

Received November 28, 2020, accepted December 5, 2020, date of publication December 9, 2020, date of current version December 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3043630

Power Outage Estimation: The Study of Revenue-Led Top Affected States of U.S.

NAVEED TAIMOOR¹, IKRAMULLAH KHOSA¹, (Member, IEEE),
MUHAMMAD JAWAD¹, JAHANZEB AKHTAR¹, IMRAN GHOUSS¹,
MUHAMMAD BILAL QURESHI², (Member, IEEE),
ALI R. ANSARI³, AND RAHEEL NAWAZ⁴

¹Department of Electrical and Computer Engineering, COMSATS University Islamabad, Lahore Campus, Lahore 54000, Pakistan

²Department of Electrical Engineering, COMSATS University Islamabad, Abbottabad Campus, Abbottabad 22000, Pakistan

³Department of Mathematics and Natural Sciences, Gulf University of Science and Technology, Mishref Campus, Kuwait City 32093, Kuwait

⁴Department of Operations, Technology, Events and Hospitality Management, Manchester Metropolitan University (MMU), Manchester M156BH, U.K.

Corresponding author: Ikramullah Khosa (ikramullahkhosa@cuilahore.edu.pk)

ABSTRACT The electric power systems are becoming smart as well as complex with every passing year, especially in response to the changing environmental conditions. Resilience of power generation and transmission infrastructure is important to avoid power outages, ensure robust service, and to achieve sustained economic benefits. In this study, we employ a two-stage model to estimate the power outage in terms of its intensity as well as the duration. We identify the top three potentially critical states of United States of America, not merely based on duration of the power outage, but by also incorporating outage related revenue loss. In the proposed model, the first stage classifies the intensity of the outage event while the second stage predicts the duration of the outage itself. Moreover, the key predictors are characterized and their association with outage duration is illustrated. We use a comprehensive and publicly available dataset, which provides the information related to historical power outage events, such as electricity usage patterns, climatological annotations, socio-economic indicators, and land-use data. Our rigorous analysis and results suggest that the power outage interval is the function of several parameters, such as climatological condition, economic indicators as well as the time of the year. The proposed study can help the regulatory authorities taking appropriate decisions for long term economic paybacks. It can also help disaster management authorities to take risk-informed resilient decisions for system safety.

INDEX TERMS Prediction, power outages, natural disasters, classification, support vector machine, random forest.

I. INTRODUCTION

Energy requirement of the world is increasing day by day; therefore, several sources of energy are being explored to discover a sustainable and efficient way of producing energy. Modern world demands continuous electric power supply since a minor disruption in the supply highly affects all the sectors of society, such as security, health, education, industry, and socio-economic activities. In the United States (U.S.), the nature of power infrastructure is complex and extensive where the power demand is rising consistently [1]. Such extensive power infrastructure is more prone to failure

due to natural disasters whom occurrence is also on a rise over the last two decades due to significant changes in climate conditions [2], [3], [7]. The disaster events damage the electric power infrastructure and cause prolonged power outages [4]–[6]. The potential reasons for occurrence of such events include (a) increase in sea-water temperature leading to floods, (b) increase in global high mean temperature causing heat waves and droughts, and (c) change in air pressure and humidity producing thunderstorms, hurricanes, wind storms, typhoons, and tropical cyclones [8]. For instance, in 2012 Derecho storm affected approximately 4.2 million people in eleven states of the U.S. The southwest blackout in 2011 was prolonged to 12 hours affecting 2.7 million people. The super storm sandy in 2012 came in urban areas of north-eastern

The associate editor coordinating the review of this manuscript and approving it for publication was Shen Yin.

states of the U.S. and impacted more than 8.5 million people by causing interruptions in electric power supply [9], [10].

Among all the affected by electricity disruption, one of the key sectors is the utility industry. The power blackout undoubtedly affects the economic progress of the region in terms of heavy revenue losses. The U.S. economy has borne the loss US\$ 20-55 billion due to severe-weather related power outages from 2003 to 2012 [11]. Over the period 37 years: 1981-2017, the U.S. has borne the economic loss worth of U.S. \$219 billion caused by climate change and resulting severe weather disasters [9]. In the year 2017 alone, sixteen disasters occurred across the U.S. and put a dent of billion dollars on economy [9]. These statistics indicate the alarming economic risk associated with power outages. In literature, several models have been presented to forecast power outage duration, mainly due to hurricanes and storm-induced natural disaster [12]–[14]. The machine learning techniques are widely being used in these studies due to their suitability for inter-disciplinary research [15]–[19]. The financial loss due to power outages is also linked with demand of consumer sectors: residential, commercial, and industrial where the demand varies with the seasons. Electricity consumption has been enhanced by 75% in the U.S. since 1980. The deployment of heating and cooling systems is the main reason behind this increase in residential and commercial demands [20]. A wide range of literature exists to understand the relationship of the electricity demand and the climate sensitivity. A study presents the sensitivity of the electricity and natural gas demand to climate for the top eight energy concentrated states of the U.S. comprising California, Louisiana, Texas, Florida, Washington, Illinois, Ohio, and New York [21]. Another study presents the analysis of the U.S. level energy demand by using the two-stage least square methodology. Three energy sectors: residential, commercial, and industrial were considered for the analysis of relationship between energy demand and price elasticity [22]. A similar study presents impact of climate sensitivity on the residential and commercial energy demand [21]. In the context of climate change, most of the studies discuss the impact of temperature variation on the energy demand, excluding other factors such as wind speed, precipitation, and humidity. In a recent study, a multi-hazard approach was presented for risk assessment due to power outages [3]. Prediction of power outage duration was performed using data of outage event associated with natural disasters with prolonged durations.

In most of previous studies, the focus has been on analyzing the impact of a specific type of event on the power infrastructure. For instance Nateghi estimated the power outage duration for the events occurred due to hurricanes Dennis, Katrina and Ivan in the central gulf coast state [6]. Another study presents the damage estimate to the communication networks infrastructure during the hurricane Sandy [23]. Ali proposed the solution to minimize potential damages to power systems for the upcoming hurricane [24]. Similarly there are a number of recent studies where the research focus is to assess the impact on electric power systems specifically

during or after a hurricane has occurred [25]–[30]. There are several studies where the research motivation is to assess the impact of geomagnetic storms on the electric power system [31]–[33]. In context with natural disasters, there are several studies in the literature which present the impact assessment of other kind of severe-weather related events such as thunderstorm, rainstorms and heavy winds [34]–[38]. Most of the existing studies present the analysis and prediction of outages in context to specific kind of event which trigger the power outage as discussed earlier. The work is limited where the impact of outage event is assessed due to all possible reasons to trigger the outage event at the same time, and moreover exclusively targeted for a specific region. The economic risk assessment has been investigated based on the public feedback in terms of money they are willing to pay (WTP) for uninterrupted power supply [39], [40]. In such studies, the economic loss has been estimated based on the amount which the public is ready to pay (WTP) for continuous power supply. However, there is a gap in analyzing the loss of revenue an electric supply company bears in case of an outage event occurs. To address these shortcomings, we focus our research in finding the regions bearing large revenue losses due to power outages events. We use the electricity consumption patterns in the United States (U.S.) and its price with the outage duration to formulate the revenue loss on which we lay our foundation of research in identifying the potential states.

Previously, outage risk is widely assessed under natural disaster events mostly such as hurricanes, thunderstorms, heavy wind, winter storms etc. Apart from weather-related disasters, there are other reasons too for power outages in the U.S., such as equipment failure, fuel supply emergency, and public appeal. Such accidental or manual shut down of the power system also have significant impact. We use the historical data of power outage events associated not only with natural disasters but other reason too including equipment failure, fuel supply emergency, intentional attacks, system operability disruption, islanding and public appeal.

Therefore, we predict the outages which may be triggered not by a specific reason but the whole bunch of possible reasons at a time.

We explore the data of 50 states of the U.S., and then to focus our search, we identify and analyze the three most affected states for evaluation of results and discussion. As discussed earlier, this selection is not merely based on the total duration of the outages rather on the base of revenue loss bared by the electric supplying companies. Then we predict the outage intensity as well as the outage duration using a two stage machine learning model. The contributions of our research are as follows:

- We perform the exploratory data analysis to identify and put the foundation of our research.
- We calculate the revenue loss of electric supply companies for all the individual states of the U.S. and identify the top three economically affected states of the U.S. due to power outages.

TABLE 1. Data categories and respective sources.

S No	Data Type	Data Sources
1	State level population (yearly)	U.S. Census Bureau
2	Electricity consumption patterns	National Oceanic and Administrative Administration (NOAA)
3	Type of the year	NOAA
4	Climate regions (by state)	NOAA
5	Percentage customers served	US Energy Information Administration (EIA 826)
6	Economic characteristics	Bureau of Economic Analysis
7	Percentage of urban and rural areas	U.S. Census Bureau
8	Percentage of land and water mass	U.S. Census Bureau
9	State-level climate and water data	NOAA's National Climate Data Centre (NCDC)

- We develop a two-stage model where in the first stage we classify the intensity of power outage event among three target categories: minor, moderate and extreme. In the second stage, we predict of the duration of the outage specifically for extreme category.
- We characterize the key parameters for individual states which contribute most towards efficient predictions as well as illustrate their relationship with the outage duration.

The rest of the paper is organized as follows: Exploratory data analysis is presented in section 2; Section 3 includes methodology; results and discussion are presented in section 4 and conclusion is added in section 5.

II. EXPLORATORY DATA ANALYSIS

The historical data related to the power outages in the U.S. is publicly available for the period January 2000 to July 2016 [41]. The dataset contains 1534 events of power outages that triggered due to seven different reasons. The dataset comprehensively includes the information of nine different categories, and subsequent multiple indicators within each category. The nature and the sources of the data are provided in Table 1.

For the sake of explanation, foundation of this research, and to find the hidden statistics, we present the exploratory data analysis. Figure 1 shows the percentage distribution of the occurrence of seven types of events that caused power outages. Major percentage of power outage events is associated with severe weather category (50%), followed by international attacks (27%).

If we analyze the occurrence of power outage events over the years 2000-2016 with respect to interval of the year or specifically month-wise, the frequency of different events is shown in Fig. 2. It can be observed again that majority of power outage events are triggered due to weather related events. Moreover these events occur more frequently during the the peak summer and winter seasons. Statistically speaking, in summer, from May to October, an average of 76 events per month is recorded; and during the winter from December to February, on average 64 events per month is recorded.

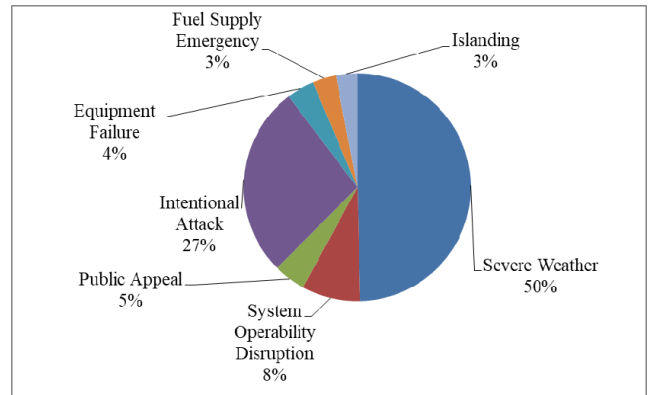


FIGURE 1. Percentage distribution of power outages caused by seven types of events occurred in the U.S. during Jan 2000 – July 2016.

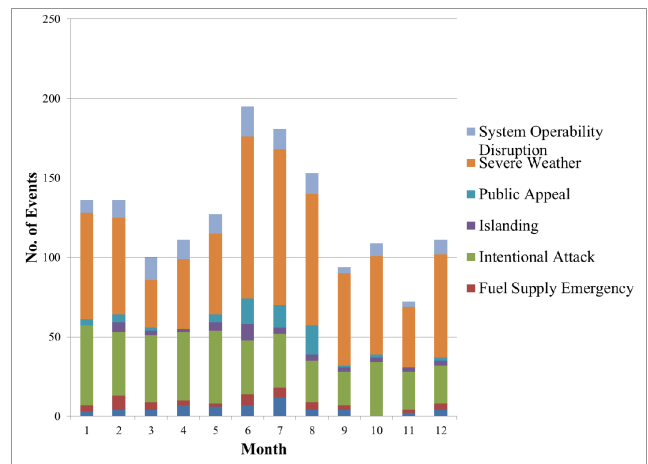


FIGURE 2. Percentage Monthly distribution of frequency of events causing major power outages in the U.S. during 2000-2016.

To analyze and quantify the revenue loss due to power outage events, we calculate the total loss in U.S. dollars using all the outage event observations in the data for the period 2000-2016.

Initially, the total income in billion dollars per minute is computed using the sale of electricity and electricity prices at the time of power outage event. The income is then multiplied with the outage duration (in minutes) to calculate the total loss during the event. We define this loss as TLO (Total Loss calculated using Observed outage duration). The Eqn. (1) is used to calculate the income at the time of event and Eqn. (2) is used to calculate the total loss during an event.

$$\begin{aligned}
 & \text{Total income (Bn. US\$ / min)} \\
 &= \text{total consump (MWh)} \times \text{unit pr (cents/KWh)} \\
 & \times \frac{1000}{100} \times 60 \times 10^{-9} \tag{1}
 \end{aligned}$$

$$\begin{aligned}
 & \text{Total loss (US\$ bn)} \\
 &= \text{total income (US\$bn/ min)} \times \text{outage duration (min)} \tag{2}
 \end{aligned}$$

where the unit of electricity consumption is MWh (Mega Watt Hour), and of price is Cents per KWh. Figure 3 depicts the

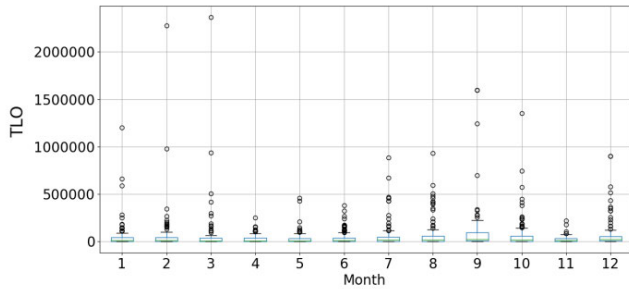


FIGURE 3. Month-wise distribution of total loss (TLO) US\$ billion in overall U.S.

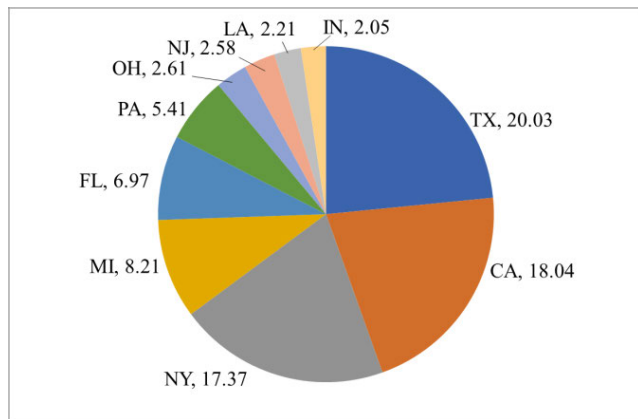


FIGURE 4. Distribution of total fiscal loss in terms of percentage for top 10 affected states of U.S.

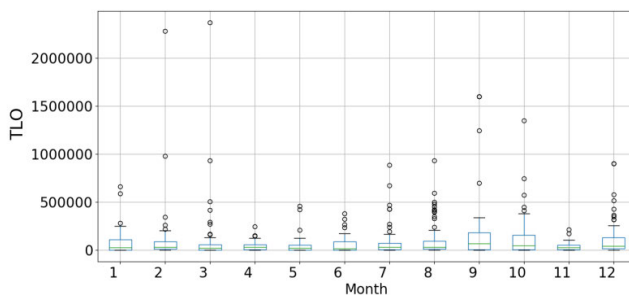


FIGURE 5. Month-wise distribution of total loss (US\$ billion) in top 10 affected states of the U.S.

box plot of month-wise financial loss distribution across the U.S. The loss is higher in the months of extreme weather due to higher number of severe-weather related disaster events.

While calculating the revenue loss for individual states, it was observed that 85% of the total loss belongs to the top 10 affected states including Texas, California, New York, Michigan, Florida, Pennsylvania, Ohio, New Jersey, Louisiana and Indiana in descending order respectively. The percentage distribution of financial loss in aforementioned states is depicted in Fig 4. Moreover, it can be observed that Texas, California, and New York are the top three affected states while their accumulated financial loss is more than 55% of total financial loss. The month-wise distribution of the financial loss in top 10 states is shown in Fig. 5.

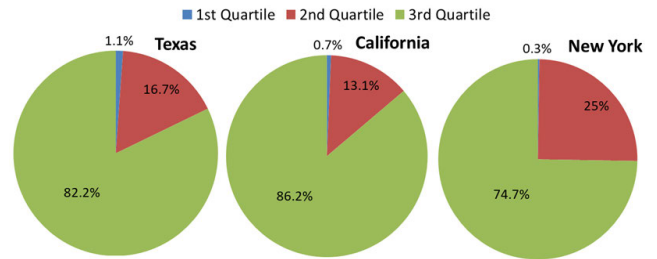


FIGURE 6. Percentage of financial loss associated with quartile range of power outage duration.

The exploratory data analysis concludes that more than 55% share of the total revenue (generated by electricity sale across U.S.) belongs to only three states, which are badly affected due to the power outage events. Therefore, the data of Texas, California and New York is considered for further analysis, and results evaluation. For exploratory data analysis, we considered the loss calculation to identify the most vulnerable states in context of economic impact. Since the revenue loss directly depends on outage duration (see Eq. 1 and 2), therefore, we shall consider the outage duration as our output or the response variable for our model.

A. RESPONSE VARIABLE NORMALIZATION

When we observed the duration of the power outage events, we found that most of the power outage events occurred for a shorter period like less than 48 hours, while fewer events occurred for longer time durations such as more than a week time. However, the impact of the longer duration outage events is more damaging for the economic pace compared to shorter duration outage events. To illustrate this impact, we divide the events as per their outage duration in quartiles. Fig. 6 shows the percentage of financial loss for top three affected states individually, calculated according to the outage duration associated with different quartiles where inter-quartile range is defined as 2nd quartile. It is observed that the major portion of loss is related to the 3rd quartile range. Therefore, it is important to predict the duration of an outage event which is likely to be in 3rd quartile i.e. the prolonged blackout. Hence, instead of using all the data for the prediction of outage duration, we shall consider the observations belonging only to third quartile range of outage duration for the prediction of outage duration.

Figure 7 shows the kernel density distribution of the power outage duration for the third quartile in the top three affected states. It can be observed that the New York (NY) faced more prolonged power outages comparatively than the other two states. The longest outages also occurred in the NY state, where California witnessed more outages with shortest duration among the three states. It can be observed that the distributions in general are skewed towards left explaining many events with short outage duration, and the long tail for all three states shows that fewer power outage events have been with prolonged duration. This left-skewed distribution

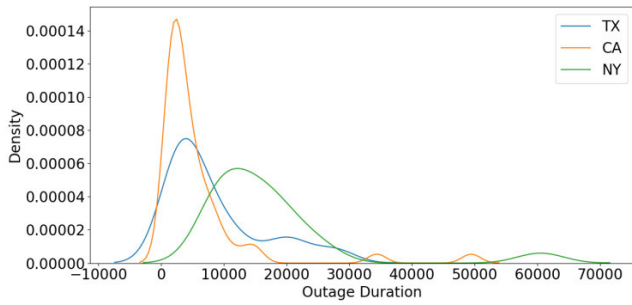


FIGURE 7. Kernel distribution of power outage duration belonging to 3rd quartile for Texas, California and New York.

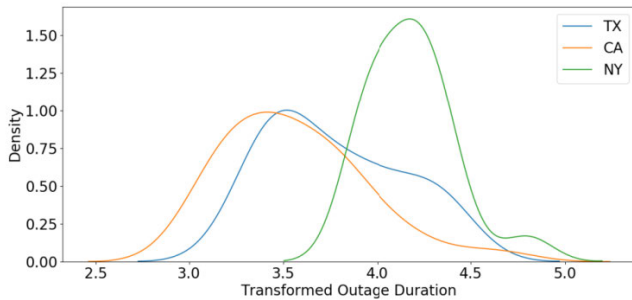


FIGURE 8. Log transformed kernel distribution of power outage duration belonging to 3rd quartile for Texas, California and New York.

indicates that the classifier may get influence of shorter duration outages and therefore may result as the biased prediction toward them. To avoid this situation and to normalize it, we instead use the logarithmic scale as follows:

$$\begin{aligned} & \text{Transformed outage duration} \\ &= \log(\text{observed outage duration} + 1) \quad (3) \end{aligned}$$

The benefit of this transformation is that it improves the data distribution and consequently the classifier provides fair prediction on the transformed data. For the top three states, the impact of the log transformation on kernel distribution is shown in Fig. 8. The visualization indicates the improved distribution.

In the Table 2, we show the statistics of the power outage duration (response variable) using the data of 3rd quartile only for each of the top three states of the U.S. The observations are calculated both with original outage duration (min) and the transformed outages duration. These statistics reveal that the longest mean outage duration occurred for the New York state while the minimum mean outage duration is observed in California. The same we indicated earlier while discussing kernel distributions of these states. The statistics for log transformed observations are also shown and the similar pattern can be observed there as well.

B. FINAL FEATURE SELECTION

In the dataset, each outage event is described by 50 features. Since the existence of linearly correlated features is likely as in case of any statistically described dataset, it produces multi-collinearity in the data. The high multi-collinearity can divert the impact of features on the response variable.

TABLE 2. Statistics of 3rd quartile of outage duration for top three affected states.

Stats of outage duration	Statistical Parameters					
	Mean	Median	Std Dev.	Min	Max	
Texas	Outage duration (min)	8489.5	5595.0	7718.1	2220.0	27698.0
	Transformed outage duration	3.8	3.7	0.4	3.3	4.4
California	Outage duration (min)	5766.3	2997.5	8299.9	1054.0	49427.0
	Transformed outage duration	3.6	3.5	0.4	3.0	4.7
New York	Outage duration (min)	17131.6	14519.5	12006.5	8221.0	60480.0
	Transformed outage duration	4.2	4.2	0.2	3.9	4.8

To reduce the multi-collinearity, we normalize each feature and then select those features that have VIF index (variance inflation factor) less than 4 [42]. The VIF is a statistical measure to assess the severity of multi-collinearity in the least square regression analysis. We performed the procedure individually for each of our model stage i.e. classification and regression. Initially for each of the three states, all the data is normalized, and features are selected for first (classification) stage based on VIF index. In the second (regression) stage, the data associated only with the 3rd quartile of the outage duration is used and normalized, and the procedure is repeated for features selection. Since the first stage of the model classifies the outage event as one of three quartile’s events, therefore all the data is used and the event to be classified as falling in one of the three quartiles. The final features for classification stage as well as for regression stage are presented in Table 3. The features with zero standard deviation are removed from the data. The feature selection procedure was done individually for each stage.

III. METHODOLOGY

In this section, we describe the two-stage classification-regression model for the power outage intensity categorization and its interval prediction. In the first stage, the intensity of outage duration is classified as minor, moderate, and extreme based on quartile division of outage duration: Minor: 1st quartile range, Moderate: inter-quartile range, Extreme: 3rd quartile range. There are many classifiers having been used in the existing literature, however, two most popular and extensively used among them in last two decades are Support Vector Machines and Artificial Neural Network. The SVM is known as large margin classifier which fits $c - 1$ hyperplanes for c classes in the data [43]. It has the advantage of capturing highly complexity in the data in the presence of outliers. The ANN is the other common classifier inspired by the biological brain, being used widely across all research areas [44]–[47]. It has the capability to adopt non-linear

TABLE 3. Final set of parameters for classification and regression stages.

Feature	Final sets of features					
	Classification			Regression		
	CA	NY	TX	CA	NY	TX
Anomaly Level	√	√	√	√	√	√
Cause Category (e.g. severe weather)	√	√	√	√	√	√
Cause Category Detail (e.g. heavy wind)	√	√	√	√	√	√
Climate Category	√	√	√	√	√	√
Commercial Customer Percentage	-	-	-	-	-	√
Number of Commercial Customers	√	-	√	-	-	-
Commercial Sale Percentage	-	√	√	-	√	-
Commercial Price	√	-	-	-	-	-
Commercial sale (units)	-	√	-	-	√	√
Hurricane Name	-	√	√	-	√	√
Industrial Customer Percentage	-	√	-	-	-	-
Number of Industrial Customers	-	-	-	√	-	-
Industrial Price	-	√	√	√	√	-
Industrial Sale (units)	√	-	√	-	√	√
Month	√	√	√	√	√	√
NERC Region	√	√	√	-	-	√
Percentage Real GSP Change	√	√	√	√	-	-
Percentage Real GSP Relative (state to US GSP ratio)	√	-	-	√	-	-
Percentage Real GSP of USA	-	√	-	-	-	√
Residential Customer Percentage	-	-	-	√	-	-
Number of Residential Customers	-	√	-	-	-	-
Residential Percentage Sale	√	√	√	√	-	-
Residential Price	√	√	-	-	√	√
Total Sales	-	-	-	√	-	-
Utility Sector Real GSP	√	-	√	√	-	-
Year	√	-	-	√	-	-

relationship between input and target. We used both the SVM and the ANN classifier for category classification of outage data. In the second stage, the duration of power outage event is predicted using Random Forest (RF) model [48]. The Random Forest algorithm has the advantage that it can very well capture the non-linear structure within the data while being robust to noise and outliers. The decision-tree in contrast is the low bias, high variance technique. The RF averages the predictions across all the trees which overcome the problem of high variance. The RF is simple to implement without the need of fine tuning and generally provides good prediction accuracy. Besides, it provides the parameter ranking as well based on their contribution toward prediction.

For this purpose, the data belonging to the 3rd quartile of outage duration was used. As presented in previous section, a high percentage (more than 75% for all the states) of financial loss is related to the outage events with longer durations (3rd quartile). Therefore, we considered the events of 3rd quartile only for the prediction of prolonged-duration outage event. If we consider the total data for prediction of outage duration, the large number of short duration events will bias the model, and consequently there will be high error rate for long duration outage events. As a result, the model will not be fruitful for the critical purpose that is to efficiently predict for the longer-duration power outage events. In addition to the prediction of outage duration, the important parameters

for prediction are also identified. We also present the partial dependence of response variable with the identified key predictors. The overall flow diagram of the proposed method is shown in Fig. 9.

A. SUPPORT VECTOR MACHINES

The SVM algorithm is a well-known machine learning algorithm and widely used in the applications of pattern recognition and classification [43]. The SVM is reliable and can be optimized for the data that is noisy and have outliers. It can be trained on simple as well as highly complex data. The SVM uses $n - 1$ dimensions out of n -dimensional space as a hyper plane to maximize the distance between the hyper plane and the nearest training sample. The larger the distance, the lower is the generalization error. The SVM algorithm can construct both linear and non-linear hyper planes (boundaries). For non-linear boundaries, the SVM algorithm uses different kernel approaches to capture the non-linear structure inherited in higher feature dimensions. The SVM generalization error can be reduced by different tuning parameters such as the complexity cost (to avoid the over fitting) and the kernel function types (e.g. radial, Gaussian, polynomial or exponential) to generate appropriate non-linear boundaries for classification of different classes.

B. ARTIFICIAL NEURAL NETWORK

An Artificial neural network (ANN) is a data driven machine learning algorithm which consists of interconnected nodes called the neurons. It was inspired by the working of biological central nervous system. Other than input and output layer, it may have one or more hidden layers. The features are fed to the input layer of the network which are forward to the next layer and then to the output layer where the network makes the prediction. The ANN learns by watching the true labels for the data samples and updating its own weights via back propagation. The weights of the network are optimized after several forward-backward passes by minimizing the difference between the actual output and the predicted output. The ANN has been extensively used for last two decades by the researchers across wide application areas [44]–[47].

C. RANDOM FOREST

We leverage random forest (RF) model developed by Breiman [48] for the prediction of extreme outage duration caused by the power outages. The RF is a tree-based ensemble model that can understand the nonlinear structure of the data and is robust to outliers and noise. Moreover, the RF is a non-parametric method and so it does not consider a particular distribution and performs efficient prediction for heterogeneous data. The procedure to develop an RF algorithm is as follows:

- 1) Create a training set by selecting N re-sampled structures of data, keeping the remaining samples for validation purpose (error estimate) of the tree.
- 2) Use training data to fit a regression tree by choosing m variables to split on.

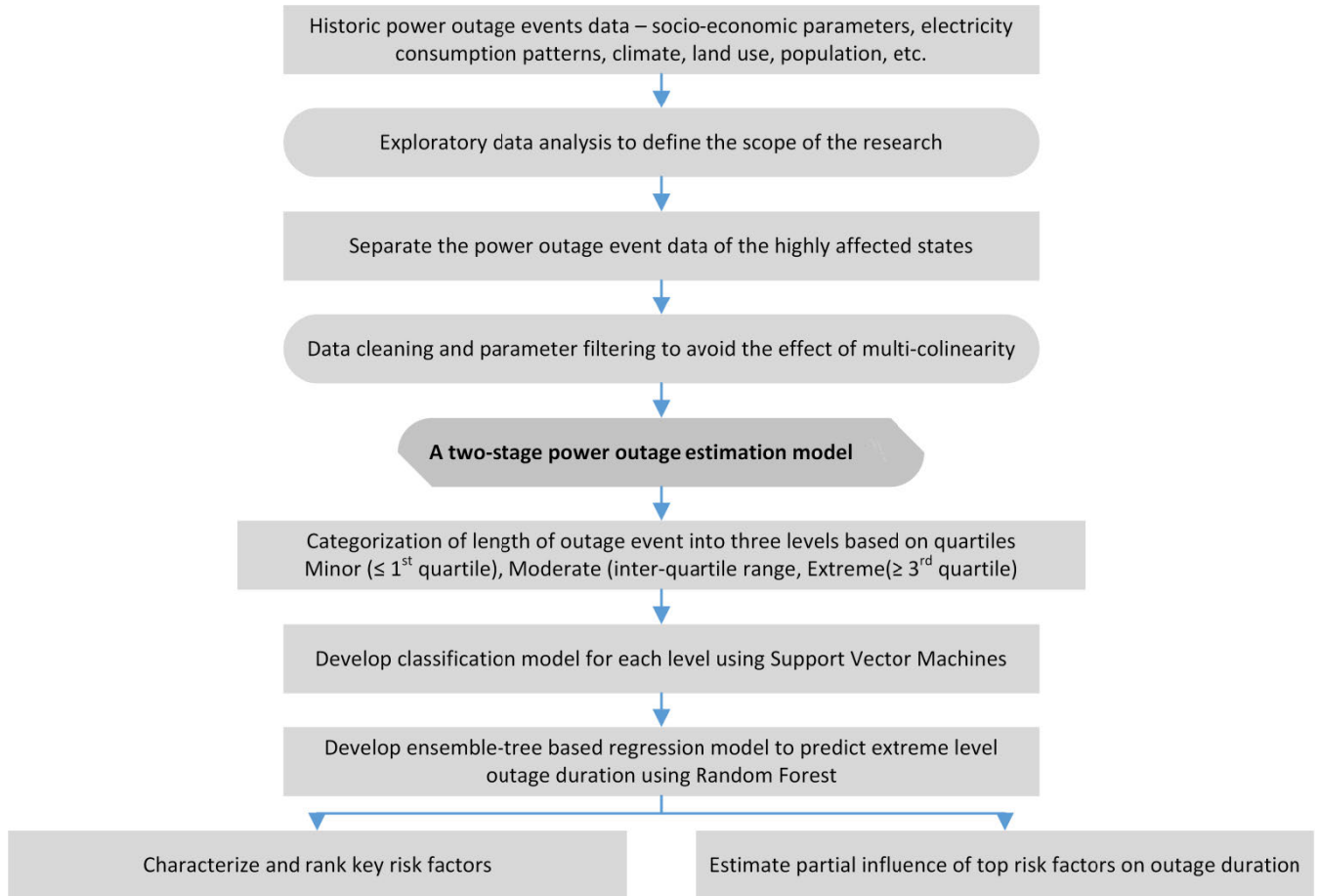


FIGURE 9. Flow diagram of the proposed two-stage model for power outage estimation.

- 3) Select the optimal splitting value, letting the tree grow completely.
- 4) Calculate the prediction error using residual data.

Repeat steps 1-4 K -times to mature K number of trees.

Random forest captures the general structure of the data with high sensitivity to outliers, leading to high-variance case. Since random subsets of data are used to fit each individual regression tree, and the split value for the tree is also random, averaging the estimates of all trees reduces the overall high-variance impact and therefore improves the accuracy. These characteristics make it an ideal algorithm to fit the complex as well as noisy data. The RF also estimates the importance of key features for the prediction of the response variable. It ranks the predictors on the base of their contribution towards the response variable's prediction.

D. PARTIAL DEPENDENCY PLOT

To get an insight of the influence of individual features on the response variable, we used Partial Dependency Plot (PDP). In nonparametric models, the PDP helps to understand the influence of individual feature variable on the output variable keeping all other factors constant [49]. The PDP is a fine way to represent the marginal effect on the response variable by changing one input feature variable at a time while keeping

all others feature variables unchanged. The PDP is mathematically computed as;

$$Y_s(X_s) = \frac{1}{K} \sum_{i=1}^K Y(X_s, x_{iR}) \quad (4)$$

where the Y_s is the output variable, X_s is the covariate for which the PDP is being estimated, and x_{iR} are all the covariates except X_s .

IV. RESULTS AND DISCUSSION

This section includes the results of the two-stage model as follows:

Stage 1: This is the classification stage where the outage duration is classified in terms of the severity levels as minor, moderate or extreme. The severity describes the outage event falling on the quartile scale such that 1st quartile range corresponds to minor category, inter-quartile range to moderate category and the 3rd quartile range corresponding to extreme outage event. For classification, we employed two algorithms: support vector machines and artificial neural network.

Stage 2: In this stage, the random forest-based model predicts the duration of power outage event. The model also

TABLE 4. SVM Classification results for outage intensity (minor – 1st quartile, moderate - interquartile, extreme - 3rd quartile).

State	Optimized Tuning Parameters	Support Vector Machine Results				Optimized Tuning Parameters	Artificial Neural Network Results			
		Predicted	Actual				Predicted	Actual		
Texas	Cost=32 Gamma=0.02041	Minor	26	2	0	Hidden layer size=8 Learning rate = 0.01	Minor	24	2	0
		Moderate	4	57	1		Moderate	5	58	2
		Extreme	1	1	28		Extreme	1	2	27
		Misclassification error: 7.5%					Misclassification error:9.9%			
California	Cost=256 Gamma=0.02	Minor	48	6	0	Hidden layer size=6 Learning rate = 0.01	Minor	46	4	0
		Moderate	2	89	2		Moderate	4	91	5
		Extreme	0	3	48		Extreme	0	3	45
		Misclassification error: 6.5%					Misclassification error: 8.1%			
New York	Cost=16 Gamma=0.02273	Minor	17	1	0	Hidden layer size=14 Learning rate = 0.01	Minor	15	0	1
		Moderate	1	33	5		Moderate	3	34	7
		Extreme	0	0	13		Extreme	0	0	10
		Misclassification error: 10%					Misclassification error: 15.7%			

identifies the key parameters along with their importance for efficient prediction of outage duration.

After identification and ranking of key parameters, their partial dependence plots are presented to illustrate their influence on the response variable.

It is worth mentioning again that the SVM classification model is developed using the data of all the outage events for intensity categorization of the event, while RF based model uses the data of extreme-level events (3rd quartile events) only, for the prediction of time duration of outage event.

A. CLASSIFICATION MODEL

For classification, we computed the results using SVM and ANN models.

1) SVM CLASSIFICATION MODEL

As discussed, the SVM classification model was developed for three intensity levels of the outage duration by dividing the data using quartile-based division: minor intensity outage (1st quartile range), moderate (2nd quartile range) and the extreme (3rd quartile range) intensity outage. The model is tested using three different kernels: linear, polynomial and radial basis function (RBF) by varying the tuning parameters of the kernels through the grid search of values while minimizing the cost. Among the kernels, RBF kernel was selected based on lowest misclassification error on the validation set. For training and results evaluation, the data was splitted with 70:30 ratio for training and validation respectively. The results of the classification model are presented in terms of confusion matrix in the Table 4. The optimized value for kernel parameter of regularization ‘Gamma’ and the

other corresponding achieved cost are also given in Table 4. The diagonal values in the confusion matrix show the true positives (correct classification). The overall misclassification error of three categories is also presented. It can be observed that for Texas and California, minor and moderate categories are mainly confused with each other leading to misclassification error. Similarly, for the state of New York, moderate and extreme categories were confused with each other. The maximum misclassification error is recorded for the NY state i.e. 10 % where the total number of outage events of NY is 71 only. The reason is the largest variance in the observations of the outage duration for NY power outage events. On the contrary, there are 210 outage events in total for the state of California while the misclassification rate is 6.5 % only. This is due to the relatively smaller variance and smoother distribution of outage duration data.

2) ANN CLASSIFICATION MODEL

We developed the ANN model for classification of the intensity levels of outage duration as in the case of SVM. We employed the feed forward multilayer neural network with one hidden layer. For network training, the scaled conjugate back propagation algorithm was used. For each of the individual state the finalized features (shown in Table 3) were fed to the input layer while output layer contains three target classes. For network architecture estimation, data was splitted into training, validation and test sets as 70%, 15% and 15% from each class respectively and at random. The network was initialized by random weights and the optimized network architecture was obtained approximated

TABLE 5. Random forest prediction results for extreme level outage duration: RF Regression model vs. mean-only model.

State	Hyper parameters For RF model	Model	MAE	MAPE
Texas	N_estimators=1600	Mean-only	0.31	0.08
	Min_samples_split = 5	RF	0.16	0.04
	Min_sample_leaf = 1	RF performance improvement over Mean-only model	48.38%	51.85%
	Max_depth = 70			
California	N_estimators = 800	Mean-only	0.30	0.08
	Min_samples_split = 10	RF	0.20	0.06
	Min_sample_leaf = 4	RF performance improvement over Mean-only model	33.33%	27.77%
	Max_depth = 100			
New York	N_estimators = 2000	Mean-only	0.16	0.04
	Min_samples_split = 2	RF	0.11	0.03
	Min_sample_leaf = 2	RF performance Improvement Over Mean-only model	31.25%	25%
	Max_depth = 20			

using validation accuracy. Finally the network results were recorded and an average of five results is presented.

The ANN classification results are summarized in Table 4 in terms of confusion matrix as well as the misclassification error percentage. In comparison with the results obtained by the SVM model, the ANN produced higher error rate in each of the individual state case. On average the error rate of ANN is 3.2% higher than the SVM considering the results of all three states. It can be observed that the However, the accuracy of the ANN is better than SVM for the moderate class. It is evident from the confusion matrix that the ANN showed higher accuracy for moderate category than the SVM results for each of the individual state. The moderate category contains the data of inter-quartile range, and there is less chance of presence of outliers. This is the range where ANN outperformed the SVM model. The confusion among classes leading to misclassification is almost similar both in SVM and ANN; however, the error rate of ANN is higher. The results suggest that the SVM can perform better in case of complex and noisy data while ANN can be a better option where the data is filtered and free from outliers.

B. RANDOM FOREST REGRESSION MODEL

As discussed in the Section 2.1, the response variable is normalized using log transformation to normalize and improve the data distribution. We use the transformed observations of outage duration to train our RF model as well for prediction. For a comparison, we produce the prediction results using mean-only model. The results of RF model for outage duration are evaluated in terms of Mean

Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). The MAPE rate recorded for both RF model and mean-only model is less than 10% as shown in Tab. 5.

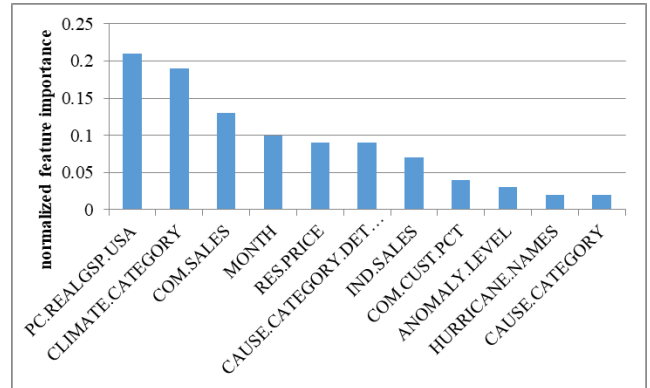


FIGURE 10. Feature importance ranking for the state of Texas.

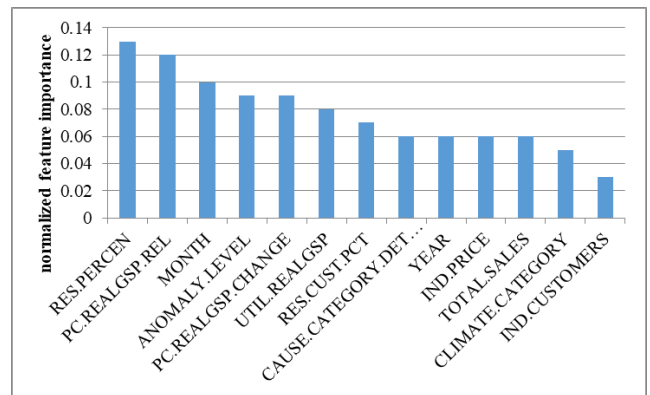


FIGURE 11. Feature importance ranking for the state of California.

However, the RF model produced better than the mean-only model in each of the individual state.

1) KEY PREDICTORS RANKING

We characterize the input features in terms of their contribution towards prediction using RF model. The random forest algorithm calculates the importance of features and therefore we can rank the features w.r.t. their significance. The importance of parameters is computed and normalized on the scale between 0 and 1. Figure 10, 11 and 12 show the most important parameters along with their importance toward prediction of outage duration for the state of Texas, California, and New York, respectively. Features with normalized value of their importance equal to 0.02 or higher are shown in Fig. 10-12.

Figure 10 shows the top 11 important features for the prediction of outage duration in Texas. The per capita real GSP (gross state product) of the U.S. is the most important features toward prediction with normalized importance value of 0.21. For the California, there are 13 important features where the top predictor variable identified as percentage of residential population of the state having importance equal to 0.13 as shown in Fig. 11. Figure 12 shows the feature ranking results for the New York state with percentage of real GSP being most important feature. An interesting observation is

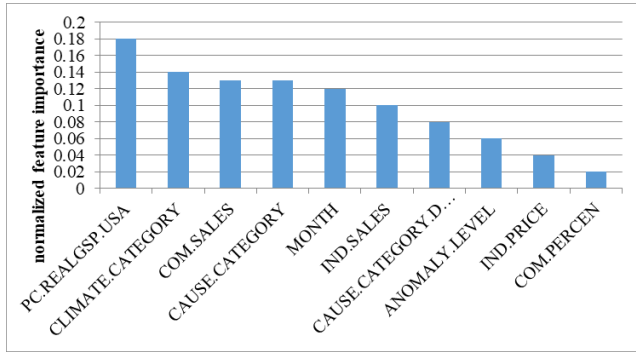


FIGURE 12. Feature importance ranking for the state of New York.

that both for the state of Texas and New York, the top three important predictors are the same. The Per capita GSP of the state is an indicator for economic health of the region, and the trends of commercial sales of electricity show the growth of utility industry. Being among top three important parameters for the state of Texas and New York, these features show the significance of economic indicators for prediction of outage duration. Apart from economic indicators, climate category is the third important parameter for power outage estimation in these states. The climate category provides the weather-related information which leads to the hint for the estimation of upcoming natural disaster in the region. Hence economy and environment related evidence was identified as most important characteristics for power outage estimation for the states of Texas as well as New York. While characterizing the key predictors for California, the month of the year identified within three top features other than economic indicators. It is also important since the severity of the season (summer or winter) lies within this indicator. On the contrary, the month of the year is ranked among top five features for power outage estimation in each of the three states.

In the next Sub-Section, we discuss the relationship between most important parameters which are identified for each state and the power outage duration. For this purpose, we only consider top three parameters and present the PDP to illustrate the dependency analysis.

C. TOP PREDICTORS IN OUTAGE INTERVAL PREDICTION FOR THE STATE OF TEXAS

1) PER CAPITA REAL GSP OF U.S

The per capita real GSP of the U.S. is the most important parameter in the prediction of outage duration for Texas. As shown in Fig. 13, we may observe with visual inspection of the PDP that keeping all features constant, the predicted outage duration (minutes) is larger while the per capita real GSP of U.S. being in its first quartile. However, in its inter-quartile range, the predicted outage duration reduces significantly, and then it remains low in the 3rd quartile range. The PDP reveals that high value of the per capita real GSP of U.S. will indicate that the outage events may occur with minimal outage duration, and hence the input and output parameters indicate an inverse relation more generally.

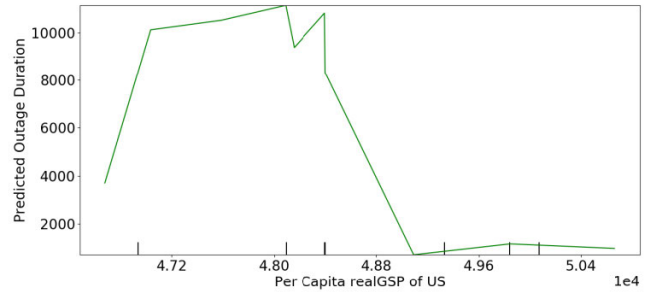


FIGURE 13. Partial dependency plot for Per Capita Real GSP of U.S. on outage duration.

2) CLIMATE CATEGORY

It was observed from the exploratory data analysis that the major reason which caused prolonged outage duration was incurrence of severe weather-related disasters. Mostly these severe weather disasters occur due to hurricanes and storms. The most important feature related to weather is identified as climate category here. While looking at climate category against power outage duration, it was observed that 76% of time duration of overall power outage events, the climate condition was normal. This is an interesting observation, which indicates that the description of state-level (global) climate category feature is less relevant. The average outage duration recorded under normal climate condition is 12,545 minutes, while the average outage duration in the 3rd quartile (where prolonged outage events occurred) is 8,489 minutes. This indicates that a lot of outage events occurred while climate category was normal. Therefore, it is important to note that even under the normal climate of the state of Texas, the prolonged outage events might occur. The normal climate condition is again a subjective assessment and one may naively select the threshold for the climate category, which may be a possibility leading to contradictive results now. The predicted outage duration under different climate conditions is shown in Fig. 14.

3) COMMERCIAL SALES

The impact of commercial sales on the predicted outage duration is in inverse relation in most of the 1st and 3rd quartile ranges of sales outage duration as shown in Fig. 15. However, there is an increase in the outage duration within the interquartile range of commercial sales of electricity. It was observed that in the months when commercial sector sale of electricity was low, prolonged power outage events were occurred due to different disasters in Texas. The PDP reveals that the elevated commercial sales will indicate the outage events with smaller outage durations. The commercial sale parameter covers a wide range of outage duration as opposed to the feature of per capita real GSP of U.S.

4) CORRELATION BETWEEN KEY PREDICTORS AND RESPONSE VARIABLE

The pair plot shows the relationship between input feature variable and the response variable (outage duration in our

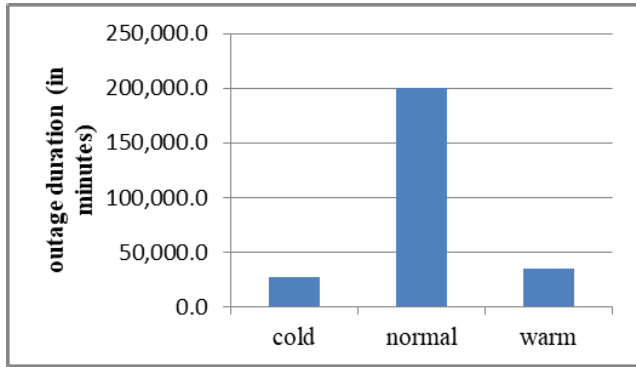


FIGURE 14. Influence of climate category for prediction of outage duration.

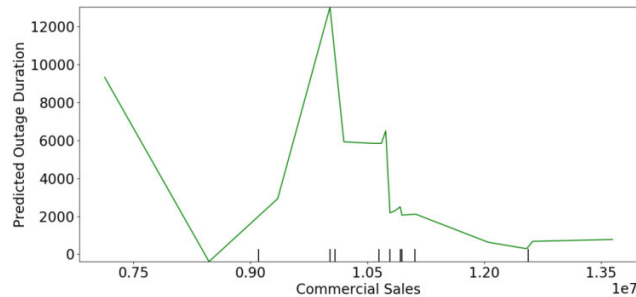


FIGURE 15. Partial dependency plot for commercial sales on outage duration.

case) while providing the index of Pearson correlation coefficient between them. We present the pair plot for the top five important features in Fig. 16. A cluster of data is made where individual entry is sub-defined based on climate category as shown in Fig. 16. From Jan 2000 till July 2016, we observe

that there is a high probability of power outages in normal climate condition as compared to cold and warm climate. The average power outage under normal climate condition is observed as 13,000 minutes. The scatter plots in Fig. 16 shows the maximum observations in normal climate (in blue color). During the normal climate, power outage events occurred for both short and long duration. Therefore, the normal climate category is more important for analysis in prediction as compared to the other categories. Besides, the average outage duration under the warm and cold climate condition is recorded as 3,500 and 4,500 minutes respectively.

The density plot of residential price shows that the variance in the price is high in normal climate followed by cold and warm, respectively. Moreover, we observe from the density plot for the parameter of month that there is high probability for the climate to be warm during the months of March to June. Considering the parameter of commercial sales, it is observable that variance is high in context with warm climate, which indicates the abrupt changes in the demand of electricity during warm climate conditions.

The Pearson correlation index of per capita real GSP and commercial sales with the outage duration is negative and this can be observed in Fig. 13 and 15 that their relationship in inverse in general.

D. TOP PREDICTORS IN OUTAGE INTERVAL PREDICTION FOR THE STATE OF CALIFORNIA

1) PERCENTAGE OF ELECTRICITY CONSUMPTION OF RESIDENTIAL SECTOR

The influence of percentage of residential electricity consumption on the outage duration can be observed from

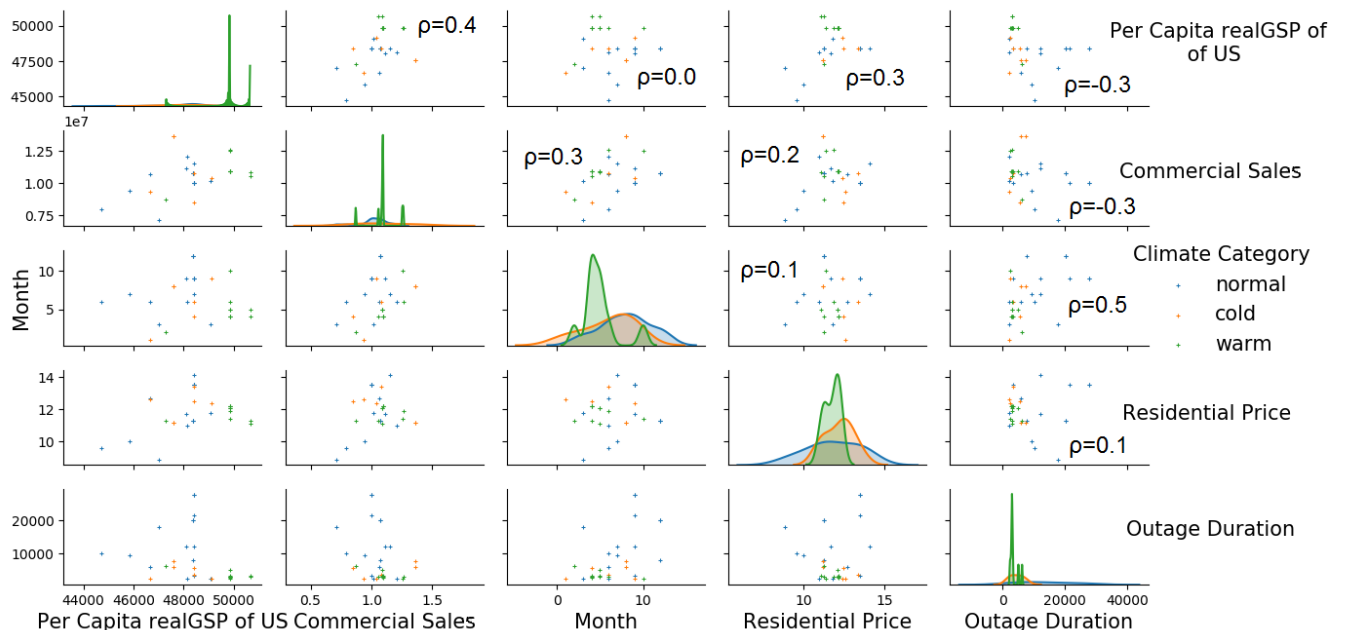


FIGURE 16. Pair-plot of top five predictors for outage interval prediction in the state of Texas.

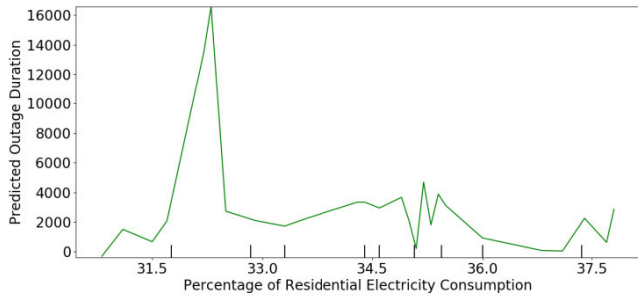


FIGURE 17. Partial dependency plot for percentage of residential electricity consumption on outage duration.

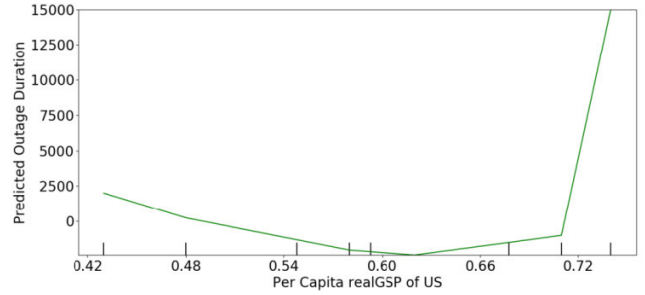


FIGURE 20. Partial dependency plot for per capita real GDP of U.S. on outage duration.

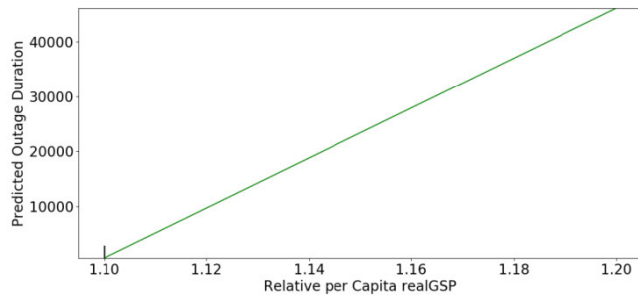


FIGURE 18. Partial dependency plot for relative per capita real GDP on outage duration.

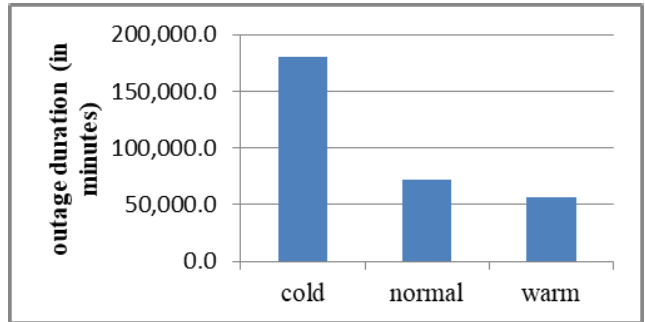


FIGURE 21. Influence of climate category for prediction of outage duration.

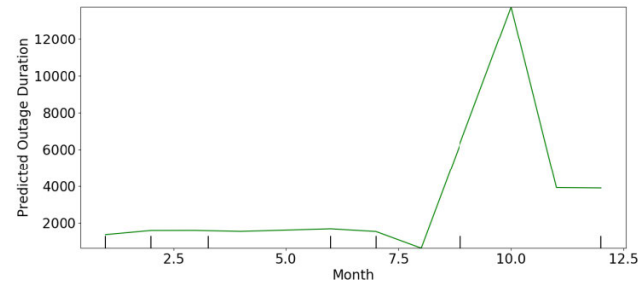


FIGURE 19. Partial dependency plot for month (interval of the year) on outage duration.

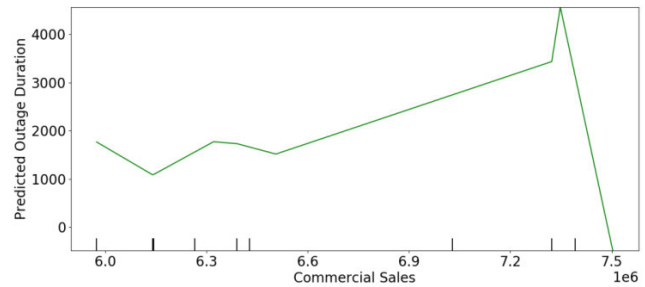


FIGURE 22. Partial dependency plot for commercial sales of electricity on outage duration.

Fig. 17. The predicted outage duration approaches to its peak in the 1st quartile range of residential electricity consumption. In the interquartile and the 3rd quartile, the predicted outage duration remains low except in the end where it increases once again for a short range of extreme electricity consumption. Therefore, except extreme cases of residential electricity consumption pattern, the outage duration is to be predicted as small.

2) RELATIVE PER CAPITA REAL GSP

The PDP shown in Fig. 18 illustrates a linear relationship between relative Per Capita real GSP (PC real GSP REL) and the outage duration. The reason for this is that the dataset from Year 2000 to Year 2016 for the California state has only two (rounded) values of PC real GSP REL: 1.1 and 1.2. Although the input feature has low variance, yet the predicted outage duration covers the largest range in comparison with previous indicators.

3) MONTH OF THE YEAR

Figure 19 shows the partial dependence of time duration of the year on the outage duration. It can be observed that the dependence scale is almost constant from January to June, and in later months from August to October, the predicted outage duration continuously and significantly increases. Overall, the influence of time interval of the year on outage duration is moderate but the predicted outage duration increases in Aug-Nov due to extreme weather condition developed by storms and hurricanes in this region which trigger the power outages for longer durations.

E. TOP PREDICTORS IN OUTAGE INTERVAL PREDICTION THE STATE OF NEW YORK

1) PER CAPITA REAL GSP OF U.S

Likewise, Texas, the per capita real GSP of U.S. is the most influential feature toward prediction of outage duration for the state of New York also. The partial dependency plot

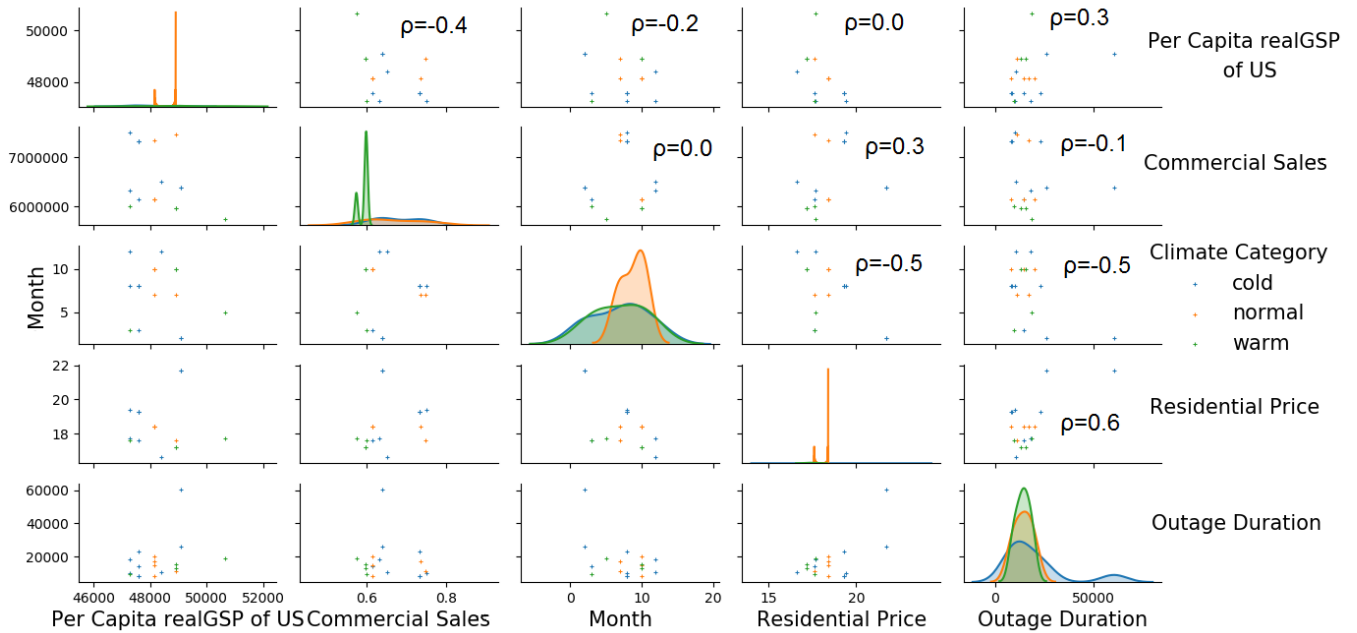


FIGURE 23. Pair-plot of top five predictors for outage interval prediction in the state of New York.

shown in Fig. 20 illustrates that the predicted outage duration decreases up to mean value of per capita real GSP of US and after that it keeps on increasing. At the end of 3rd quartile, the outage duration increases significantly. Statistically speaking, every 100 dollars increment in the per capita real GSP of U.S. causes the mean outage duration decreased to 3Hrs in the below average range (of per capita real GSP of US). Then it causes an increase of 10 Hrs (on average) of outage duration in the above average range of the per capita real GSP of U.S. as shown in Fig. 20.

2) CLIMATE CATEGORY

The climate category is the 2nd most important variable in prediction of outage duration for New York state. Climate category has 3 kinds: cold, normal and warm. In New York, the longest (cumulative) power outage duration occurred in cold climate followed by normal and warm as shown in Fig. 21. Therefore, more chances of disasters and power outages prevail in cold climate. In New York, high probability of cold climate is during the period December to February. While under normal and warm climate categories, most of the power outages were yet occurred due to severe weather-related disasters.

3) COMMERCIAL SALES

The partial dependency plot shown in Fig. 22 illustrates the influence of commercial sector sales of electricity towards the duration of outage of a power outage event. It can be observed that the predicted outage duration is low in the 1st quartile of commercial sales while it keeps on increasing in the inter-quartile range, and finally it reduces to minimum when the sale is at its peak. A linear relation is observed

here except at the time of maximum sales where the outage duration to be predicted minimum.

4) CORRELATION BETWEEN KEY PREDICTORS AND RESPONSE VARIABLE

For the state of New York, we show a pair plot in Fig. 23 for the top 5 most important features used in prediction and make clusters while representing the data based on climate categories. The pair plot provides the Pearson correlation coefficient values, scatter plots and the density plots between the input variables as well as between input variable and the output variable. As we observed previously that in New York, the longer duration power outage events occur during the cold climate followed by normal and then warm, the same can be observed from the scatter plots in last row of Fig. 23. The density plots of outage duration show that average power outage duration in case of normal and warm climate is less as compared to the cold climate. The variance is much higher for cold climate as compared to other two categories. In case of residential price, the impact is higher under normal climate but for the commercial sales, severe impact is under warm climate. The analysis of month from the density plot shows that the normal climate range exist from July to October and warm and cold climate remains in extreme summer and extreme winter season months, however, it may be witnessed throughout the year.

V. CONCLUSION

In this article, we analyzed the data of power outage events triggered due to different reasons: from public appeal and fuel emergency to extreme weather induced natural disasters. The exploratory data analysis unveiled that a huge amount of

revenue is lost due to power infrastructure damage, mainly (but not entirely) caused by natural disasters. The analysis revealed that 55% of this revenue pertains to only three of the fifty states of the U.S: Texas, California and New York, and therefore we focus on these three states. Initially, a power outage event is classified for its severity based on its time duration. Secondly, the duration of the power outage event is predicted, considering only the prolonged outage events.

Considering the top three key predictors, economic indicators like the per capita GSP of the U.S. and the commercial sector sales of electricity is identified to be the critical parameters for prediction of outage duration.

Moreover, time interval of the year is also recognized as an important factor. However, if we look down in the list, the outage duration is not a function of economic factors only; instead it is a function of several parameters. This is because of diverse observations in data for different states such as price of electricity, demand of electricity, economic stability of the region, population density, industrial and commercial activity etc. The weather and climate indicators of the region are also important and identified among top predictors. Contrary to common understanding, the analysis shows that there can be odd relationship between severity of weather-related disaster and the duration of power outage depending on the regional development as well as the infrastructural strength of power transmission and distribution system. In New York, the major percentage of overall outage duration was observed while climate was cold, whereas in Texas the larger portion of outage duration was recorded under normal climate conditions. In context with climate condition, the time of the year or specifically the month of the year is also related. It plays an important role for efficient prediction of occurrence of power outage event as well as the estimation of the duration of the outage.

We presented a cascaded model for the estimation of power outages for the top three states of the U.S. selected in the context of revenue-loss. This model can be used for any of the state of U.S. for power outage assessment due to the power outage event, but for one state at a time. It might be beneficial for utility companies to find better investment avenues. Finally, the results of our study can be employed as a decision-support tool for the authorities to design risk informed resilient power infrastructure, and to formulate policies accordingly.

REFERENCES

- [1] American Society of Civil Engineers. (2013). *Report Card for America's Infrastructure*. Accessed: Dec. 10, 2018. [Online]. Available: <http://2013.infrastructurereportcard.org/>
- [2] S. Mukherjee and R. Nateghi, "Climate sensitivity of end-use electricity consumption in the built environment: An application to the state of Florida, United States," *Energy*, vol. 128, pp. 688–700, Jun. 2017.
- [3] S. Mukherjee, R. Nateghi, and M. Hastak, "A multi-hazard approach to assess severe weather-induced major power outage risks in the U.S.," *Rel. Eng. Syst. Saf.*, vol. 175, pp. 283–305, Jul. 2018.
- [4] S. Mukherjee, "Towards a resilient grid: A risk-based decision analysis incorporating the impacts of severe weather-induced power outages," Ph.D. dissertation, School Civil Eng., Purdue Univ., West Lafayette, IN, USA, 2017.
- [5] A. Staid, S. D. Guikema, R. Nateghi, S. M. Quiring, and M. Z. Gao, "Simulation of tropical cyclone impacts to the U.S. power system under climate change scenarios," *Climatic Change*, vol. 127, nos. 3–4, pp. 535–546, Dec. 2014.
- [6] R. Nateghi, S. D. Guikema, and S. M. Quiring, "Forecasting hurricane-induced power outage durations," *Natural Hazards*, vol. 74, no. 3, pp. 1795–1811, Dec. 2014.
- [7] Climate Central. (2014). *Blackout: Extreme Weather, Climate Change and Power Outages*. Princeton, NJ, USA. Accessed: Oct. 18, 2018. [Online]. Available: <http://www.ourenergypolicy.org/wp-content/uploads/2014/04/climate-central.pdf>
- [8] K. T. Ingram, K. Dow, L. Carter, J. Anderson, and E. K. Sommer, *Climate of the Southeast United States: Variability, Change, Impacts, and Vulnerability*. Washington, DC, USA: Island Press, 2013, pp. 1–342.
- [9] NOAA. *Billion-Dollar Weather and Climate Disasters: Table of Events 2017*. Accessed: Jan. 2, 2019. [Online]. Available: <https://www.ncdc.noaa.gov/billions/>
- [10] (Apr. 12, 2016). *Electric Choice, 9 of the Worst Power Outages in United States History*. Accessed: Dec. 3, 2018. [Online]. Available: <https://www.electricchoice.com/blog/worst-power-outages-in-united-states-history/>
- [11] R. J. Campbell, *Weather-Related Power Outages and Electric System Resiliency*. Washington, DC, USA: Congressional Research Service, Library of Congress, Aug. 2012.
- [12] R. E. Brown, S. Gupta, R. D. Christie, S. S. Venkata, and R. Fletcher, "Distribution system reliability assessment: Momentary interruptions and storms," *IEEE Trans. Power Del.*, vol. 12, no. 4, pp. 1569–1575, Oct. 1997.
- [13] 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.
- [14] R. Nateghi, S. D. Guikema, and S. M. Quiring, "Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes," *Risk Anal.*, vol. 31, no. 12, pp. 1897–1906, Dec. 2011.
- [15] R. Yunus, O. Arif, H. Afzal, M. F. Amjad, H. Abbas, H. N. Bokhari, S. T. Haider, N. Zafar, and R. Nawaz, "A framework to estimate the nutritional value of food in real time using deep learning techniques," *IEEE Access*, vol. 7, pp. 2643–2652, 2019.
- [16] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Comput. Hum. Behav.*, vol. 104, Mar. 2020, Art. no. 106189.
- [17] F. Riaz, S. Naeem, R. Nawaz, and M. Coimbra, "Active contours based segmentation and lesion periphery analysis for characterization of skin lesions in dermoscopy images," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 489–500, Mar. 2019.
- [18] I. Safder, S.-U. Hassan, A. Visvizi, T. Noraset, R. Nawaz, and S. Tuarob, "Deep learning-based extraction of algorithmic metadata in full-text scholarly documents," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102269.
- [19] M. Jawad, A. Rafique, I. Khosa, I. Ghous, J. Akhtar, and S. M. Ali, "Improving disturbance storm time index prediction using linear and nonlinear parametric models: A comprehensive analysis," *IEEE Trans. Plasma Sci.*, vol. 47, no. 2, pp. 1429–1444, Feb. 2019.
- [20] N. L. Miller, K. Hayhoe, J. Jin, and M. Auffhammer, "Climate, extreme heat, and electricity demand in California," *J. Appl. Meteorol. Climatol.*, vol. 47, no. 6, pp. 1834–1844, Jun. 2008.
- [21] D. J. Sailor, "Relating residential and commercial sector electricity loads to climate—Evaluating state level sensitivities and vulnerabilities," *Energy*, vol. 26, no. 7, pp. 645–657, Jul. 2001.
- [22] M. A. Badri, "Analysis of demand for electricity in the United States," *Energy*, vol. 17, no. 7, pp. 725–733, 1992.
- [23] A. Kwasinski, "Lessons from field damage assessments about communication networks power supply and infrastructure performance during natural disasters with a focus on Hurricane Sandy," in *Proc. FCC Workshop Netw. Resiliency*, 2013, pp. 1–36.
- [24] A. Arab, A. Khodaei, S. K. Khator, K. Ding, V. A. Emesih, and Z. Han, "Stochastic pre-hurricane restoration planning for electric power systems infrastructure," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 1046–1054, Mar. 2015.
- [25] R. Nateghi, S. Guikema, and S. M. Quiring, "Power outage estimation for tropical cyclones: Improved accuracy with simpler models," *Risk Anal.*, vol. 34, no. 6, pp. 1069–1078, Jun. 2014.

- [26] R. Nateghi, "Multi-dimensional infrastructure resilience modeling: An application to hurricane-prone electric power distribution systems," *IEEE Access*, vol. 6, pp. 13478–13489, 2018.
- [27] Y. Sang, J. Xue, M. Sahraei-Ardakani, and G. Ou, "An integrated preventive operation framework for power systems during hurricanes," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3245–3255, Sep. 2020.
- [28] K. Feng, M. Ouyang, and N. Lin, "Modeling the impact of hurricane induced power outage in a changing climate," in *Proc. AGUFM*, Dec. 2019, p. GH13C-1065.
- [29] H. T. Nguyen, J. W. Muhs, and M. Parvania, "Assessing impacts of energy storage on resilience of distribution systems against hurricanes," *J. Modern Power Syst. Clean Energy*, vol. 7, no. 4, pp. 731–740, Jul. 2019.
- [30] A. Kwasinski, F. Andrade, M. J. Castro-Sitiriche, and E. O'Neill-Carrillo, "Hurricane maria effects on Puerto Rico electric power infrastructure," *IEEE Power Energy Technol. Syst. J.*, vol. 6, no. 1, pp. 85–94, Mar. 2019.
- [31] M. Švanda, D. Mourenas, K. Žertová, and T. Výbošťoková, "Immediate and delayed responses of power lines and transformers in the czech electric power grid to geomagnetic storms," *J. Space Weather Space Climate*, vol. 10, p. 26, May 2020.
- [32] M. Guo, X. Kang, Y. Xu, C. Zhao, and Y. Zhang, "Preliminary reliability evaluation for power system considering geomagnetic data," *J. Eng.*, vol. 2019, no. 16, pp. 1429–1433, Mar. 2019.
- [33] X. Kang, Y. Zhang, X. Zhang, Y. Xu, X. Li, and Y. Yuan, "Risk assessment of cascading failure in power grid based on brittleness theory of complex system and entropy theory under geomagnetic storm conditions," in *Proc. 8th Renew. Power Gener. Conf.*, 2019, pp. 1–7.
- [34] E. Kabir, S. D. Guikema, and S. M. Quiring, "Predicting thunderstorm-induced power outages to support utility restoration," *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 4370–4381, Nov. 2019.
- [35] B. A. Alpay, D. Wanik, P. Watson, D. Cerrai, G. Liang, and E. Anagnostou, "Dynamic modeling of power outages caused by thunderstorms," *Forecasting*, vol. 2, no. 2, pp. 151–162, May 2020.
- [36] H. Sabouhi, A. Doroudi, M. Fotuhi-Firuzabad, and M. Bashiri, "Electrical power system resilience assessment: A comprehensive approach," *IEEE Syst. J.*, vol. 14, no. 2, pp. 2643–2652, Jun. 2020.
- [37] C. Rosende, E. Sauma, and G. P. Harrison, "Effect of climate change on wind speed and its impact on optimal power system expansion planning: The case of Chile," *Energy Econ.*, vol. 80, pp. 434–451, May 2019.
- [38] D. Cerrai, M. Koukoulou, P. Watson, and E. N. Anagnostou, "Outage prediction models for snow and ice storms," *Sustain. Energy, Grids Netw.*, vol. 21, Mar. 2020, Art. no. 100294.
- [39] J. Reichl, "Power outage cost evaluation: Reasoning, methods and an application," *J. Sci. Res. Rep.*, vol. 2, no. 1, pp. 249–276, Jan. 2013.
- [40] C. Jed, K. Moeltner, J. Reichl, and M. Schmidthaler, "Effect of global warming on willingness to pay for uninterrupted electricity supply in European nations," *Nature Energy*, vol. 3, no. 1, p. 37, 2018.
- [41] *Power Outage Data*. Accessed: Dec. 3, 2018. [Online]. Available: <https://engineering.purdue.edu/LASCI/researchdata/outages/outagerisks/>
- [42] J. Gareth, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 8th ed. New York, NY, USA: Springer, 2017.
- [43] J. Gareth, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. New York, NY, USA: Springer, 2013.
- [44] J. Kuruville and K. Gunavathi, "Lung cancer classification using neural networks for CT images," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 202–209, Jan. 2014.
- [45] I. Khosa and E. Pasero, "Artificial neural network classifier for quality inspection of nuts," in *Proc. Int. Conf. Robot. Emerg. Allied Technol. Eng. (iCREATE)*, Apr. 2014, pp. 103–108.
- [46] J. D. Sitton, Y. Zeinali, and B. A. Story, "Rapid soil classification using artificial neural networks for use in constructing compressed Earth blocks," *Construct. Building Mater.*, vol. 138, pp. 214–221, May 2017.
- [47] S. Behkani, S. M. Zain, M. Gholami, and M. F. A. Khir, "Classification of cow milk using artificial neural network developed from the spectral data of single-and three-detector spectrophotometers," *Food Chem.*, vol. 294, pp. 309–315, Oct. 2019.
- [48] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [49] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.



NAVEED TAIMOOR received the master's degree in electrical engineering from the Department of Electrical and Computer Engineering, COMSATS University Islamabad, Lahore Campus, in 2019. His research interests include artificial intelligence, machine learning, and data analysis.



IKRAMULLAH KHOSA (Member, IEEE) received the Ph.D. degree in electronics and telecommunications from the Politecnico di Torino, Italy, in 2015. He is currently working as an Assistant Professor with the Electrical Engineering Department, COMSATS University Islamabad, Lahore Campus. His research interests include artificial intelligence, data analysis, machine learning, and pattern recognition.



MUHAMMAD JAWAD received the Ph.D. degree from North Dakota State University, Fargo, ND, USA, in 2015. He is currently an Assistant Professor with COMSATS University Islamabad, Lahore Campus, Pakistan. His research interests include optimization theory, prediction algorithms, machine learning, deep learning, and smart grids optimization for energy efficiency.



JAHANZEB AKHTAR received the Ph.D. degree in electrical engineering (advance control systems) for his research on efficient particle tracking in high-energy physics experiments in 2015. He moved to the Department of Electrical and Computer Engineering, COMSATS University Islamabad, Lahore Campus, as an Assistant Professor, in 2016. His research interests include control theory applications, state estimation (bad data and outlier detection and treatment), Kalman filtering, and non-linear tracking.



IMRAN GHOUS received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2011 and 2013, respectively, and the Ph.D. degree in control science and engineering from the Nanjing University of Science and Technology, China, in July 2016. He has been serving as an Assistant Professor with the Department of Electrical Engineering, COMSATS University Islamabad, Lahore Campus, Pakistan.



MUHAMMAD BILAL QURESHI (Member, IEEE) received the Ph.D. degree from NDSU, USA, in 2017. He is currently an Assistant Professor with COMSATS University Islamabad, Abbottabad Campus, Pakistan. His research interests include control systems, optimization, and biomedical engineering.



ALI R. ANSARI has an academic with over 28 years-experience in teaching and research in the area of applied mathematics. He was the Dean of the College of Arts and Sciences, Gulf University for Science and Technology, Kuwait, and served for ten years. He is currently a Professor of applied mathematics. He has more than 100 publications in the area of applied mathematics.



RAHEEL NAWAZ is currently the Director of Digital Technology Solutions and a Reader in analytics and digital education with Manchester Metropolitan University (MMU). He has founded and/or headed several research units specializing in artificial intelligence, data science, digital transformations, digital education, and apprenticeships in higher education. He has led on numerous funded research projects in U.K., EU, South Asia, and Middle East. He has held adjunct or honorary positions with several research, higher education, and policy organizations, both in U.K., and overseas. He regularly makes media appearances and speaks on a range of topics, especially artificial intelligence and higher education. Before becoming a full-time academic, he served in various senior leadership positions in the private higher and further education sector; and was an Army Officer before that.

...