the causes of power outages involve a wide variety of factors, e.g., bad weather, human behaviors, and equipment failures [2], [3], [4], [5]. To improve customer satisfaction, accurate prediction of outage recovery time is becoming a top priority for utilities.

In recent years, research studies have focused on inferring the quantity and duration of outages by using various approaches and data sources, which can be broadly classified into two groups based on prediction targets: *Group I* - Prediction of outage *duration*. The authors in [6] analyzed the restoration duration of outages using statistical and quantitative methods with several factors including the time of outages, consequences, and environmental conditions. In [7], an accelerated failure time model using severe weather records was developed to estimate the duration of outages. In [8], the authors summarized six years' historical outage data and proposed a deep neural network to predict repair and restoration time with respect to severe weather events. In [9], the authors utilized radar observation data and further proposed a generalized weather-dependent failure rate model, based on the Bayesian prediction algorithm, to provide a prediction of outage duration. *Group II* - Prediction of outage *numbers*. In [10], the authors performed a Poisson regression model to predict an average number of outages over a period under normal weather conditions. The authors in [11] estimated distribution system outage numbers caused by wind and lightning using an artificial neural network. In [12], a graph neural network is proposed to predict power outage numbers by utilizing weather variables. A summary of the literature is shown in Table I.

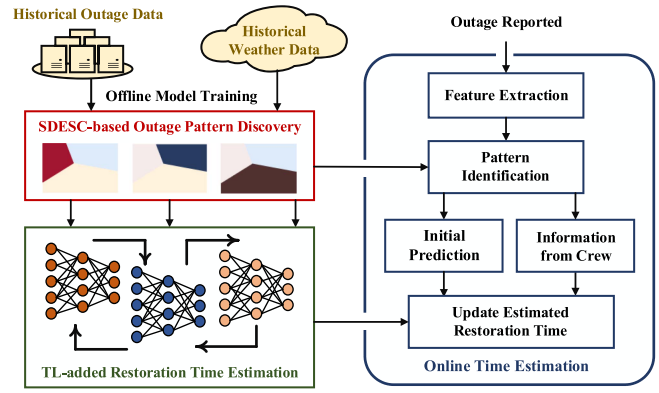


Fig. 1. The flowchart of the proposed method.

Even though the previous works demonstrate valuable results, some research challenges remain outstanding in this area. First, most studies in group I are generally based on the assumption that each outage recovery can be treated as an isolated process. In other words, with respect to multiple outages that occur in the same systems, the existing methods estimate the restoration time separately without consideration of the correlation among multiple coinciding outages. However, in actual grids, while maintenance crews are repairing an outage, the repair of another outage occurring in the neighboring area may be delayed due to crew shortages. Thus, such an assumption in group I reduces the accuracy of these prediction models. Second, some studies in group II trained a global model for the whole historical outage dataset, which ignores the uncertainty caused by the heterogeneity of outage events under different scales and factors, thus reducing prediction performance for unseen outages. Also, given that real outage reports are scarce, most existing works rely on weather information to develop their models and assume the weather conditions across each small area are homogeneous. Nevertheless, while severe weather is one of the leading causes of power outages, other factors, such as equipment failures and human factors, should not be ignored.

To this end, we propose a novel data-driven method to predict outage restoration time that incorporates a combination of cluster ensembles and transfer learning techniques. The flowchart

TABLE I
LITERATURE REVIEW ON OUTAGE PREDICTION IN DISTRIBUTION SYSTEMS

Reference	Approach	Data source	Case study	Cons
[6]	Poisson regression model	Time and weather data	Prediction of historical outages	Without developing a prediction model for future data samples
[7]	Accelerated failure time model	Severe weather records	Estimate duration of historical outages	Data distribution assumption, uses only weather data as variables, limited data source
[10]	Poisson regression model	Normal weather records	Predict average number of outages over a period under normal weather conditions	
[9]	Bayesian prediction algorithm	Radar observations data	Provide an estimation of outage numbers	
[8]	Deep neural network	Historical outage data with severe weather records	Predict repair and restoration time with respect to severe weather events	Single global model, each outage recovery is treated as an isolated process
[11]	Artificial neural network	Wind and lighting records	Estimate distribution system outages numbers caused by wind and lighting	
[12]	Graph neural network	Historical normal and severe weather records	Predict upcoming power outages numbers	

of our proposed method is depicted in Fig. 1. Compared with existing studies, the contributions of our work are summarized as follows: 1) To investigate the interaction of simultaneous outage events during a period, we extract the statistics from real-world outage reports and calculate the cumulative number of coinciding outages and affected customers. The temporal information in the outage dataset, such as the outage start time, end time, and restoration time, is utilized to summarize the real-time numbers of outages and customers affected in a time span. This feature can be explored to approximate the utility's stress for repairing the outage. 2) Unlike previous methods that train a global predictor, the proposed method estimates the restoration time in a cluster-wise manner to deal with the uncertainty caused by the heterogeneity of outage events. Specifically, a SDESC method is developed to efficiently group the historical outage events. A prediction model for each data subset is trained to construct an end-to-end mapping between the outage-related information and the restoration time by leveraging machine learning techniques. 3) According to our investigation of real-world outage datasets, there may be significant gaps between the amount of data accessible for various patterns and scales of outages. Therefore, a transfer learning strategy is integrated with the proposed prediction framework to deal with the class imbalance problem caused by the data scarcity of specific outage patterns. 4) The proposed approach leverages not only high-precision weather information but also exploits outage-related features collected by our utility partner, which are the time and location of outages, number of customers interrupted, cause code, and duration of repair/restoration process. The proposed prediction methodology has been tested and verified using real outage data.

The rest of this paper is organized as follows: Section II introduces the problem definition and describes the available outage dataset in detail. Section III proposes the sparse dictionary-based ensemble spectral clustering method. Section IV presents the transfer learning-added outage restoration time prediction model. The numerical results are analyzed in Section V. Section VI concludes the paper.

II. PROBLEM DEFINITION & OUTAGE DATA DESCRIPTION

A. Problem Definition

When an outage occurs, one of the most common questions that customers may ask is: How long will it take to restore the power [6]? If utilities can answer this question at an early stage, customers can plan ahead and accordingly to avoid inconvenience caused by power outages. In practice, outage restoration time is defined as the time from the start of the outage to the service fully recovered to the customers [8]. In actual grids, one common solution is to formulate a mathematical relationship between the restoration time and the number of interrupted customers [6], [7], [9]. Such a solution embodies a set of assumptions, such as the pre-define statistical models with fixed parameters (e.g., Poisson distribution). However, the distribution of outage restoration time may not follow the pre-defined pattern; meanwhile, some variables and features cannot be considered in the statistical models due to the curse of dimensionality, which reduces the accuracy of the prediction. Hence, machine learning-based methods are receiving increasing attention due to the unprecedented amount of data collected by utilities. Basically, a learning-based solution is based solely on using historical outage data without an assumption of data distributions to develop a supervised prediction model.

B. Available Outage Dataset

The outage data under study includes over 16 000 records over a six-year period in New York State. Fig. 2 describes the structure of the available outage dataset. The original information on each outage record includes: the start and end time of the outage accurate to seconds, the number of customers interrupted, repair and restoration time accurate to seconds, cause code, location, and distribution network circuit number. Specifically, the start and end times of the outage events are reported by fuse cards that record the time at which an outage starts and ends according to the loss of power. The cause information is summarized into two types: *Cause key* - There are 63 causes, of which 5% of the

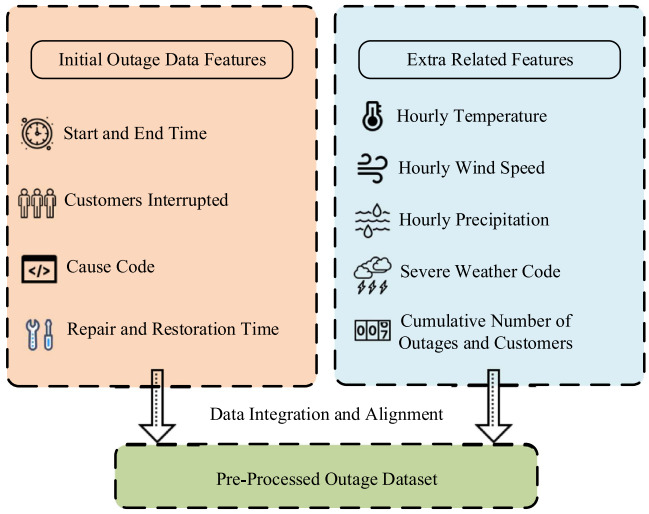


Fig. 2. The description of the outage dataset.

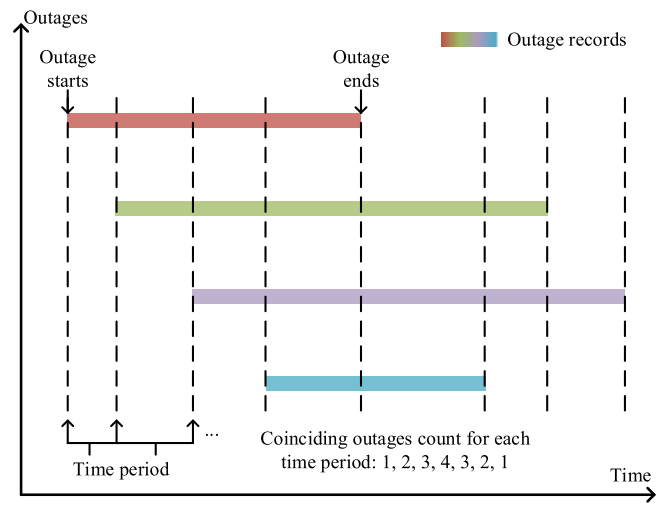


Fig. 3. Exemplary of the cumulative number of coinciding outages.

records are weather-related, another 14% are animal-related, and the rest are mainly due to component malfunctions, tree limbs, and debris. The top two causes of all outage events are tree limbs near the clearance zone of equipment and squirrels. The top weather-related causes are precipitation, wind, and lightning strike. *Equipment cause key* - The outage caused by equipment failure. The top equipment-related causes are system failure, conductor disconnection, and transformer malfunction. These two types of cause information are recorded in each outage and represented by a digital code.

In this work, data preprocessing includes two steps: 1) *Missing and bad data cleaning* - Given equipment failures or human mistakes, missing and bad data in historical outage datasets are typically unavoidable and can lead to misclassification, which negatively affects the performance of data-driven methods. Hence, the available dataset is initially processed to clean missing and bad outage samples. Data samples with empty entries are removed first. Then, following the engineering intuition, data samples with logically incorrect entries (e.g., the restoration time is greater than the total outage time) and grossly erroneous values (e.g., the restoration time is extremely illogically high) are removed. 2) *Outage-related features investigation* - Leveraging cross-domain insights from public weather data and the geographic information of systems, the raw dataset is explored by adding hourly temperature, precipitation, and wind speed. These data are collected from the National Oceanic and Atmospheric Administration (NOAA) [13], [14]. The hourly data from the NOAA is aligned with each outage sample based on the start time of each outage. Given that severe weather events play a crucial role in outage restoration time prediction, the weather condition of each outage record is also marked as a discrete code, consisting of normal conditions, snow storm, lightning strikes, high-speed wind, and flood [1]. Other features we have added are *cumulative number of coinciding outages and the number of customers affected*. Specifically, in this work, the cumulative number of coinciding outages is the quantity of outages presented at a certain time period that has not yet been resolved. This number varies over time while the outages occur and are restored. When this number remains near zero, it indicates that the system was in normal condition prior to outages, or that outages happened infrequently and were resolved quickly. Conversely, this number can be relatively high under certain stressed conditions, indicating that numbers of outages have stacked and affected the restoration time. For example, on December 10th, 2014, the cumulative number of coinciding outage events was over 100, and the average outage restoration time was above 300 minutes. In contrast, the cumulative number of coinciding outage events is relatively low on April 15th, 2015, the average outage restoration time is below 60 minutes. Note that in this work, the number of outages is actually the number of failures requiring repair. It is determined by the records of the outage management system. For example, if two feeders connected to a medium-voltage substation lose power, it can be counted as two faults or more, depending on the number of damaged devices.

Fig. 3 demonstrates an example for calculating the cumulative number of coinciding outages. Each dashed line on the graph is a timestamp that is recorded at the start and end of each outage. By comparing each two adjacent dashed lines, a specific time period can be defined to explore the cumulative number of coinciding outages by counting the number of outages. Following this process, the cumulative number of coinciding outages for each period is calculated by sorting the timestamps by their orders of occurrence:

$$c_{outages}^{t_i} = c_o^{t_i} - c_r^{t_i} \tag{1}$$

where $c_o^{t_i}$ is the cumulative total outages at time t_i, $c_r^{t_i}$ is the cumulative total restorations at time t_i, and $c_{outages}^{t_i}$ is the number of simultaneous outages at time t_i. The cumulative number of coinciding outages is used to define the stress of the system and can provide additional dimension information for outage grouping, as well as enhance the variability of the metric. Similar to the above definition and (1), the cumulative

number of customers interrupted can be calculated by replacing the number of coinciding outages with the number of customers.

III. Historical Outage Data Discovery Using SDESC

Currently, utilities constantly attempt to collect as much information as possible on power outages. However, the vast majority of outages in distribution systems are small-scale and medium-scale; in contrast, large-scale outages are still rare, thus leading to a data imbalance problem[1] [15]. In this case, it could result in overfitting problems when the raw outage data is used to train a global prediction model.

To address the challenge posed by the real-world imbalanced outage dataset, a novel unsupervised method, known as SDESC, is leveraged to distinguish the hidden outage features and partition the historical dataset into distinct subsets. The proposed method follows the line of unsupervised research utilizing a spectral analysis to discover the latent features. Furthermore, the sparse coding technique is adapted to decrease the complexity of outage event-based adjacency matrix construction and eigen-decomposition, thus greatly reducing the cost of practical implementation [16].

A. Feature Selection

The purpose of feature selection is to discover features that may have a large impact on the recovery time and to remove features that may have duplicate information about each other [17]. In this work, we applied two simple but effective methods to achieve this goal. Specifically, to find valuable features, the correlation criteria method is a bivariate analysis that measures the strength of association between outage-related features and outage restoration time. For this approach, the Pearson correlation coefficient is utilized which is defined as:

$$\Gamma_{\mathbf{f},\mathbf{u}} = \frac{cov(\mathbf{f},\mathbf{u})}{\sqrt{var(\mathbf{f}) * var(\mathbf{u})}} \tag{2}$$

where $\mathbf{f}$ and $\mathbf{u}$ are the input feature vector and output vector, respectively, $cov(\cdot)$ is the covariance, and $var(\cdot)$ is the variance. The correlation criteria method is used to determine which input features are more important for the output prediction based on the Pearson correlation coefficients between input features and the output.

For the mutual information method, the ranking criteria examines the dependency measurement between two features. Let $\tilde{X}$ and $\tilde{Y}$ be random feature variables, $\tilde{x}$ and $\tilde{y}$ are their realizations. We start with Shannon's definition of entropy, defined as:

$$M(\tilde{Y}) = -\sum_{\tilde{y}} P(\tilde{y}) * \log\left(P(\tilde{y})\right) \tag{3}$$

$P(\tilde{y})$ is the probability of obtaining the value $\tilde{y}$, (3) represents the uncertainty (information content) in variable $\tilde{Y}$. Suppose we observe a variable $\tilde{X}$, then the conditional entropy is given by:

$$M(\tilde{Y}|\tilde{X}) = -\sum_{\tilde{x}}\sum_{\tilde{y}} P(\tilde{x},\tilde{y}) * \log\left(P(\tilde{y}|\tilde{x})\right) \tag{4}$$

This equation indicates that by observing a variable $\tilde{X}$, the uncertainty in the output $\tilde{Y}$ is reduced. The decrease in uncertainty is written as:

$$U(\tilde{Y},\tilde{X}) = M(\tilde{Y}) - M(\tilde{Y}|\tilde{X}) \tag{5}$$

where (5) presents the mutual information index between two variables $\tilde{Y}$ and $\tilde{X}$. If the mutual information index is zero, two variables are independent. Otherwise, they are dependent, implying that one variable can provide useful information about the other variable.

Based on our outage dataset structure, which contains a large size of outage features, firstly, the correlation criteria method is an efficient and feasible algorithm to identify which features are important concerning the outage restoration time. In the next step, when the total number of features is reduced, we use the mutual information method to further determine if two features have too much overlapping information. After obtaining the results, we first set up a threshold (i.e., 0.3) of the Pearson correlation coefficient to identify which features are important concerning the outage restoration time. The feature with the Pearson correlation coefficient lower than the threshold is removed. Then, the mutual information index is used to determine if any two features have overlapped information. In this case, if the mutual information index between two features is higher than a threshold (i.e., 0.9), the two features are considered to have too much overlapped information. The feature with the lower Pearson correlation coefficient is removed.

B. The SDESC Algorithm

Let $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_m] \in \mathbb{R}^{n\times m}$ represent the historical outage dataset, where n is the total number of outage-related features and m is the total number of outage events, $\mathbf{o}_i \in \mathbb{R}^{n\times 1}$ represent the outage-related information in the ith outage event including n features. And let o_{ij} denote the ith row and j^{th} column entry in $\mathbf{O}$. In actual grids, given the accumulation of historical outage data over many years, m is a large real integer. In this situation, to reduce the computational complexity, the sparse coding technique is adapted to decrease the complexity of outage event-based adjacency matrix construction and eigen-decomposition. Sparse coding is utilized to find a sparse representation of $\mathbf{O}$ that consists of a dictionary $\mathbf{D} \in \mathbb{R}^{n\times p}$ and a representation $\mathbf{R} \in \mathbb{R}^{p\times m}$ [16], and to make $\mathbf{O} \approx \mathbf{D} * \mathbf{R}$. It's worth noting that the value of p is substantially lower than that of m. Therefore, each column of $\mathbf{D}$ contains the features of the outage samples and is called a basis vector. Each column of $\mathbf{R}$ represents the p-dimensional representation of the raw data inputs with respect to the new basis vectors. To this end, adopting the sparse coding, a high dimensional matrix $\mathbf{O}$ is factorized to lower-dimensional representation $\mathbf{D}$ and $\mathbf{R}$.

[1]The data imbalance problem refers to an unequal distribution of classes in the training dataset. When the dataset is imbalanced, the trained model typically fails to capture the hidden features of minority groups. Thus, the performance of a supervised model may suffer from the fact that the distribution of the target variable is skewed.

To minimize the approximation error, this process is regulated as an optimization problem using the Frobenius norm :

$$\underset{\mathbf{D},\mathbf{R}}{\text{minimize}} \quad ||\mathbf{O}-\mathbf{D}\mathbf{R}||_F^2 \tag{6a}$$

$$\text{subject to} \quad ||\mathbf{r}_i||_F^2 < \epsilon \tag{6b}$$

where $\mathbf{r}_i$ denotes the ith column of the representation matrix $\mathbf{R}$, ϵ is a parameter representing the maximum value of the Frobenius norm of the representation matrix $\mathbf{R}$.

The optimization problem (6) is non-convex in $\mathbf{D}$ and $\mathbf{R}$, which is difficult to solve. Unlike most of the existing approaches that compute dictionary and representation iteratively, we first obtain $\mathbf{D}$ by finding the dictionaries of $\mathbf{O}$ and then solve (6) by fixing $\mathbf{D}$. Suppose $\mathbf{D}$ is fixed, the optimization problem (6) becomes a regularized least squares problem, which is convex in $\mathbf{R}$. In addition, $\mathbf{R}$ is always expected to be sparse to provide a better data representation. However, there are still two challenges for this regularized least squares problem with sparse requirements: 1) The column number of $\mathbf{R}$, i.e., the total number of outage events, can be very large, leading to a large amount of computational burden. In our case, given that the outage dataset is updated by the utility every year, the increasing data size will gradually affect the computation time of the problem; 2) To favor the sparse $\mathbf{R}$, one common way is to include the L_1 regularization in the objective function, which is known to produce sparse solutions. But for the L_1 regularization, it is not continuously differentiable, and the most straightforward gradient-based methods are difficult to apply [18]. Thus, in our case, we adopt the Nadaraya-Watson kernel regression approach to handle the above challenges.

Before determining $\mathbf{R}$, a cluster ensemble framework is introduced to combine various clustering algorithms [19] to determine $\mathbf{D}$. Specifically, a density-based spatial clustering of applications with noise is first utilized to cluster all the data points, and then use the cluster centers to form $\mathbf{D}$, which is tabulated as Algorithm 1 [20]. Two user-defined hyperparameters, a threshold for the minimum number of neighbors, γ, and the radius, ξ, are utilized to perform a minimum density level estimation. $\mathbf{o}_i$ with more than γ neighbors within ξ distance are considered to be the centroid. All neighbors within the ξ radius of the centroid are considered to be part of the same group as this centroid. This method is capable of finding clusters with arbitrary shapes and sizes and shows robustness and practicality because it does not require *a priori* specification on the number of clusters. By selecting p centroids from the data points as dictionary points, we can form the dictionary matrix $\mathbf{D}$.

Let $\mathbf{d}_j$ be the jth column vector of $\mathbf{D}$, r_{ji} be the jth row and i^{th} column entry of $\mathbf{R}$. A natural assumption is that r_{ji} should be larger if $\mathbf{o}_i$ is closer to $\mathbf{d}_j$. To emphasize this assumption, we set the r_{ji} to zero as $\mathbf{d}_j$ is not among the $\hat{r}$ ($\hat{r} \leqslant p$) nearest neighbors of $\mathbf{o}_i$. This restriction also satisfies the dimension condition of the sparse representation matrix $\mathbf{R}$.

Let $\mathbf{D}_{\langle i\rangle} \in \mathbb{R}^{n\times\hat{r}}$ be a sub-matrix of $\mathbf{D}$ contains $\hat{r}$ nearest dictionaries of $\mathbf{o}_i$. Let $\mathbb{D}_{\langle i\rangle}$ denote the set consisting of all the column vectors in $\mathbf{D}_{\langle i\rangle}$. r_{ji} can be calculated using the following equation:

$$r_{ji} = \frac{K_h(\mathbf{o}_i, \mathbf{d}_j)}{\sum_{\mathbf{d}_{j'}\in\mathbb{D}_{\langle i\rangle}} K_h(\mathbf{o}_i, \mathbf{d}_{j'})} \qquad \mathbf{d}_j \in \mathbb{D}_{\langle i\rangle} \tag{7}$$

with

$$K_h(\mathbf{o}_i, \mathbf{d}_j) = \exp(-||\mathbf{o}_i - \mathbf{d}_j||^2/2\,h^2)$$

where $K_h(\cdot,\cdot)$ is a pre-defined Gaussian kernel function with a bandwidth h.

After applying the sparse coding, the representation matrix $\mathbf{R}$ can be represented as an undirected similarity graph, $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ with vertex set $\mathbb{V}$, where each vertex $\mathbf{v}_i$ in this set represents a data point $\mathbf{r}_i$. $\mathbb{E}$ is a set of edges connecting different vertices. We assume that the graph $\mathbb{G}$ is weighted, that means each edge between two vertices $\mathbf{v}_i$ and $\mathbf{v}_j$ carries a non-negative weight w_{ij}. The weighted adjacency matrix of the graph $\mathbb{G}$ is the matrix $\mathbf{W} = [w_{ij}]$, where $i, j = 1, \ldots, m$. While $w_{ij} > 0$ indicates the similarity between two selected vertices. $w_{ij} = 0$ indicates that two selected vertices $\mathbf{v}_i$ and $\mathbf{v}_j$ are not connected by an edge. To build the entry of the adjacency matrix $\mathbf{W}$, we have adopted the Gaussian kernel function written as follows:

$$w_{ij} = \exp\left(\frac{-||\mathbf{v}_i - \mathbf{v}_j||^2}{\alpha^2}\right) \tag{8}$$

where α is a scaling parameter that indicates how fast the weight decreases with the distance between the two vertices $\mathbf{v}_i$ and $\mathbf{v}_j$. To avoid the error caused by manual parameter selection, a localized scaling parameter α_i is calculated for each vertex, which allows self-tuning of the point-to-point distances based on the local distance of the neighbor of $\mathbf{v}_i$ [21]:

$$\alpha_i = ||\mathbf{v}_i - \mathbf{v}_\beta|| \tag{9}$$

where $\mathbf{v}_\beta$ is the βth neighbor of $\mathbf{v}_i$. Then, the weight (8) can be reformulated as follows:

$$w_{ij} = \exp\left(\frac{-||\mathbf{v}_i - \mathbf{v}_j||^2}{\alpha_i\alpha_j}\right). \tag{10}$$

By defining a couple of vertices and weight matrix $\mathbf{W}$, the outage data grouping is converted to a graph partitioning problem. The graph partitioning process divides a graph into disjoint sets of vertices by removing the edges connecting two groups. An optimized graph partitioning process is achieved when the edges between different sets have low weights, and the edges within a set have high weights. The objective function of the graph partitioning is to maximize both the dissimilarity between the disparate groups and the total similarity within each group:

$$S(\mathbb{G}) = \min_{\mathbb{A}_1,\mathbb{A}_2,\ldots,\mathbb{A}_\theta} \sum_{i=1}^{\theta} \frac{\tilde{s}\left(\mathbb{A}_i, \overline{\mathbb{A}_i}\right)}{d\left(\mathbb{A}_i\right)} \tag{11}$$

where θ is the number of vertices, $\mathbb{A}_i$ is a subset belonging to $\mathbb{V}$, $\tilde{s}(\mathbb{A}_i, \overline{\mathbb{A}_i})$ denotes the sum of the weights between vertices in $\mathbb{A}_i$ and vertices in $\overline{\mathbb{A}_i}$, i.e., the complement of $\mathbb{A}_i$. $d(\mathbb{A}_i)$ denotes the sum of the weights of vertices in $\mathbb{A}_i$. According to [22], the minimum of $S(\mathbb{G})$ is obtained at the second smallest eigenvector of the Laplacian matrix. Then, the graph Laplacian matrix is

Algorithm 1: Dictionary Selection in SDESC.

Initialization: Initialize $i \leftarrow 1$, γ, ξ
repeat
[**S1**]: Select the ith column of $\mathbf{O}$.
[**S2**]: Pick $\mathbf{o}_i$ and retrieve all direct density-reachable points in $\mathbf{O}$ using ξ.
[**S3**]: Based on γ, if $\mathbf{o}_i$ is a core point, a cluster is formed; otherwise, assign $\mathbf{o}_i$ to noise
[**S4**]: Update $i \leftarrow i + 1$.
until $i = m$

formulated based on the weight matrix $\mathbf{W}$:

$$\mathbf{L} = \mathbf{I}^{-\frac{1}{2}}\mathbf{W}\mathbf{I}^{-\frac{1}{2}} \tag{12}$$

where $\mathbf{I}$ is a diagonal matrix which the ith row and i^{th} column element of $\mathbf{I}$ is the sum of elements in the ith row of $\mathbf{W}$. Therefore, to solve the graph partition problem (minimize $S(\mathbb{G})$), according to the *Rayleigh-Ritz Theorem*, the solution is acquired by using k ($2 \leqslant k \leqslant m$) smallest eigenvectors of the Laplacian matrix, which guarantees an approximate value of the *optimal cut* [23], [24]. The value of k can be determined by various clustering evaluation metrics, such as the Silhouette coefficient, Dunn's index, and Davies-Bouldin validation index (DBI). In this work, to set the optimal k, we adopt the DBI, which purposes to minimize the overlap of different groups and maximize the conformance within each group. When we modify the k value to find the smallest eigenvalues of the Laplacian matrix, the corresponding DBI value is recorded for each k. The optimal value of k is determined when the DBI is minimized [25]. When the value of k is assigned, a new matrix $\mathbf{N} \in \mathbb{R}^{m\times k}$ is built based on the k smallest eigenvalues of the $\mathbf{L}$ matrix. Based on the properties of the graph Laplacians, the data point $\mathbf{r}_i$ is reconstructed using the ith row of the matrix $\mathbf{N}$, which enhances the cluster properties of the data [21]. After the data reconstruction, any clustering methods can be used to easily obtain the results. In this work, we used the k-means algorithm to obtain the final solution from the matrix $\mathbf{N}$. The overall process of the SDESC method is tabulated as Algorithm 2.

C. Method Comparison

Theoretically, typical pattern extraction of the outage dataset can be performed using any clustering algorithms, such as spectral clustering (SC), k-means, self-organizing maps, and hierarchical clustering. However, considering that the available dataset is high-dimensional and contains noise and extreme cases, several factors were prioritized in the selection of the clustering algorithm, including the curse of dimensionality and robustness against data noise. Hence, spectral clustering outperformed other conventional clustering algorithms by its outstanding data reconstruction and graph partitioning features. Specifically, the SC algorithm employs eigenvectors of graph matrices for data reconstruction. This data reconstruction process enhances the cluster properties of complex and unknown distributed datasets,

Algorithm 2: Sparse Dictionary-Based Ensemble Spectral Clustering.

Input: m data points $\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_m, \in \mathbb{R}^{n\times m}$
cluster number k
Output: k clusters
[**S1**]: Produce p dictionary points using random selection following **Algorithm 1** to form $\mathbf{D}$.
[**S2**]: Construct a sparse representation matrix $\mathbf{R} \in \mathbb{R}^{p\times m}$ between data points and dictionary points according to equation (7).
[**S3**]: Build weighted adjacency matrix $\mathbf{W}$ according to equation (10).
[**S4**]: Use k smallest eigenvectors of the Laplacian matrix from equation (12) to generate a new matrix $\mathbf{N} \in \mathbb{R}^{m\times k}$.
[**S5**]: Each row of $\mathbf{N}$ is a data point and apply k-means to get the clusters.

so that clusters can be easily detected from the reconstructed datasets [21], [22].

However, given the increasing size of the outage dataset for each year, traditional spectral clustering gradually shows a critical drawback: extremely high computational burden. Such a drawback is caused by the fact that spectral clustering requires the construction of an adjacency matrix and the computation of the eigen-decomposition of the Laplace matrix. For large-scale cases, the above two steps can cause an overwhelming computational burden [26], [27]. Hence, this suggests an urgent need for a new algorithm to solve this task.

The SDESC algorithm has three unique advantages over the conventional spectral clustering (SC) method in this task: 1) The enhanced cluster property of the reconstructed dataset reduces the sensitivity of the clustering process to outliers, which are unavoidable in real-world applications. 2) The proposed method introduces the dictionary-based weight matrix of the dataset rather than computing the high-dimensional profiles of all available outage data directly. Such a strategy can significantly reduce the complexities of computing the adjacency matrix and graph Laplacian matrix from $O(m^2)$ and $O(m^3)$ to $O(pm)$ and $O(p^3 + p^2m)$ due to the fact that the value of p is substantially lower than that of m, which is beneficial in the Big Data age. The complexity analysis can be found in [28]. 3) The graph partitioning problem could be settled without making any assumptions about the data distribution. This step enhances the robustness of the clustering method, thus resulting in better performance for complicated outage data structures.

IV. Outage Restoration Time Prediction

To choose our baseline algorithm to predict the outage restoration time, several state-of-art methods, such as artificial neural network (ANN) and convolutional neural network (CNN), are evaluated and compared based on the dataset we acquired. CNN is one of the most successful advanced deep learning algorithms and has been widely utilized in the image and video fields. The basic idea of CNN is to automatically extract the important

features from the input using convolutional operations. The algorithm should be used preferentially when the data has spatial characteristics. However, the data type for this work is sequential and time-serial, and we do not consider spatial information at this stage. Also, considering the limited outage-based features recorded by the utilities, it is not necessary to use CNN as the baseline model in this work [29]. Moreover, CNN generally requires more training data than the ANN to obtain a good model. This creates a hindrance when utilities apply CNN models to predict restoration time in actual grids, especially for small utilities with limited outage data. Hence, we believe that ANN is a more suitable baseline model for this work.

Upon the outage data partitioning results, each outage subset is first assigned with an artificial neural network to estimate the outage restoration time. As discussed in Section I and Section II-B, there may be huge gaps between the amount of data available for different patterns and scales of outage events. However, there are always internal relationships between different outage event patterns. To this end, a transfer-learning-based method is then employed to transfer the learned knowledge from one prediction model to enhance the performance of the rest models. To help the reader understand the proposed model, we first briefly revisit the concept and properties of ANN, then describe the transfer learning strategy in detail.

A. Artificial Neural Network

In this work, We adopt the multi-layer feed-forward ANN structure, in which the sigmoid active function is used in the hidden layers. We first performed the feature selection procedure to select the important outage-related features. Then, the outage dataset with selected features is used to train the ANN to predict the outage restoration time. The loss function is the total mean squared error between the ANN prediction and the ground truth, i.e., the real outage restoration time in the dataset. In this work, the choice of the optimizer is determined by the highest prediction accuracy on the validation set.

For calibration purposes, stochastic gradient descent, Adam, Levenberg-Marquardt (LM), and RMSprop have been tested using the same dataset. The LM method shows the highest detection accuracy on the validation set. Therefore, the LM backpropagation method is utilized in this work to update the weights and threshold parameters [30]. The LM method is derived from Newton's method to minimize sum-of-squares error functions [31]. It can automatically adjust the learning rate in the direction of the gradient using the Hessian matrix. Compared to backpropagation methods with a constant learning rate, the LM method significantly boosts the training speed [32]. To calibrate the parameters of ANN, the optimal set of hyper-parameters is determined by the grid search method [33].

B. Transfer Learning-Added Outage Restoration Time Prediction Model

When the training targets are multiple related tasks (i.e., restoration time prediction for outages under different scales and factors), conventional machine learning-based methods need to train multiple models from scratch, thus requiring a large and comprehensive dataset. Such a requirement renders their practical implementation costly. In contrast, transfer learning-based models greatly reduce the amount of data required for training by leveraging prior knowledge gained from previous training tasks [34]. Therefore, in this work, a transfer learning strategy is adopted to discover domain-invariant intrinsic outage features and structures under different but related domains, which establishes the re-utilization of data information across domains.

In this work, the *source model* is defined as a pre-trained outage restoration time estimation model with the neural network parameters, while the data samples and their predicted values are stored in a *knowledge matrix*. A pre-trained model is a model that is trained on a large benchmark dataset to solve a task that is similar to the one that we want to solve accordingly. The *learning task* is defined as the upcoming training assignment of an untrained outage subset, while the *learning matrix* is filled with the data samples in the corresponding subset. To choose the pre-trained source model in this work, we adopt one of the cluster-wised subsets that we obtained from Section III-B. In general, to find a fair source model, it is important to choose a dataset that is relatively large and contains a variety of data patterns. By observing our clustered subsets, two subsets have a relatively scarce amount of data points than the largest subset, it is not sufficient to train a fair model. Therefore, the largest subset is utilized to treat as the source model. The rationale behind this is that it consists of the most frequently occurring event patterns that provide a baseline to train a good source model. Once the source model is determined, the knowledge matrix with predicted values of this subset is first obtained during the training process. The neural network parameters are reserved and exploited as the initial parameters for a new learning task. Specifically, the neural network for the new learning task is formed by mapping the reserved learning parameters to layers. The prediction value for the new learning task is then obtained by training the re-formed model. After obtaining all predicted values of the new learning task, they are combined with the original data samples into a complete learning matrix that has the same structure as the knowledge matrix.

This transfer learning process can be depicted in Fig. 4. In detail, after the source model and learning tasks are distinguished by an evaluation of the cluster-wised outage subsets, the transfer learning process gathers the outage-related features and the output (i.e., actual restoration time) in the pre-trained model, and stores them as a knowledge matrix. Similarly, each learning task, which is an untrained model, generates a learning matrix containing data points with outage features in the particular subset. After the first training process, by updating the parameters for the new learning task, the predicted restoration time can be obtained by training the new model. When the first prediction task is completed, the learned model can be utilized in a recursive manner when dealing with a new learning task [35]. For example, the training tasks for cluster-wised subsets are signed as task 1, 2, and 3. Task 2 is learned by exploiting task 1 as a source model, then task 2 can be used as the source model for training task 3.

Compared to the conventional machine learning-based method, the proposed transfer learning-added outage restoration time prediction method using cluster-wised datasets has greatly

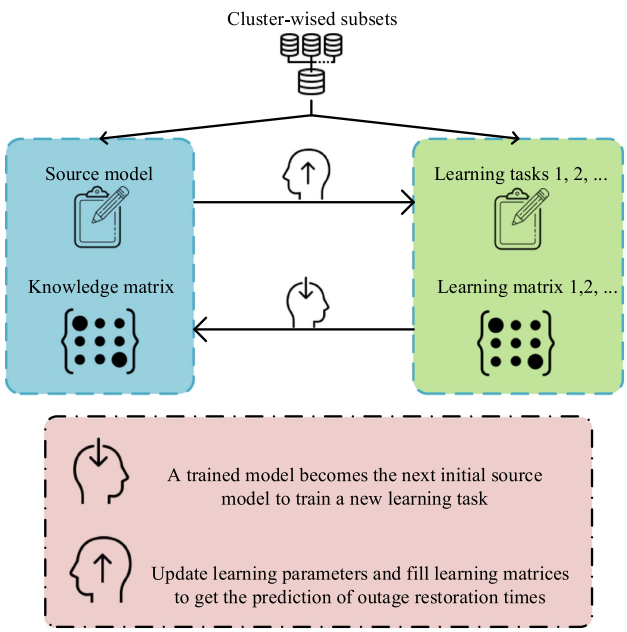


Fig. 4. Exemplary of the transfer learning process.

reduced the overfitting risk caused by the data scarcity of the specific outage prediction patterns and distributions. This can be confirmed using the numerical results.

C. Unseen Outage Classification

To identify and allocate the corresponding outage pattern and related prediction model to unseen outage samples, a t-distributed stochastic neighbor embedding (t-SNE) method is utilized. Initially, the Euclidean distance is involved in mapping the high-dimensional data. In the t-SNE algorithm, it starts by converting the high-dimensional Euclidean distance between two data points into conditional probabilities which represent their similarity. This conditional probability between two data points $\mathbf{o}_j$ and $\mathbf{o}_i$ is represented as $P(\mathbf{o}_j|\mathbf{o}_i)$. It means that $\mathbf{o}_i$ would pick $\mathbf{o}_j$ as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at $\mathbf{o}_i$. The Euclidean distance is smaller when the conditional probability is larger. Data with higher similarity in high-dimensional space is closer to each other after being embedded in a low-dimensional space [36], [37]. By optimizing the conditional probability between original data and analog data, the t-SNE can convert high-dimensional outage data into low-dimensional representations. Mathematically, the conditional probability between any two outage samples $\mathbf{o}_j$ and $\mathbf{o}_i$ in high-dimensional space can be formulated as:

$$P(\mathbf{o}_j|\mathbf{o}_i) = \frac{\exp\left(\frac{-||\mathbf{o}_i-\mathbf{o}_j||^2}{2\lambda_i}\right)}{\sum_{\kappa=1,\kappa\neq i}^{m}\exp\left(\frac{-||\mathbf{o}_i-\mathbf{o}_\kappa||^2}{2\lambda_i}\right)} \tag{13}$$

where κ is the number of neighbor points, λ_i is the vector variance of the Gaussian function centered on the data $\mathbf{o}_i$. Next, t-SNE utilizes symmetrized probability to alleviate the crowding problem that is illustrated in [38]. The symmetrical conditional probability between two data samples $\mathbf{o}_j$ and $\mathbf{o}_i$ is represented as:

$$P(\mathbf{o}_i|\mathbf{o}_j) = \frac{\exp\left(\frac{-||\mathbf{o}_j-\mathbf{o}_i||^2}{2\lambda_j}\right)}{\sum_{\kappa=1,\kappa\neq j}^{m}\exp\left(\frac{-||\mathbf{o}_j-\mathbf{o}_\kappa||^2}{2\lambda_j}\right)} \tag{14}$$

where κ is the number of neighbor points, λ_j is the vector variance of the Gaussian function centered on the data $\mathbf{o}_j$. The joint probability of $\mathbf{o}_i$ and $\mathbf{o}_j$, within a Gaussian space is:

$$P(\mathbf{o}_i,\mathbf{o}_j) = \frac{P(\mathbf{o}_i|\mathbf{o}_j)+P(\mathbf{o}_j|\mathbf{o}_i)}{2\,m}. \tag{15}$$

In a low-dimensional space, the t distribution is applied with one degree of freedom. The joint distribution of two simulated data $\mathbf{z}_i$ and $\mathbf{z}_j$ is calculated as follows:

$$P(\mathbf{z}_i,\mathbf{z}_j) = \frac{\left(1+||\mathbf{z}_i-\mathbf{z}_j||^2\right)^{-1}}{\sum_{\kappa=1,\kappa\neq l}^{m}\left(1+||\mathbf{z}_\kappa-\mathbf{z}_l||^2\right)^{-1}}. \tag{16}$$

The Kullback Leibler (KL) divergence is utilized to quantify the similarity between $P(\mathbf{z}_i,\mathbf{z}_j)$ and $P(\mathbf{o}_i,\mathbf{o}_j)$:

$$\phi = \sum_{i=1}^{m}\sum_{j=1}^{m} P(\mathbf{o}_i,\mathbf{o}_j)\log_2\frac{P(\mathbf{o}_i,\mathbf{o}_j)}{P(\mathbf{z}_i,\mathbf{z}_j)}. \tag{17}$$

The optimal low-dimensional data is obtained by minimizing the KL divergence using the gradient descent method, which is denoted as:

$$\frac{d\phi}{d\mathbf{z}_i} = 4\sum_{j=1}^{m}\left(P(\mathbf{o}_i,\mathbf{o}_j)-P(\mathbf{z}_i,\mathbf{z}_j)\right) \cdot (\mathbf{z}_i-\mathbf{z}_j)\left(1+||\mathbf{z}_i-\mathbf{z}_j||^2\right)^{-1} \tag{18}$$

When an outage occurs, the high-dimensional data is first converted into a two-dimensional map using the t-SNE method. Then, the distances between the edges of each subset with this point are calculated to quantify the similarity between the unseen outage and the learned subsets. Based on the distance values, the most probable class is chosen as the correct underlying subset for an unseen outage.

V. Numerical Results

This section explores the practical effectiveness of the proposed outage restoration time prediction method. A real-world dataset with 16,000 outage samples is utilized in this case study, which includes six years of data collected by a utility in New York State. After data preprocessing, the whole dataset is randomly divided into three parts for training, testing, and validation by 70%, 15%, and 15% of the total data, respectively.

A. SDESC Algorithm Performance

Fig. 5 presents the data distribution of the number of customer interruptions versus outage restoration time. In real-world scenarios, most people would believe that there should be a clear relationship between the restoration time and the number of affected customers. As described in this figure, the number of customers interrupted has a clear impact on the restoration time.

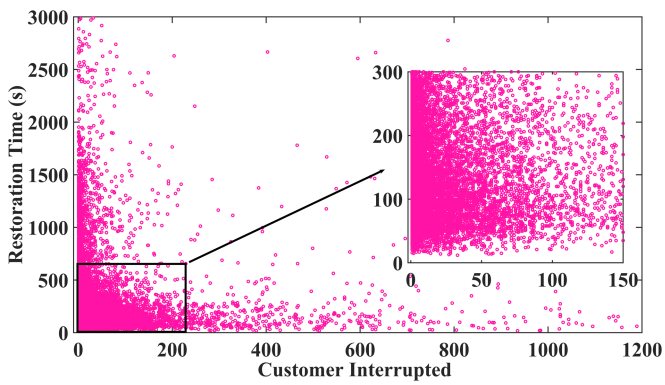


Fig. 5. Data mapping by customer interrupted vs. restoration time.

TABLE II
CLUSTERING STATISTICS

Cluster	Samples	Avg. CI	Avg. RT(min)
$\mathbf{C}_1$	2379	170	740.5
$\mathbf{C}_2$	5302	21	288.4
$\mathbf{C}_3$	2884	16	144.5
$\mathbf{C}_4$	5872	22	82.2

The rationale behind this is that when the number of affected customers is high, the utility usually prioritizes these events. The corresponding outage restoration time is often at a low level, as shown on the right side of Fig. 5. In contrast, estimating restoration time for small-scale outages is more challenging, as they are more likely to be affected by a variety of factors. This can also be confirmed using real-world data, as shown in Fig. 5. Therefore, it is necessary to implement the proposed SDESC method to distinguish the outage groups that are constrained by various features.

Basically, the SDESC calibration is a trial and error process using a specific cluster evaluation metric. In this work, the optimal number of subsets, k, is assigned as 4 based on the minimum DBI value. The grouping results with the corresponding k value are marked in Fig. 5. Specifically, each color represents a subset of the outage data, namely $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$, and $\mathbf{C}_4$. Table II demonstrates the statistics in the grouping results, including the number of data samples in each subset, the average number of customers interrupted (i.e., Avg. CI), and the average restoration time in minutes (i.e., Avg. RT). As shown in table, $\{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_4\}$ consist of $\{2379, 5302, 2884, 5872\}$ outage data samples, respectively. Such results promise the data imbalance problem: subset $\mathbf{C}_2$ and $\mathbf{C}_4$ have twice as many data samples as $\mathbf{C}_1$ and $\mathbf{C}_3$, while the average recovery time (i.e., 740.5 minutes) and the number of customers interrupted (i.e., 170) of $\mathbf{C}_1$ are significantly higher than for the other subsets. In our view, $\mathbf{C}_1$ refers to severe outages with a higher Avg. RT and Avg. CI, but relatively infrequent. $\mathbf{C}_2$ and $\mathbf{C}_4$ represent intermediate and least serious outages, which are twice as frequent as severe outages. $\mathbf{C}_4$ represents a subset of minor outages, which occur frequently but can typically be resolved in a timely manner.

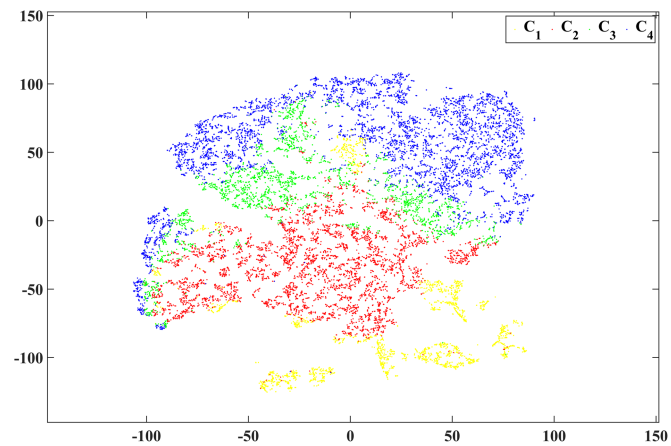


Fig. 6. t-SNE plot of clustered data using the proposed SDESC method.

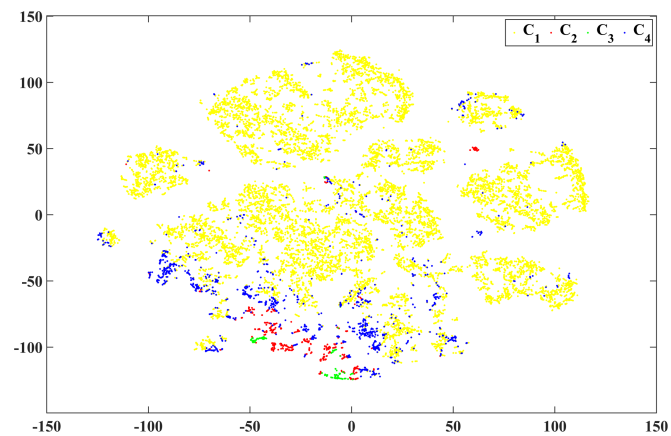


Fig. 7. t-SNE plot of clustered data using the advanced k-means method.

B. Data Virtualization in Low-Dimensional Representation

In this paper, the main function of t-SNE is used for data visualization (Figs. 6 and 7) and to enhance the overall interpretability of the framework. Observing the distributions of the cluster-wised data in a low-dimensional map will help us to visualize and briefly evaluate the performance of the proposed clustering algorithm. Theoretically, the similarity between the unseen outage and the learned outage subsets can be calculated based on specific similarity scores, such as cosine similarity, to classify unseen outages to the existing typical patterns. However, considering the satisfactory result of our clustering process, this task is not a critical challenge in this work. Therefore, at this stage, we did not investigate which method or metrics is one of the optimal solutions to this task. This will be one of the future research directions.

The t-SNE plot for cluster-wised outage data using the SDESC is shown in Fig. 6. The shortest distances between the edges of each subset with each unseen outage are calculated to measure the similarity between the unseen outage and the learned patterns. Based on the testing dataset (15% of the total data), the classification error margin can achieve 5% when assigning outage patterns to unseen outages.

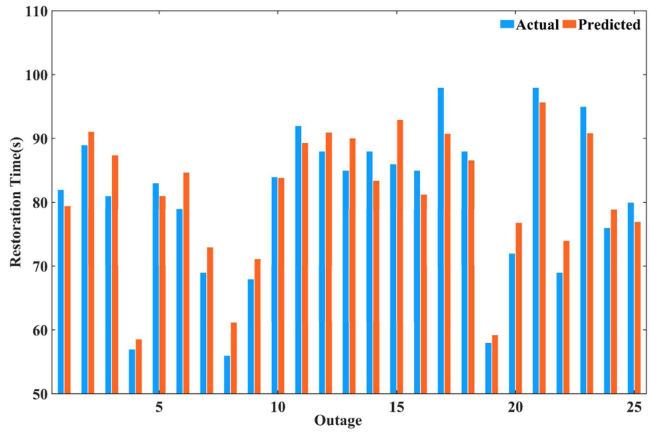
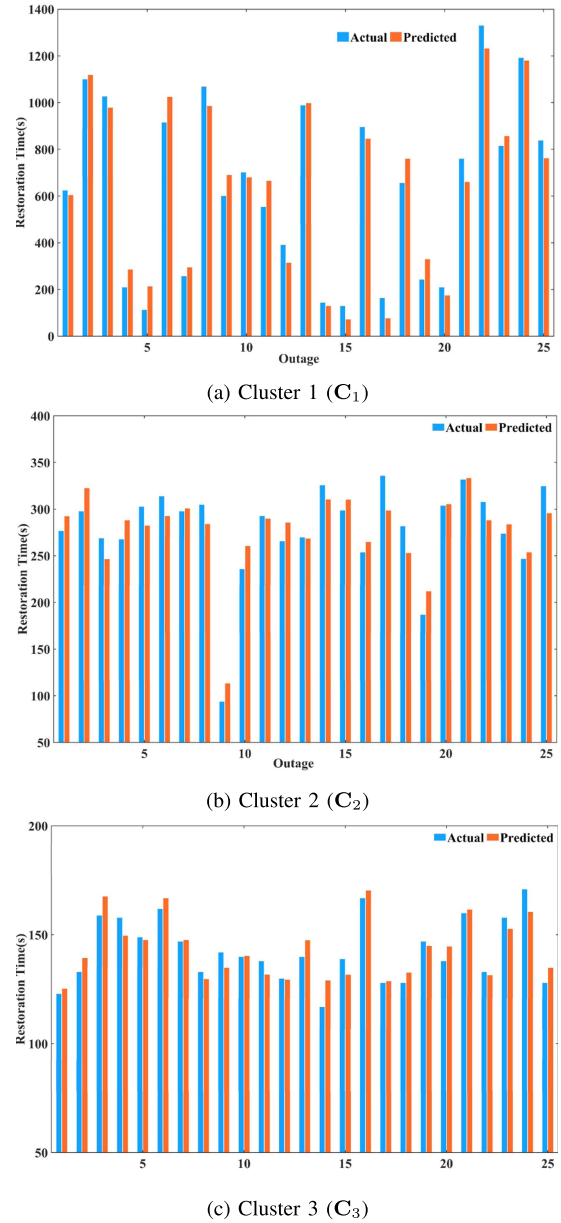


Fig. 8. Comparison result between actual and predicted restoration time for the source model ($\mathbf{C}_4$).

We have conducted a numerical comparison with an advanced k-means algorithm [39] to show that the proposed SDESC method can offer a dramatic improvement in outage data grouping. Fig. 7 shows the t-SNE plot of the result using the advanced k-means algorithm. By using this state-of-the-art clustering method, over 70% of the total data has fallen into a single cluster subset, $\mathbf{C}_1$. Such a result increases the overfitting risk caused by data insufficient in other subsets. Moreover, it is clear that the data points in different subsets of Fig. 7 are more difficult to be classified than in Fig. 6. This indicates that the homogeneity of each subset obtained from the advanced k-means is much lower than that of each subset obtained from the proposed method. Therefore, when an unseen outage occurs, it is highly likely to misclassify, thus resulting in a decrease in restoration time prediction accuracy.

C. Outage Restoration Time Prediction Performance Analysis

When the outage dataset is separated using the SDESC method, the ANN with a transfer learning embedded model is utilized to predict the outage restoration time. By observing our clustered subset in Table II and Fig. 6, two subsets have a significantly lower amount of data points than the subset $\mathbf{C}_4$, it is not sufficient to train a good model. Therefore, we utilize $\mathbf{C}_4$ as the source model, this is because $\mathbf{C}_4$ represents the most commonly occurred event (about 35.7% of the total data), and this provides the baseline to train a good model. Then, training predictive models on other data subsets are considered as learning tasks. Specifically, we train a model using the training set in $\mathbf{C}_4$, and the validation set in $\mathbf{C}_4$ is used to optimize the model's hyperparameters. After the complete training, $\mathbf{C}_1$ is trained using the pre-trained model $\mathbf{C}_4$. This process is repeated for other subsets by leveraging the previous pre-trained combined model. To evaluate the prediction performance of the ANN, the mean absolute percentage error (MAPE) is utilized in this paper. In addition to MAPE, the percentage of predicted restoration time that falls within the reasonable range from the actual restoration time is also calculated to further evaluate the performance of our method. Fig. 8 describes the comparison between the actual and predicted restoration time for 25 randomly selected samples in $\mathbf{C}_4$. After predicting the restoration time of the test data, the predicted restoration time range is 22.8 minutes, which is below the 30 minutes threshold. 3% of the predicted time is more than 60 minutes of the actual repair time, and only one particular outage showed that the predicted time is more than 90 minutes of the actual repair time.



(a) Cluster 1 ($\mathbf{C}_1$)

(b) Cluster 2 ($\mathbf{C}_2$)

(c) Cluster 3 ($\mathbf{C}_3$)

Fig. 9. Comparison result between actual and predicted restoration time for learning tasks.

The proposed method is utilized to train the prediction models for $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$ using the source model. The MAPE for $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$ is 23%, 24%, and 11.7%, respectively. The predicted restoration time range for all three subsets is 48.8 minutes, while subset $\mathbf{C}_3$ had an outstanding prediction range of 21.9 minutes. Fig. 9 demonstrates the comparison between the actual and predicted restoration time for 25 randomly selected

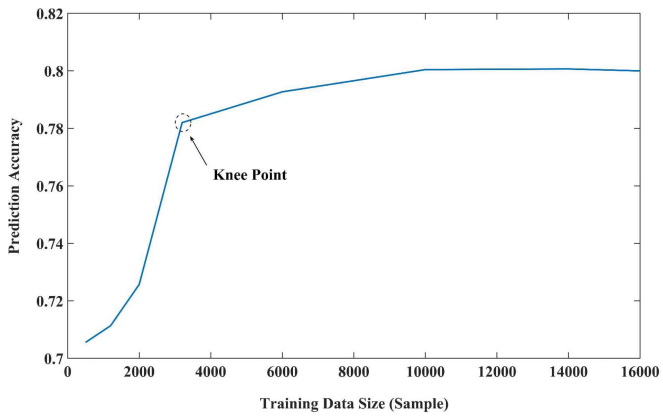


Fig. 10. Sensitivity analysis between the number of training data and accuracy.

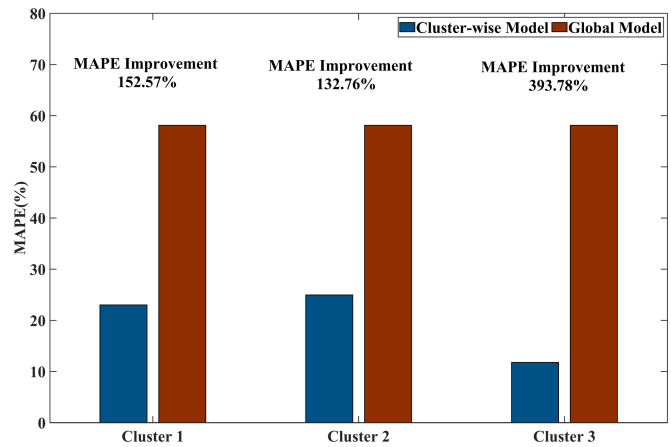


Fig. 11. Restoration time comparison between cluster-wised model and global model.

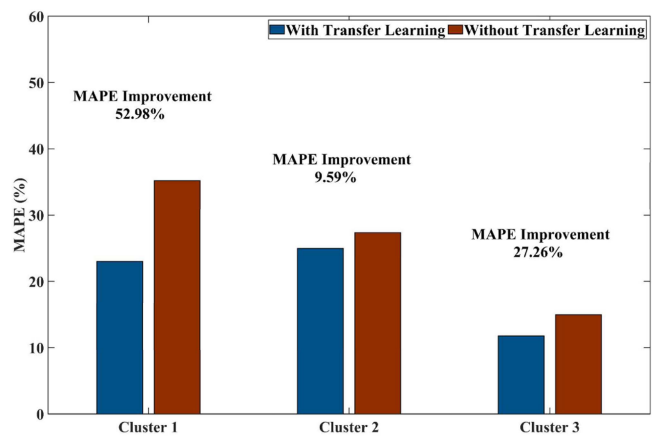


Fig. 12. Prediction of restoration time with and w/o transfer learning approach.

outages in $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$. Note that the variance in restoration time for each subset is 1300 seconds, 350 seconds, and 170 seconds. Despite the high variance of $\mathbf{C}_1$, the prediction model still performs decently. The prediction accuracy of $\mathbf{C}_1$ is slightly lower than that of $\mathbf{C}_2$ and $\mathbf{C}_3$. This result is reasonable because higher data variance usually leads to a higher risk of overfitting and reduces the accuracy of the prediction model.

D. Sensitivity and Uncertainly Analysis

To demonstrate the sensitivity of the proposed method to the size of training data, we have tested the average performance of the proposed model under various sizes of training dataset as shown in Fig. 10. As is demonstrated in the figure, the performances of our model can reach acceptable prediction accuracy on a training set with around 3,000 data samples. This is equal to a half year to one year of the data samples, depending on the varying frequency of outages. After reaching the knee point on the figure, our method tends to become stable with increasing numbers of data samples. Based on the literature reviews in Table I, most of the recently published works require data sizes beyond this knee point [6], [7], [8].

In this work, uncertainty analysis is reflected by the imperfection of the data by adding noise. Specifically, noise samples were generated from a normal distribution with zero mean and 1% variance and added to the continuous features to represent standard measurement deviations. In terms of the categorical features, 2% of the data were randomly adjusted to represent error-induced uncertainty.

E. Method Comparison

1) Proposed Model Vs. Global Model: We have conducted a comprehensive comparison between the proposed cluster-wised model with the previous restoration time prediction model [8]. Note that the previous model follows a global training fashion and is developed using all outage records without clustering. Where possible, we attempted to tune the parameters for each algorithm to give a fair comparison. The MAPE improvement compared to the global model for each of the subsets are 152.57%, 132.76%, and 393.78%, respectively, as shown in Fig. 11. This result indicates that our SDESC method can substantially improve the prediction performance by reducing the uncertainty of the real-world outage data compared to the global model.

2) Proposed Model Vs. Conventional Cluster-Wise Learning-Based Model: Another numerical comparison has been conducted between the proposed method with the conventional cluster-wise-based method. This conventional method trained independent neural networks for each cluster-wised subset. Such a comparison can further demonstrate that our method can achieve good prediction performance. Both methods are evaluated based on the same neural network configurations to ensure a fair comparison. As is demonstrated in Fig. 12, for the three different outage subsets $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$, compared to the conventional cluster-wise-based method, our transfer-learning-added model has 52.98%, 9.59%, and 27.26% MAPE improvement, respectively. The results show that the transfer learning strategy can meet the challenges posed by real-world unbalanced outage datasets and greatly improve the accuracy of repair time prediction.

3) Proposed Model Vs. Previous Related Works: To show the ability of our proposed method with previous related works, we have conducted numerical comparisons with three methods, including a state-of-the-art regression method, support vector

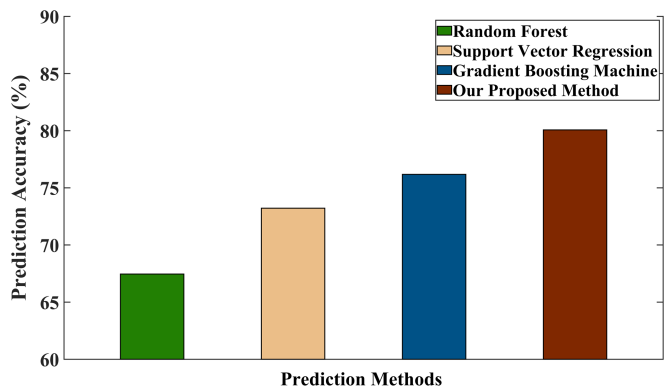


Fig. 13. Comparison of prediction results with three existing methods.

regressions (SVR) [40], and two recently published models for outage restoration time prediction [41] and [42]. Specifically, in [41], a random forest (RF) based approach is proposed to predict the outages with related weather variables. In [42], a gradient boosting machine (GBM) learning model is utilized to discover the power outage data, including outage frequency and duration. The comparison results are demonstrated in Fig. 13.

It can be observed that [42] and our proposed method can achieve good accuracy with the available outage dataset, while [41] and the SVR approaches do not exhibit good performance. The result between the GBM and our method is competitive, and our method is slightly better than the GBM. The difference between those two approaches is that the proposed method overcomes the overfitting risk caused by data imbalance and scarcity problems using the clustering ensemble and the transfer learning strategy. In contrast, gradient boosting models can overemphasize data with noise and easily cause overfitting. Meanwhile, our proposed method provides an efficient processing facility for future unseen data. The characteristics of the data sample can be rapidly identified according to the clustering result; this will help the utility select the corresponding restoration plan for the specific data pattern, and the restoration time can be estimated using the trained subset simultaneously.

VI. Conclusion

This article presents a novel data-driven approach to accurately predict outage restoration time using transfer learning with cluster ensembles. In this paper, six years of real-world outage dataset from our utility partner is investigated for model development and validation. The proposed SDESC approach utilizes the sparse coding technique and cluster ensemble mechanism to first decompose the large-scale datasets, which has good computational efficiency and scalability. Based on the learned outage patterns, the developed transfer-learning-added model can not only accurately predict the outage restoration time in each subset, but also addresses two fundamental challenges: 1) neglect the uncertainty caused by the heterogeneity of outage events with different scales and factors; 2) data imbalance problem in different data subsets. Based on the available real-world utility data, the results show that the proposed method has improved performance compared to existing methods and has overcome large-scale data challenges.

References

[1] A. Jaech, B. Zhang, M. Ostendorf, and D. S. Kirschen, "Real-time prediction of the duration of distribution system outages," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 773–781, Jan. 2019.

[2] Y. Zhou, A. Pahwa, and S. S. Yang, "Modeling weather-related failures of overhead distribution lines," *IEEE Trans. Power Syst.*, vol. 21, no. 4, pp. 1683–1690, Nov. 2006.

[3] F. Yang, P. Watson, M. Koukoula, and E. N. Anagnostou, "Enhancing weather-related power outage prediction by event severity classification," *IEEE Access*, vol. 8, pp. 60029–60042, 2020.

[4] P. Kankanala, S. Das, and A. Pahwa, "AdaBoost$^+$: An ensemble learning approach for estimating weather-related outages in distribution systems," *IEEE Trans. Power Syst.*, vol. 29, no. 1, pp. 359–367, Jan. 2014.

[5] M. S. Bashkari, A. Sami, and M. Rastegar, "Outage cause detection in power distribution systems based on data mining," *IEEE Trans. Ind. Informat.*, vol. 17, no. 1, pp. 640–649, Jan. 2021.

[6] M.-Yuen Chow, L. S. Taylor, and Mo-Suk Chow, "Time of outage restoration analysis in distribution systems," *IEEE Trans. Power Del.*, vol. 11, no. 3, pp. 1652–1658, Jul. 1996.

[7] H. Liu, R. A. Davidson, and T. V. Apanasovich, "Statistical forecasting of electric power restoration times in hurricanes and ice storms," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 2270–2279, Nov. 2007.

[8] A. Arif and Z. Wang, "Distribution network outage data analysis and repair time prediction using deep learning," in *Proc. IEEE Int. Conf. Probabilistic Methods Appl. to Power Syst.*,, 2018, pp. 1–6.

[9] M. Yue, T. Toto, M. P. Jensen, S. E. Giangrande, and R. Lofaro, "A Bayesian approach-based outage prediction in electric utility systems using radar measurement data," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6149–6159, Nov. 2018.

[10] A. Domijan Jr et al., "Effects of norman weather conditions on interuptions in distribution systems," *Int. J. Power Energy Syst.*, vol. 25, no. 1, pp. 54–61, 2005.

[11] P. Kankanala, A. Pahwa, and S. Das, "Estimation of overhead distribution system outages caused by wind and lightning using an artificial neural network," in *Proc. Int. Conf. Power Syst. Operation Plan.*, vol. 545, 2012.

[12] D. Owerko, F. Gama, and A. Ribeiro, "Predicting power outages using graph neural networks," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2018, pp. 743–747.

[13] National Oceanic and Atmospheric Administration, "Severe weather," 2021. [Online]. Available: https://www.ncei.noaa.gov/products/severe-weather

[14] National Oceanic and Atmospheric Administration, "Climate data online," 2021. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/

[15] L. Xu, M.-Y. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification e-algorithm," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 164–171, Feb. 2007.

[16] X. Chen and D. Cai, "Large scale spectral clustering with landmark-based representation," in *Proc. 25th AAAI Conf. Artif. Intell.*, 2011, pp. 313–318.

[17] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[18] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2006, pp. 801–808.

[19] J. Azimi and X. Fern, "Adaptive cluster ensemble selection," in *Proc. 21st Int. Joint Conf. Artif. Intell.*, 2009, pp. 992–997.

[20] J. Jang and H. Jiang, "DBSCAN++: Towards fast and scalable density clustering," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3019–3029.

[21] U. Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Mar. 2007.

[22] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.

[23] F. R. Chung and F. C. Graham, "Spectral graph theory," *Amer. Math. Soc.*, vol. 92, pp. 113–123, 1997.

[24] C. C. Aggarwal and C. K. Reddy, "Data clustering," in *Algorithms and Applications* (Chapman&Hall/CRC Data Mining and Knowledge Discovery Series). Londra: Springer, 2014, pp. 1–31.

[25] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Trans. Syst., Man, Cybern., Part B. Cybern*, vol. 28, no. 3, pp. 301–315, Jun. 1998.

[26] S. M. Bidoki, N. Mahmoudi-Kohan, M. H. Sadreddini, M. Z. Jahromi, and M. P. Moghaddam, "Evaluating different clustering techniques for electricity customer classification," in *Proc. IEEE PES T&D*, 2010, pp. 1–5.

[27] D. Vercamer, B. Steurtewagen, D. Van den Poel, and F. Vermeulen, "Predicting consumer load profiles using commercial and open data," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3693–3701, Sep. 2016.

[28] J. Liu, C. Wang, M. Danilevsky, and J. Han, "Large-scale spectral clustering on graphs," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1486–1492.
[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, pp. 321–350.
[30] S. Sapna et al., "Backpropagation learning algorithm based on Levenberg Marquardt Algorithm," *Comp. Sci. Inform. Technol.*, vol. 2, pp. 393–398, 2012.
[31] C. Lv et al., "Levenberg–marquardt backpropagation training of multilayer neural networks for state estimation of a safety-critical cyber-physical system," *IEEE Trans. Ind. Inform.*, vol. 14, no. 8, pp. 3436–3446, Aug. 2018.
[32] B. M. Wilamowski and H. Yu, "Improved computation for Levenberg-Marquardt training," *IEEE Trans. Neural Netw.*, vol. 21, no. 6, pp. 930–937, Jun. 2010.
[33] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.
[34] L. Torrey and J. Shavlik, "Transfer learning," *Handbook Res. Mach. Learn. Appl. Trends: Algorithms, Methods, Techn.*, pp. 242–264, 2010.
[35] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.
[36] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, "Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets," *Nature Commun.*, vol. 10, no. 1, pp. 1–12, 2019.
[37] X. Fang et al., "Blast furnace condition data clustering based on combination of T-distributed stochastic neighbor embedding and spectral clustering," *Proc. IEEE 15th Int. Conf. Control Automat.*, 2019, pp. 1608–1613.
[38] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[39] R. Yadav and A. Sharma, "Advanced methods to improve performance of k-means algorithm: A review," *Glob. J. Comput. Sci. Technol.*, vol. 12, no. 9, pp. 47–52, 2012.
[40] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.
[41] D. Cerrai et al., "Predicting storm outages through new representations of weather and vegetation," *IEEE Access*, vol. 7, pp. 29639–29654, 2019.
[42] T. Lawanson, V. Sharma, V. Cecchi, and T. Hong, "Analysis of outage frequency and duration in distribution systems using machine learning," in *Proc. 52nd North Amer. Power Symp.*, 2021, pp. 1–6.

Dingwei Wang (Graduate Student Member, IEEE) received the B.S. degree in electrical and computer engineering from Marquette University, Milwaukee, WI, USA, in 2018, and the M.S. degree in electrical and computer engineering from The George Washington University, Washington, D.C., USA, in 2020. He is currently working toward the Ph.D. degree with Iowa State University, Ames, IA, USA. His research interests include distribution system resilience, data analytics, and machine learning.

Yuxuan Yuan (Graduate Student Member, IEEE) received the B.S. degree in electrical and computer engineering in 2017 from Iowa State University, Ames, IA, USA, where he is currently working toward the Ph.D. degree. His research interests include distribution system state estimation, synthetic networks, data analytics, and machine learning.

Rui Cheng (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Hangzhou Dianzi University, Hangzhou, China, in 2015, and the M.S. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2018. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA. His research interests include the intersection of optimization, learning, and power systems, with particular applications to voltage/var control, demand response, and transactive energy markets.

Zhaoyu Wang (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiaotong University, Shanghai, China, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA. He is an Northrop Grumman Endowed Associate Professor with Iowa State University. His research interests include optimization and data analytics in power distribution systems and microgrids. He was the recipient of the National Science Foundation CAREER Award, Society-Level Outstanding Young Engineer Award from IEEE Power and Energy Society, Northrop Grumman Endowment, College of Engineering's Early Achievement in Research Award, and Harpole-Pentair Young Faculty Award Endowment. He is also a Principal Investigator for a multitude of projects funded by the National Science Foundation, the Department of Energy, National Laboratories, PSERC, and Iowa Economic Development Authority. He is the Co-TCPC of IEEE PES PSOPE, Chair of IEEE PES PSOPE Award Subcommittee, Vice Chair of PES Distribution System Operation and Planning Subcommittee, and Vice Chair of PES Task Force on Advances in Natural Disaster Mitigation Methods. He is an Associate Editor for IEEE TRANSACTIONS ON POWER SYSTEMS, IEEE TRANSACTIONS ON SMART GRID, IEEE OPEN ACCESS JOURNAL OF POWER AND ENERGY, IEEE POWER ENGINEERING LETTERS, and *IET Smart Grid*.