Outage Cause Detection in Power Distribution Systems based on Data Mining

Mohammad Sadegh Bashkari, Ashkan Sami, and Mohammad Rastegar, *Member IEEE*

***Abstract*— Realizing the factors involved in power system outages can be effective in reliability improvement. This paper analyzes the distribution power network outage data to find dominant factors in occurring vegetation-, animal-, and equipment-related outages. After their integration, real outage, weather, and load as input data are used to extract associated features. In this paper, visualization techniques are initially utilized to show the impact of features on the outage occurrence and then association rule mining is used to find factors correlated with each outage type as well as each other. Association rules are mined using Apriori technique, considering the chi-square and lift index as the measures of interestingness. The outage analyses are also performed for each equipment separately to find the associated rules. The results showing the effectiveness and validity of the proposed method to identify factors connected with outage occurrences can be used for future planning and the operation schedule of distribution power networks.**

***Index Terms*— Artificial intelligence, Association Rules, Data mining, Power System, Reliability.**

I. Introduction

VAST amount of data collected in a system can make opportunities to gain a better understanding of the system. Regular methods for data analysis such as traditional statistical analysis may not be applicable due to the large amounts of data. Recently, data mining algorithms have presented promising artificial intelligence tools to discover useful and actionable knowledge from the available data. Among the most well-known data mining analyses, one can refer to visualization, classification, clustering, association rule mining, and outlier detection [1], [2].

Power networks are also faced with various and huge data such as voltages, current, voltage phases, active and reactive power, component failures, and system outages. Accordingly, data mining methods can be implemented to discover valuable knowledge from power network data. For example, meter and consumer power usage data were classified in [3]–[6] to detect electricity theft and in [7] and [8] to classify power quality disturbances. Visualization techniques were also used in [9]–[11] to monitor power network states [9] and to identify temporal and spatial clusters of faults in Scottish distribution and transmission systems [10]. A set of clustering methods was further utilized in [12] to optimally design the topology of a wind farm.

To evaluate the power system reliability, many researchers have focused on the mining of recorded outage data in power networks [13]-[22]. For example, to build an accurate model to predict outage components in the power grid. Researchers in [13] used the support vector machine algorithm. Since most failures occur in the distribution systems [14], researchers have mostly paid attention to distribution system outages. For example, [15] and [16], respectively, used artificial neural network and support vector machine classification algorithm to estimate the fault location in distribution systems. In [17], outage duration was predicted and refined over time through the artificial neural network. Feature selection methods were utilized in [18] both to find a subset of useful factors in different fault causes in different regions and to identify the cause of outages. Classification methods were also implemented in [19] to predict the weather-related outages in distribution systems. In [20], the number of growth-related and weather-related vegetation outages were estimated for the future through regression methods. In the same vein, logistic regression was also utilized in [21] to examine the factors involved in vegetation related outages. To predict the failure rate of line sections in power distribution networks, the researchers in [22] employed and evaluated four data mining models namely, linear regression, exponential regression, linear multivariable regression and neural network. The animal-related incidents on overhead distribution feeders were studied in [23] and the main causes of outages in distribution systems such as bad weather conditions, vegetation incidents, animal-related incidents, and equipment failures were totally classified in [24]. In [25], equipment related outages were initially examined and 12 important factors in equipment-related failure were identified through a feature selection algorithm. Afterward, three classification methods were used to relate the outages to equipment or non-equipment related outages. However, fault causes in different equipment are not separately investigated. Although investigated factors might have different influences on each equipment, these differing influences would be lost in the results. An association rule mining algorithm has been also implemented in [24] to determine the effects of various environmental factors on the outage causes. Despite its significant impacts on the operation of equipment, distribution system loads have not been properly addressed in the literature of fault cause analysis. Association rule mining is the most appropriate method to

Manuscript submitted October 20, 2019; revised December 31, 2019.

M. S. Bashkari, A. Sami (Corresponding Author) are with Department of Computer Science and Engineering and IT; and M. Rastegar (corresponding Author) is with Department of Power and Control, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran. Tel: +98-71-36133569 (E-mail: s.bashkari@shirazu.ac.ir, sami@shirazu.ac.ir, mohammadrastegar@shirazu.ac.ir).

determine the important factors in each outage type, since other methods were not able to investigate the importance of co-occurrence of the factors better than association rule mining. Although feature selection methods could yield high correlation between factors, some non-correlated factors could cause a condition that might consequently leads to an outage.

This paper utilizes association rule mining to discover frequent patterns in vegetation, animal, and equipment-related outages in a real power distribution system. To this end, in addition to faults and weather data that have been previously analyzed in the literature [18], [19], we aim to investigate load data. In this regard, various features such as temperature, humidity, hourly load, date, and wind speed are selected. Then, fault datasets are labeled by their causes, (i.e., vegetation-, animal-, and equipment-related) and these labeled data are transformed from a multi-class dataset to one-versus-all datasets. Several 2-D visualization techniques are also used to get a better understanding of the inner relations of datasets. Finally, after a preprocessing phase, which includes under-sampling procedure association rules are mined in each fault cause class against other causes. To inspect associations between outage causes and factors affecting them, we should count each of their co-occurrences in the data Since this procedure could be time-consuming due to a large number of records and parameters, the Apriori algorithm which avoids generating unnecessary candidate rules [26] is used to improve performance. Besides, in the equipment-related fault category, various equipment such as transformers, poles, jumpers, overhead lines and cables, cable terminations, and insulators are separately examined, to identify the most frequent patterns of failure occurrence in each equipment. In addition to support, confidence, and lift indexes, the current study also proposes chi-square as a statistical technique [27] to determine the strength of the relation between outage and its factors. In conclusion, utilizing the load dataset in the rule mining methods for fault cause analysis, mining fault cause rules for each equipment of the distribution systems, and testing the interestingness of the rules using both lift and chi-square tests are the main contributions of the current study to the related literature.

The proposed method is examined on a real system to show its effectiveness and applicability in real-world data, obtainable in many distribution networks.

II. Data and Features

The real power distribution network considered in this paper consists of 64 substations, 491 feeders with a total length of 10981 Km. Table I shows some characteristics of this network.

TABLE I
Test Distribution Network Characteristics

Components	Measurement Unit	Amount
Substations	Count	64
Feeders	Count	491
Feeder Length	Length (km)	10981
Distribution Transformers	Count	26155
Underground Posts	Count	1593
Overhead Posts	Count	23207

Three datasets are integrated and used in this paper. The first and foremost dataset, called the outage dataset throughout this paper, is a five-year outage record collected from this network from March 2013 to March 2018. The second dataset contains weather conditions for the same period which is collected from [28] and named weather dataset here. Hourly loads collected from the same distribution network for a year from March 2017 to March 2018 comprise the third dataset.

TABLE II
Datasets Used in This Research

Dataset	Number of Features	Number of Records	Important Features	Period
Outage Dataset	18	31079	Outage Cause, outage date and time, feeder number	2013 to 2018
Weather Conditions	29	14565	Time and date, temperature, humidity, pressure, observed weather, precipitation, wind speed	2013 to 2018
Load	4	477548	Hourly Load of each substation	2017 to 2018

When an outage occurs in this network, a record is added to the outage dataset. Various features which are recorded in the dataset are regional features showing the location of the outage, temporal features determining the date and time of the outage, restoration time and outage duration, main reasons for the outage, faulty feeder number, and the amount of energy not supplied

Based on the existing literature, bad weather conditions may cause some outages. Since weather attributes are not available in the main dataset, these features are collected from other sources. Each weather record represents the weather conditions of a three-hour period. In this weather dataset, there are various features including time and date, temperature, humidity, pressure, observed weather, precipitation, and wind speed.

Another important feature that may cause outages is the substation load. The load profile as the hourly metered electrical energy consumptions of each substation is used in this research. The load data are available only from March 2017 to March 2018. Table II shows the characteristics of the three datasets used in this research.

III. Proposed Framework for association rule mining

The objective of this paper is to discover meaningful patterns in the outage occurrences. To this end, several data sources are initially chosen to study their probable effects on outage occurrences. While weather and outage datasets are commonly addressed in the related literature, the impact of load dataset on the outage occurrences has not been touched upon. To eliminate unreliable values from datasets, we used a cleaning process on each dataset prior to their joining and integration. Since these datasets are gathered from different sources, they should be integrated based on a common field. Although the datasets contain time and date values, the time stamps are different; therefore, an alignment procedure is performed. To prepare the data for investigation, continuous features are discretized in several ways such as equal count and equal width bins. In this paper, the equal count method is used, so that the final discretized feature would have a similar population among categories. After the initial preprocessing phase, visualization is used to gain an understanding of the inner relations between different variables. In the final phase, reclassification is done several times to generate multiple datasets each of which is used to investigate one fault cause. In addition to general categories of vegetation, animal and equipment investigated in the literature [20]-[25], each equipment is separately investigated in this paper. Each dataset is then balanced to have the same number of samples from each class. Finally, the association rules are extracted to discover interesting patterns. Flowchart of the proposed data mining procedure is presented in Fig. 1 to better clarify the structure of the paper.

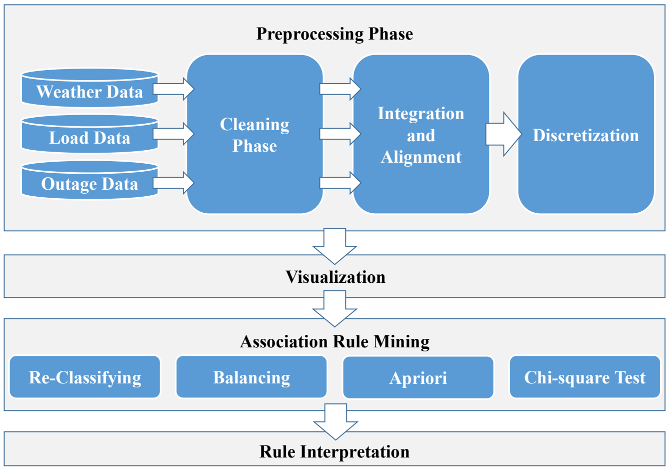

Fig. 1. Flowchart of the proposed method.

A. *Preprocessing*

Three different sources of data used in this study are shown in Table II. Preprocessing step is described in the following three phases:

1) *Cleaning*

The weather dataset is already clean and in a flat-comma separated format. The outage dataset has been stored with an unusual structure. Each record consists of three rows and seven columns and has been sorted by the feeder number and date. Some cells are blank or contain irregular values. For example, substation names were not stored consistently. In such cases, if additional information through which substation names could be modified was available, modification would be done; otherwise, the record would be eliminated. After eliminating or fixing these inconsistent values, the dataset is converted into a flat comma-separated file.

Each row in the load dataset contains 24 values, representing an hourly load for one substation in a day. For cleaning this dataset, load values are transposed and a time field is added to each row to represent the load measured for an hour for one substation. In addition, another feature is added to the dataset as '*Normalized Load'* which is

$$Normalized\ Load = \frac{Hourly\ \ Load}{Average\ Substation\ Load} \qquad (1)$$

In (1) *Hourly Load* is the actual measured load for that hour in the substation and *Average Substation Load* is the average measured load for that substation in a year.

2) *Integration and Alignment*

To perform analysis of data, the researcher in the current research did the alignment and integration of the datasets. Since substation load does not have direct effects on animal and vegetation-related outages, vegetation and animal-related datasets are produced from the integrating of both outage and weather datasets from 2013 to 2018. However with regard to both the possible load effects and the inaccessibility of the load data prior to March 2017, equipment-related datasets have been generated from the combination of outage, weather and load datasets over a one year period from March 2017 to March 2018. As mentioned before, each record in the datasets has a timestamp. Since weather and load datasets are time-based and the outage dataset is event-based, timestamps are not exactly aligned. Hence, to integrate these datasets, it is necessary to match outage records with weather conditions and the load causing the outage. To match outage and weather datasets, it needs to be pointed out that each record in weather dataset shows the weather conditions in the last three hours. As a result, the weather condition of the outage time is recorded after the outage timestamp. Therefore, to align outage records with weather conditions, we consider the next recorded weather data to be relevant. To match outage and load datasets, we also regarded the load recorded at each hour as the outage feature for the next hour. Fig. 2 shows the presented alignment method for the datasets.

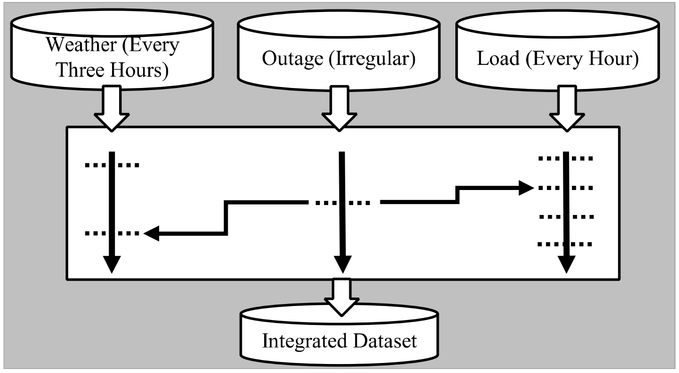

Fig. 2. Alignment method used to merge three datasets.

3) *Discretization*

To prepare data for association rule mining, continuous features are discretized. Because association rule mining is greatly disturbed by having imbalanced features and classes, the equal frequency-binning method is used to convert each continuous feature into a categorical feature with nearly equal count categories. Therefore, the newly generated features are completely balanced. Features like *hour* and *month* are also replaced by *Daytime* (*Morning*, *Evening*, *Night*) and *Season* (*spring*, *summer*, *autumn*, *winter*) respectively. Table III shows continuous features and their discretized values.

B. *Visualization*

In this section, several visualization techniques are used to visualize and get a better understanding of the inner relations of data fields. In Fig. 3, the relation between outage classes and the measured load before the incident is illustrated. This shows that equipment failures in higher *loads* occur more than other failures. This result shows that the load amount might have an impact on equipment failure.

TABLE III
FEATURES AND ASSOCIATED DISCRETIZED VALUES

Continues Feature	Discretized Feature	Condition	Value
Hourly Load	Hourly Load (D)	Hourly Load < 8.75	1
		8.75 <= Hourly Load < 15.6	2
		15.6 <= Hourly Load	3
Normalized Load	Normalized Load (D)	Normalized Load < 0.83	1
		0.83 <= Normalized Load < 1.13	2
		1.13 <= Normalized Load	3
Temperature	Temperature (D)	Temperature < 14.2	1
		14.2 <= Temperature < 24.6	2
		24.6 <= Temperature	3
Humidity	Humidity (D)	Humidity < 18	1
		18 <= Humidity < 37	2
		37 <= Humidity	3
Time	Day Time	8:00 <= Time < 16:00	Morning
		16:00 <= Time <= 23:59	Evening
		00:00 <= Time < 8:00	Night
Month (Jalali)	Season	1 <= Month <= 3	Spring
		4 <= Month <= 6	Summer
		7 <= Month <= 9	Autumn
		10 <= Month <= 12	Winter
Wind	Windy	Wind < 1	False
		1 <= Wind	True
Precipitation	Precipitation (D)	Precipitation = 0	False
		0 < Precipitation	True

In Fig. 4, for the sake of comparison, failures of insulator, pole and cable termination versus *season* and *load* data are illustrated. This shows that *load* amount and *season* can have different impacts on the different equipment failures. For pole, the *season* seems to be irrelevant as the failure rate is assumed to be fixed throughout the year. However, for other equipment like insulator and cable termination, the failure rate varies in different seasons.

C. *Association Rule Mining*

Determining features causing animal, vegetation or equipment-related outages and investigating features affecting each of the equipment separately are two general goals explored in this paper. These goals are fulfilled in four steps:

1) Re-Classification

In this step, nine sub-datasets of the main dataset are created. Each sub dataset is created to investigate one of the main causes of outage. These datasets are:

a) Animal:

In this sub-dataset, every animal-related outages are labeled *animal* and every other class is labeled by *other*. The goal of this dataset is to distinguish the parameters having more effect on animal-related incidents than other types of outages.

b) Vegetation:

This sub-dataset is generated by aggregating all trees and vegetation-related outages in one class named *vegetation* against all other types of outages labeled *other*.

c) Equipment:

In equipment failure outages, every outage caused by at least one faulty equipment is aggregated in one class named *equipment* and other outage causes are labeled *other* as the keyword.

d) Separate equipment sub-datasets:

To further investigate the causes of equipment failure, we labeled the outage for each specific equipment by that equipment name. For example, transformer faults and cut-out fuse-related outages were labeled as *transformer* and *cut-out fuse* respectively. Accordingly, 12 different equipment sub-datasets are created but only six have enough occurrence frequency to be considered for the final step.

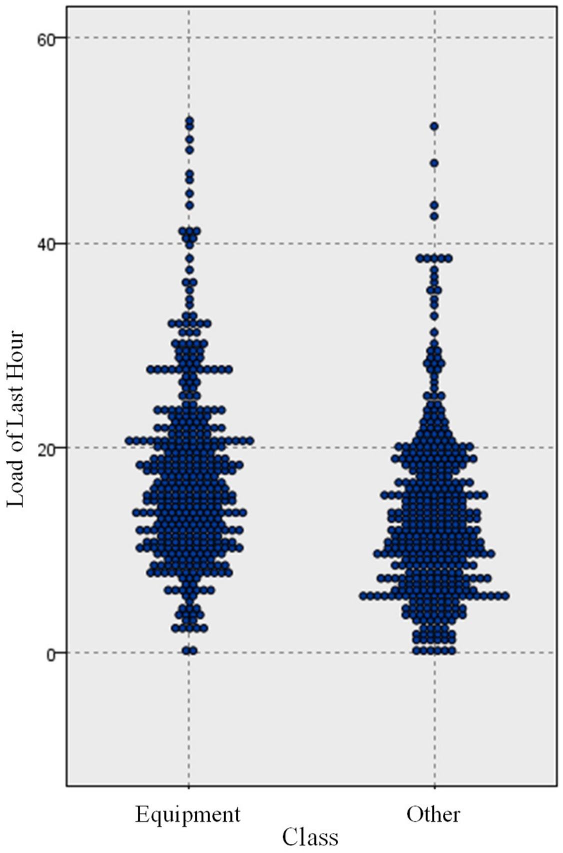

Fig. 3. 2-D illustration of the relation between load and outage class

2) Balancing

As mentioned in the previous section, nine different binary sub-datasets are generated. Since the class labeled as *other* is more frequent than the main class in every dataset, prior to any further investigation, every sub-dataset is balanced through the reduce method. To apply this method to the nine sub-datasets created in the reclassification phase, we remove samples from the class with higher frequency (*other*), so that the frequency of the main class and *other* class would be the same. As a result, every sub dataset includes a class labeled 'other' representing instances of all other classes. In both classes the number of records are approximately equal. Table IV shows the investigated outage causes, the frequency of their occurrences and the date range of the sub dataset based on the data availability.

3) Apriori

In this paper, the Apriori algorithm is used to find features of data that occur in a correlated matter to determine important features in each type of outage. Apriori algorithm was proposed in 1994 to find frequent itemsets in a transactional database in a bottom-up approach. To measure the interestingness of each rule in the Apriori technique, we utilize

TABLE IV
INVESTIGATED CAUSES, TIME INTERVAL AND TOTAL COUNT

Outage Cause	Time Interval	Total Count
Vegetation	5 Years	210
Animal	5 Years	251
Equipment	1 Year	405
Transformer	1 Year	32
Pole	1 Year	41
Jumper	1 Year	38
Cable Termination	1 Year	103
Distribution Lines and Cables	1 Year	77
Insulator	1 Year	38

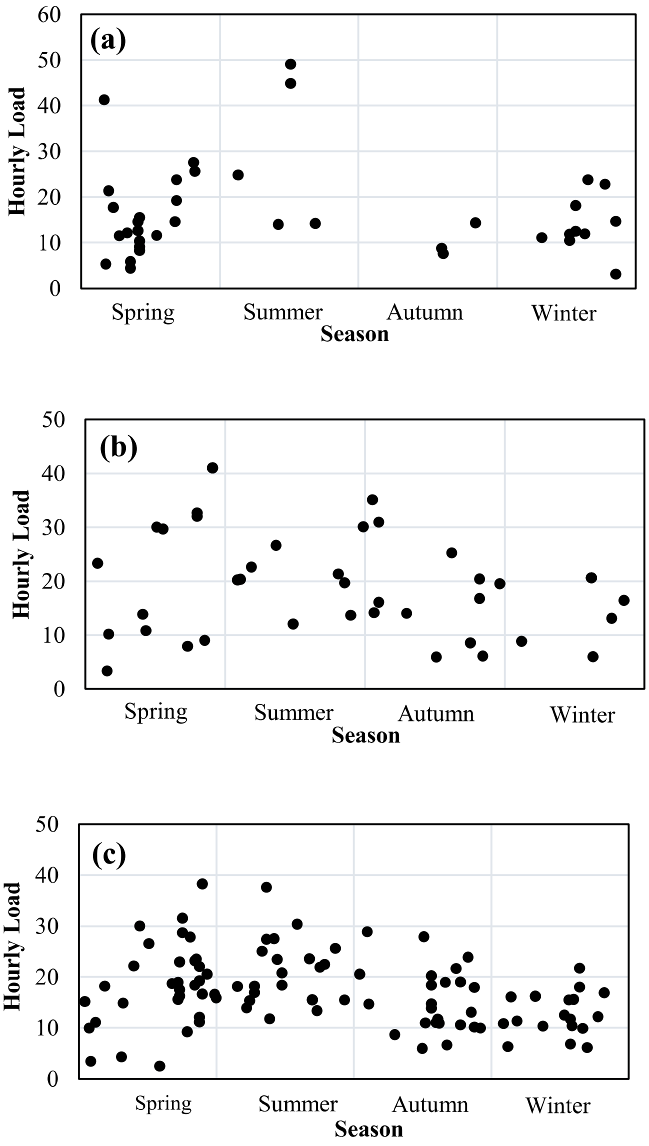

Fig. 4. 2-D illustration of relation between load and season from spring to winter; a) Insulator, b) Pole and c) Cable Termination

three parameters of support, confidence, and lift [1]. In association rule mining, for a rule with a form of $A \rightarrow B$, support of a rule is the percentage of transactions including both antecedent (*A*) and consequent (*C*) of the rule, as follows:

$$\text{support}(A \rightarrow C) = P(A \cup C) \quad (2)$$

In this case, each row of the sub-datasets is considered as a transaction. As a result of the balancing phase, support value, which is calculated here, is higher than the actual support. But, since the target class of each sub-dataset is fixed, the support values are comparable and their order is unchanged. The confidence parameter is the percentage of transactions including *A* within which *C* also lies, as follows:

$$confidence(A \rightarrow C) = P(C \mid A) \quad (3)$$

Similar to support value, the value of calculated confidence is higher than the actual value due to the balancing process, The lift value as the common metric in association rule mining is the ratio of $P(A \cup C)$ to $P(A)P(C)$ which will be close to one when *A* is independent of *C*, as follows:

$$lift(A, C) = \frac{P(A \cup C)}{P(A)P(C)} \quad (4)$$

If lift value is higher than one it means that *A* and *C* are positively correlated. To avoid generating too many rules which would be hard to inspect, we initially restrict the Apriori algorithm to find rules with support values higher than a minimum threshold (*minsup*) which is set to 10 percent in the present study. Additionally, any rule having a confidence value below 60% is omitted from final rulesets. All rules with the consequent of *other* are also removed since they are not helpful in achieving the objectives of the study.

4) Chi-square test

Since sampling can have an effect on the values of support, confidence and lift, it is necessary to further evaluate the interestingness of all rules to statistically filter out the insignificant ones. To this end, the chi-square test [27] is used here. For each rule, a confusion matrix is initially calculated for the whole population, and chi-square is then computed as:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{eij} \quad (5)$$

Where o_{ij} is the observed frequency of each ij^{th} cell (actual count), and e_{ij} is the expected count computed as:

$$e_{ij} = \frac{count(Antecedent = a_i) \times count(Consequent = c_j)}{n} \quad (6)$$

To compute the probability of the independence of *A* and *C* based on chi-square value, degree of freedom (*k*) is needed which computed as:

$$k = (r-1) \times (c-1) \quad (7)$$

Where *r* and *c* are the number of rows and columns in the confusion matrix, respectively. The probability of independence based on the calculated chi-square is computed through the following equation:

$$probability = \int_{\chi^2}^{\infty} f(x;k)dx = \int_{\chi^2}^{\infty} \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} dx \quad (8)$$

It could be proved that Apriori is a complete solution that finds all the association rules having their support and confidence more than specified minimum support and minimum confidence respectively [26]. In other words, all popular association rule mining algorithms like Apriori, Eclat and FP-growth produce the same results on a similar data. Their difference is on their efficiency and speed. All remaining rules are inspected by experts to extract meaningful results. Fig. 5 shows the pseudocode used in this paper.

IV. Results and Discussion

In this section, the presented results show the final rulesets for each sub-datasets. For each rule, support, confidence and lift values are calculated after running the Apriori algorithm. Then, the significant relation between antecedent and consequent which is determined by the chi-square test is bolded in tables. The confusion matrix is then used to calculate chi-square for each rule. As an example, consider the following rule:

'Season = Winter and Temperature = 1 and Humidity = 3 and Windy' → 'Vegetation'

Where the antecedent of the rule is *'Season = Winter and Temperature = 1 and Humidity = 3 and Windy'* and the consequent of the rule is *'Vegetation'*.

```
Apriori algorithm with lift and chi-square computation
$C_k$: Candidate itemset of size k
$L_k$ : Frequent itemset of size k

$L_1$ = {frequent items};
for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
  $C_{k+1}$ = candidates generated from $L_k$;
  for each transaction $t$ in database do
    increment the count of all candidates in $C_{k+1}$ that are contained in $t$
    $L_{k+1}$ = candidates in $C_{k+1}$ with minsup
    for each transaction in $L_{k+1}$ do
      compute lift = p(antecedent and consequent) / (p(antecedent)*p(consequent))
      compute chi-square = sum((observed – expected)$^2$ / expected)
  end
return $\cup_k L_k$;
```

Fig. 5. Pseudocode used to mine the association rules in this paper

TABLE V
A Sample Confusion Matrix: Expected Values (Observed Counts)

	Consequent	
	Vegetation	**Other**
Antecedent = True	9.56 (22)	1708.44 (1696)
Antecedent = False	200.44 (188)	35820.56 (35833)

According to (5), in this example, using values from Table V, chi-square is equal to 17.055, and the probability that a pair of irrelevant antecedent and consequent would generate a confusion matrix with the degree of freedom = 1 would be smaller than 0.001 based on equation 8. This probability is compared with a minimum threshold of significance (p-value) to determine whether the rule is significant or not. In the current research, the p-value is assumed to be 0.05. Chi-square and p-value of other rules are also calculated in the same way.

Vegetation-related ruleset presented in table VI is generated through the Apriori algorithm with *minsup* of 10 percent. It is observed that most vegetation-related outages happen in bad weather conditions. *Precipitation*, low *temperature*, *wind,* and high *humidity* are frequent patterns in vegetation-related outages. *Winter* which is also the most frequent season in this type of outages might be related to the fact that tree branches are more fragile in this season. Although morning and evening *Daytime* appear in the ruleset, this field does not seem to be relevant to vegetation-related outages.

The ruleset for animal-related outages is represented in table VII. It is perceived that animal-related incidents are the least frequent in winter. This might refer to fewer activities of animals, especially birds in cold days. Furthermore, since birds and animals are more active in Morning and Evening *Daytime*, animal-related incidents are observed to be more frequent. High *temperature*, low *humidity*, and *wind* are also observed in the rules. This can be justified in the light of the correlation of high *temperature* and low *humidity* with the season on the one side and the impact of the wind on changing birds' flight paths on the other.

Table VIII shows the results when all equipment is considered as one class. In equipment-related rulesets, *minsup* value is assumed to be 10%. It is also observed that *Load*-related features appear to be the predominant cause of failures in this ruleset. *Hourly Load which* seems to be more correlated with the equipment failure can be used to change the topology of the network and distribution of the load between feeders. *Spring* season, *wind*, morning *Daytime*, and low *humidity* are other frequent conditions in this ruleset.

It is also significant to separately analyze each equipment's failure. Table IX shows frequent features causing *transformer* failures. As expected, morning *Daytime*, high *load*, and summer *season* are frequent in this ruleset. This shows that when *load* is higher than a maximum threshold, it might cause *transformer* failure. In addition, low *humidity* is observed to cause transformer failure in all conditions. *Wind,* which is another frequent feature in this ruleset, might affect the transformers' performance in an indirect way.

Table X shows that *wind* is the most dominant factor for pole-related failures. *Load*, low *humidity,* and high *temperature* seem to be other frequent causes of outages. Although having lower chi-square compared with other tables, the rules of this ruleset are all validated by lift and chi-square values. In Table XI, jumper-related incidents which are presented in the rulesets shows that jumpers are affected by

TABLE VI
Ruleset of Vegetation-Related Outages

Consequent	Antecedent	Support %	Confidence %	Lift	Chi-Square	Probability
Vegetation	Precipitation (D) and Season = Winter	10.38	87.23	1.88	**220.186**	**< 0.001**
Vegetation	Precipitation (D) and Temperature (D) = 1 and Humidity (D) = 3	14.13	82.81	1.79	**265.272**	**< 0.001**
Vegetation	Season = Winter and Daytime = Morning	11.04	76.00	1.64	**43.958**	**< 0.001**
Vegetation	Daytime = Evening and Humidity (D) = 3	13.25	68.33	1.47	**18.681**	**< 0.001**
Vegetation	Daytime = Evening and Temperature (D) = 1	11.70	64.15	1.38	**3.962**	**0.047**
Vegetation	Season = Winter and Temperature (D) = 1 and Humidity (D) = 3 and Windy	10.15	63.04	1.36	**17.055**	**< 0.001**

TABLE VII
RULESET OF ANIMAL-RELATED OUTAGES

Consequent	Antecedent	Support %	Confidence %	Lift	Chi-Square	Probability
Animal	Humidity (D) = 1 and Windy	11.69	77.05	1.60	**27.687**	**< 0.001**
Animal	Season = Summer and Humidity (D) = 1 and Daytime = Morning and Temperature (D) = 3	11.49	73.33	1.53	**72.641**	**< 0.001**
Animal	Daytime = Morning and Windy	12.45	72.31	1.50	**65.715**	**< 0.001**
Animal	Windy and Temperature (D) = 3	14.56	67.11	1.40	**37.567**	**< 0.001**
Animal	Season = Autumn and Daytime = Morning	12.26	64.06	1.33	**44.847**	**< 0.001**
Animal	Season = Summer and Windy and Temperature (D) = 3	10.15	60.38	1.26	**15.523**	**< 0.001**
Animal	Season = Spring and Humidity (D) = 1 and Temperature (D) = 3	10.54	60.00	1.25	**8.212**	**0.004**
Animal	Daytime = Evening and Humidity (D) = 1 and Temperature (D) = 3	10.54	60.00	1.25	1.353	0.245

TABLE VIII
RULESET OF EQUIPMENT-RELATED OUTAGES

Consequent	Antecedent	Support %	Confidence %	Lift	Chi-Square	Probability
Equipment	Season = Spring and Hourly Load = 3	11.78	73.11	1.42	**34.985**	**< 0.001**
Equipment	Daytime = Morning and Hourly Load = 3 and Windy	13.43	72.64	1.41	**75.749**	**< 0.001**
Equipment	Temperature (D) = 3 and Daytime = Morning and Hourly Load = 3 and Windy	10.01	70.88	1.38	**21.698**	**< 0.001**
Equipment	Humidity (D) = 1 and Temperature (D) = 3 and Hourly Load = 3 and Windy	14.06	70.27	1.36	**8.098**	**0.004**
Equipment	Normalized Load = 3 and Humidity (D) = 1 and Daytime = Morning and Hourly Load = 3	11.02	70.11	1.36	**69.689**	**< 0.001**
Equipment	Normalized Load = 3 and Temperature (D) = 3 and Hourly Load = 3 and Windy	13.68	69.44	1.35	**39.292**	**< 0.001**
Equipment	Season = Spring and Daytime = Morning	13.18	69.23	1.34	**47.971**	**< 0.001**
Equipment	Normalized Load = 3 and Humidity (D) = 1 and Temperature (D) = 3 and Daytime = Morning and Hourly Load = 3	10.64	69.04	1.34	**62.87**	**< 0.001**
Equipment	Normalized Load = 3 and Humidity (D) = 1 and Temperature (D) = 3 and Hourly Load = 3 and Windy	11.66	68.47	1.33	**29.779**	**< 0.001**
Equipment	Humidity (D) = 1 and Temperature (D) = 3 and Daytime = Morning and Hourly Load = 3	13.05	67.96	1.32	**60.519**	**< 0.001**

TABLE IX
RULESET OF TRANSFORMER-RELATED OUTAGES

Consequent	Antecedent	Support %	Confidence %	Lift	Chi-Square	Probability
Transformer	Hourly Load = 3 and Humidity (D) = 1 and Temperature (D) = 3 and Daytime = Morning	21.31	92.30	1.75	**43.476**	**< 0.001**
Transformer	Season = Summer and Humidity (D) = 1 and Normalized Load = 3 and Daytime = Morning	18.03	90.90	1.73	**36.633**	**< 0.001**
Transformer	Season = Summer and Normalized Load = 3 and Temperature (D) = 3 and Daytime = Morning	18.03	90.90	1.73	**34.393**	**< 0.001**
Transformer	Hourly Load = 3 and Humidity (D) = 1 and Daytime = Morning and Windy	18.03	90.90	1.73	**32.93**	**< 0.001**
Transformer	Season = Summer and Humidity (D) = 1 and Normalized Load = 3 and Temperature (D) = 3 and Daytime = Morning	18.03	90.90	1.73	**36.633**	**< 0.001**
Transformer	Hourly Load = 3 and Humidity (D) = 1 and Temperature (D) = 3 and Daytime = Morning and Windy	16.39	90	1.71	**30.442**	**< 0.001**
Transformer	Season = Summer and Hourly Load = 3 and Humidity (D) = 1 and Daytime = Morning	14.75	88.88	1.69	**34.365**	**< 0.001**
Transformer	Season = Summer and Hourly Load = 3 and Temperature (D) = 3 and Daytime = Morning	14.75	88.88	1.69	**31.859**	**< 0.001**
Transformer	Season = Summer and Hourly Load = 3 and Humidity (D) = 1 and Normalized Load = 3 and Daytime = Morning	14.75	88.88	1.69	**41.911**	**< 0.001**
Transformer	Hourly Load = 3 and Humidity (D) = 1 and Normalized Load = 3 and Daytime = Morning and Windy	14.75	88.88	1.69	**32.045**	**< 0.001**

wind, *load*, *temperature,* and low *humidity*. As illustrated in Table XII, cable termination is also affected by *load*. Low *humidity*, morning *time*, and spring *season* also seem to be negatively affecting cable terminations. Unlike most of the equipment, overhead lines seem to be resistant to *load*-related factors as presented in Table XIII. Medium to high level *temperature*s in warmer times of a day, warmer *seasons* of the year and *wind* speed are the major causes of overhead line failure. The high *temperature* would increase the line sag and *wind* might cause a short circuit fault. Table XIV displays insulator-related outage ruleset. As observed, insulators are affected by high *temperature*s, *precipitation*, and *wind* in spring days.

TABLE X
RULESET OF POLE-RELATED OUTAGES

Consequent	Antecedent	Support %	Confidence %	Lift	Chi-Square	Probability
Pole	Daytime = Evening and Temperature (D) = 3 and Windy	10.84	88.88	1.79	**5.733**	**0.017**
Pole	Season = Spring and Daytime = Morning and Windy	12.04	80	1.61	**13.137**	**< 0.001**
Pole	Season = Spring and Humidity (D) = 1	10.84	77.77	1.57	3.503	0.061
Pole	Daytime = Night and Hourly Load = 3	10.84	77.77	1.57	**5.423**	**0.02**
Pole	Season = Spring and Humidity (D) = 1 and Temperature (D) = 3	10.84	77.77	1.57	**4.587**	**0.032**
Pole	Temperature (D) = 3 and Windy	28.91	75	1.51	**9.274**	**0.002**
Pole	Humidity (D) = 1 and Normalized Load = 3 and Temperature (D) = 3 and Hourly Load = 3 and Windy	13.25	72.72	1.47	**7.086**	**0.008**

TABLE XI
JUMPER-RELATED RULESET

Consequent	Antecedent	Support %	Confidence %	Lift	Chi-Square	Probability
Jumper	Season = Winter and Temperature (D) = 2 and Windy	13.58	90.90	1.93	**30.357**	**< 0.001**
Jumper	Season = Winter and Temperature (D) = 2 and Hourly Load = 3	11.11	88.88	1.89	**62.254**	**< 0.001**
Jumper	Temperature (D) = 3 and Hourly Load = 3 and Normalized Load = 3 and Daytime = Morning and Humidity (D) = 1	11.11	88.88	1.89	**18.186**	**< 0.001**
Jumper	Hourly Load = 3 and Daytime = Morning and Humidity (D) = 1 and Windy	14.81	83.33	1.77	**25.113**	**< 0.001**
Jumper	Normalized Load = 2 and Hourly Load = 3 and Windy	11.11	77.77	1.65	**10.859**	**0.001**
Jumper	Temperature (D) = 3 and Hourly Load = 3 and Normalized Load = 3 and Humidity (D) = 1	14.81	75	1.59	**6.902**	**0.009**
Jumper	Hourly Load = 3 and Normalized Load = 3 and Humidity (D) = 1 and Windy	12.34	70	1.49	**4.541**	**0.033**
Jumper	Temperature (D) = 3 and Hourly Load = 3 and Normalized Load = 3 and Humidity (D) = 1 and Windy	11.11	66.66	1.42	2.971	0.085

TABLE XII
CABLE TERMINATION-RELATED RULESET

Consequent	Antecedent	Support %	Confidence %	Lift	Chi-Square	Probability
Cable Termination	Season = Spring and Hourly Load = 3	12.09	76.92	1.60	**15.482**	**< 0.001**
Cable Termination	Season = Spring and Daytime = Morning	10.69	69.56	1.45	**7.107**	**0.008**
Cable Termination	Humidity (D) = 1 and Temperature (D) = 3 and Daytime = Morning and Hourly Load = 3	10.23	63.63	1.32	**6.139**	**0.018**
Cable Termination	Temperature (D) = 3 and Normalized Load = 3 and Hourly Load = 3 and Windy	11.16	62.5	1.30	**2.819**	**0.093**
Cable Termination	Season = Spring and Temperature (D) = 2	12.09	61.53	1.28	**2.282**	**0.131**
Cable Termination	Temperature (D) = 3 and Hourly Load = 3	20.46	61.36	1.28	**7.089**	**0.008**
Cable Termination	Humidity (D) = 1 and Normalized Load = 3 and Daytime = Morning	10.69	60.86	1.27	2.139	0.134
Cable Termination	Daytime = Morning and Hourly Load = 3 and Windy	10.69	60.86	1.27	3.288	0.07

TABLE XIII
DISTRIBUTION LINE-RELATED RULESET

Consequent	Antecedent	Support %	Confidence %	Lift	Chi-Square	Probability
Class = Distribution Line	Season = Summer and Daytime = Night and Normalized Load = 2	10	80	1.55	**38.384**	**< 0.001**
Class = Distribution Line	Hourly Load = 2 and Temperature (D) = 3	10	73.33	1.42	2.206	0.138
Class = Distribution Line	Season = Summer and Temperature (D) = 2	10.66	68.75	1.33	**7.626**	**0.006**
Class = Distribution Line	Daytime = Morning and Temperature (D) = 2	10.66	68.75	1.33	1.103	0.294
Class = Distribution Line	Temperature (D) = 3 and Hourly Load = 3 and Windy	14.66	68.18	1.32	**4.08**	**0.043**
Class = Distribution Line	Season = Spring and Temperature (D) = 3	12	66.66	1.29	1.861	0.173
Class = Distribution Line	Daytime = Evening and Normalized Load = 3 and Hourly Load = 3 and Windy	10	66.66	1.29	**6.611**	**0.01**
Class = Distribution Line	Hourly Load = 2 and Windy	15.33	65.21	1.27	0.689	0.407

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.2966505, IEEE Transactions on Industrial Informatics

TABLE XIV
INSULATOR-RELATED RULESET

Consequent	Antecedent	Support %	Confidence %	Lift	Chi-Square	Probability
Class = Insulator	Season = Spring and Humidity (D) = 1 and Windy	10.84	88.88	1.94	**10.643**	**0.001**
Class = Insulator	Season = Spring and Daytime = Morning and Temperature (D) = 3 and Windy	10.84	88.88	1.94	**24.679**	**< 0.001**
Class = Insulator	Precipitation (D) and Normalized Load = 1	12.04	80	1.74	**78.703**	**< 0.001**
Class = Insulator	Humidity (D) = 2 and Temperature (D) = 3 and Windy	10.84	77.77	1.69	**19.967**	**< 0.001**
Class = Insulator	Normalized Load = 2 and Hourly Load = 2	12.04	70	1.52	0.994	0.319
Class = Insulator	Humidity (D) = 3 and Hourly Load = 2	10.84	66.66	1.45	**0.413**	**0.008**

V. CONCLUSION

In this study, real distribution system fault data are analyzed to find frequent causes of failures. First, in addition to weather data commonly viewed as an important factor in power outage related problems, load data is also taken into account. Then, the impacts of various features on different failures were visualized. After visualization, the Apriori technique was used to extract frequent fault cause rules for vegetation, animal, and equipment related faults. According to the final rulesets, load data is one of the significant factors having an impact on equipment related outages. Then, different kinds of equipment were separately analyzed because they have different characteristics, produced from different materials and used in different circumstances.

The *season* feature seems to be effective in almost all of the investigated fault causes. Spring season is correlated with cable termination and insulator-related faults. Vegetation-related outages happen more in winter, whereas animal-related incidents are less frequent in this season. In terms of day time, animal-related failures mostly occur in the *morning* and *evening*. In *the morning*, transformers and cable termination failures occur more than others do. Noteworthy to say that low *temperature* is a frequent reason for vegetation-related faults while medium to high temperatures commonly cause overhead lines, insulators, jumpers, and animal-related outages. In light of the findings, it can be stated that the *humidity* of lower than 18% is the frequent cause of animal-related and most of the equipment-related faults. In the same line, the *humidity* of higher than 37% is the main reason for vegetation-related outages. Except for cable termination, all of the investigated failures, have a correlation with the *wind*. The studies have also shown that most of the equipment failures happen due to the *load* amount, and as a result, high *load* demand is regarded as the main rule in almost all equipment failures. The results of the current study can be utilized to plan future network, condition-based maintenance scheduling, and a better selection of the network components.

REFERENCES

[1] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[2] T. Soukup and I. Davidson, *Visual data mining: Techniques and tools for data visualization and mining*. John Wiley & Sons, 2002.

[3] K. Zheng, Q. Chen, Y. Wang, C. Kang, and Q. Xia, “A Novel Combined Data-Driven Approach for Electricity Theft Detection,” *IEEE Trans. Ind. Informatics*, 2018.

[4] P. Jokar, N. Arianpoo, and V. C. M. Leung, “Electricity theft detection in AMI using customers consumption patterns,” *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.

[5] S. K. Singh, R. Bose, and A. Joshi, “PCA based electricity theft detection in advanced metering infrastructure,” in *7th International Conference on Power Systems (ICPS)*, pp. 441–445, 2017.

[6] R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. S. Shen, “Energy-theft detection issues for advanced metering infrastructure in smart grid,” *Tsinghua Sci. Technol.*, vol. 19, no. 2, pp. 105–120, 2014.

[7] F. A. S. Borges, R. A. S. Fernandes, I. N. Silva, and C. B. S. Silva, “Feature extraction and power quality disturbances classification using smart meters signals,” *IEEE Trans. Ind. Informatics*, vol. 12, no. 2, pp. 824–833, 2015.

[8] B. Biswal, P. K. Dash, and B. K. Panigrahi, “Power quality disturbance classification using fuzzy C-means algorithm and adaptive particle swarm optimization,” *IEEE Trans. Ind. Electron.*, vol. 56, no. 1, pp. 212–220, 2008.

[9] B. Lundstrom, P. Gotseff, J. Giraldez, and M. Coddington, “A high-speed, real-time visualization and state estimation platform for monitoring and control of electric distribution systems: Implementation and field results,” *IEEE Power Energy Soc. Gen. Meet.*, 2015.

[10] A. W. McMorran, G. W. Ault, and J. R. McDonald, “Solving data integration challenges for web-based geographical power system data visualization using CIM,” *IEEE Power Energy Soc. 2008 Gen. Meet. Convers. Deliv. Electr. Energy 21st Century, PES*, 2008.

[11] J. N. Bank, O. A. Omitaomu, S. J. Fernandez, and Y. Liu, “Visualization and classification of power system frequency data streams,” *ICDM Work. 2009 - IEEE Int. Conf. Data Min.*, pp. 650–655, 2009.

[12] S. Dutta and T. J. Overbye, “Optimal wind farm collector system topology design considering total trenching length,” *IEEE Trans. Sustain. Energy*, vol. 3, no. 3, pp. 339–348, 2012.

[13] R. Eskandarpour and A. Khodaei, “Leveraging

accuracy-uncertainty tradeoff in SVM to achieve highly accurate outage predictions," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 1139–1141, 2018.

[14] R. E. Brown, *Electric power distribution reliability*. CRC press, 2008.

[15] P. Warlyani, A. Jain, A. S. Thoke, and R. N. Patel, "Fault Classification and Faulty Section Identification in Teed Transmission Circuits Using ANN," *Int. J. Comput. Electr. Eng.*, vol. 3, no. 6, pp. 807–811, 2012.

[16] H. A. Bin Illias, L. J. Awalin, S. S. Gururajapathy, H. Mokhlis, and A. H. Abu Bakar, "Fault location in an unbalanced distribution system using support vector classification and regression analysis," *IEEJ Trans. Electr. Electron. Eng.*, vol. 13, no. 2, pp. 237–245, 2017.

[17] 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, 2019.

[18] Y. Cai, M.-Y. Y. Chow, W. Lu, and L. Li, "Statistical feature selection from massive data in distribution fault diagnosis," *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 642–648, 2010.

[19] 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, 2014.

[20] M. Doostan, R. Sohrabi, and B. Chowdhury, "A Data-Driven Approach for Predicting Vegetation-Related Outages in Power Distribution Systems," *arXiv Prepr. arXiv1807.06180*, pp. 1–11, 2018.

[21] L. Xu, M. Chow, and L. S. Taylor, "Data mining and analysis of tree-caused faults in power distribution systems," in *2006 IEEE PES Power Systems Conference and Exposition*, pp. 1221–1227, 2006.

[22] D. T. Radmer, P. A. Kuntz, R. D. Christie, S. S. Venkata, and R. H. Fletcher, "Predicting vegetation-related failure rates for overhead distribution feeders," *IEEE Trans. Power Deliv.*, vol. 17, no. 4, pp. 1170–1175, 2002.

[23] P. Kankanala, A. Pahwa, and S. Das, "Estimating Animal-Related Outages on Overhead Distribution Feeders Using Boosting," *IFAC-PapersOnLine*, vol. 48, no. 30, pp. 270–275, 2015.

[24] M. Doostan and B. H. Chowdhury, "Power distribution system fault cause analysis by using association rule mining," *Electr. Power Syst. Res.*, vol. 152, pp. 140–147, 2017.

[25] M. Doostan and B. H. Chowdhury, "Power distribution system equipment failure identification using machine learning algorithms," in *IEEE Power and Energy Society General Meeting*, pp. 1–5, 2017.

[26] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, pp. 487–499, 1994.

[27] K. Pearson, "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 50, no. 302, pp. 157–175, 1900.

[28] Raspisaniye Pogodi Ltd, "Reliable Prognosis. Weather and Climate Change." [Online]. Available: https://rp5.ru/.

Mohammad Sadegh Bashkari Received the B.Sc. and M.Sc. degrees in software engineering from Shiraz University, Shiraz, Iran, in 2011 and 2013, respectively. He is currently a PhD candidate of software engineering at Shiraz University. His primary research interests are data analysis and application of data mining methods to interdisciplinary research area.

Ashkan Sami Obtained his B.Sc. from Virginia Tech; USA (1991), M.Sc. in AI and Robotics from Shiraz University; Iran (1996) and PhD from Tohoku University; Japan (2006). Ashkan's PhD thesis became the main idea of a national project funded by Japanese government and earned him a tenured faculty position at Tohoku University. Currently an Associate Professor of CSE and IT department at Shiraz University, Dr. Sami conducts high quality interdisciplinary research on applied Artificial Intelligence, data mining, cyber security, and software engineering. Dr. Sami has graduated more than 100 M.Sc. and PhD students under his direct supervision, served in various program committees of reputable national and international conferences and is a technical committee member of IEEE Software Engineering and IEEE Industrial Electronics.

Mohammad Rastegar Received the B.Sc., M.Sc., and Ph.D. degrees from the Sharif University of Technology, Tehran, Iran, in 2009, 2011, and 2015, respectively, all in electrical engineering. He is currently an Assistant Professor with the School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran. His current research interests include modeling home energy management systems, plug-in hybrid electric vehicle operation, and power system reliability and resiliency studies.