Distribution Power System Outage Diagnosis based on Root Cause Analysis

Mohammad Sadegh Bashkari, Ashkan Sami*, Mohammad Rastegar*, Mohammad Javad Bordbari

School of Electrical and Computer Engineering, Engineering Department, Shiraz University, Zand Avenue, Shiraz, Iran; Postal Code: 71348-51154

*Corresponding Authors: mohammadrastegar@shirazu.ac.ir and sami@shirazu.ac.ir

Phone number: +98 917 3041503

Abstract

This paper proposes data mining-based models to diagnose outage data in distribution power systems. In this work, outage data from a local distribution company is gathered and aligned with weather data. Then, a subset of features is selected to reduce the processing time and simplifying purposes. To increase the fairness of final models and to account for differences in misclassification cost, using a customized cost matrix is proposed. Two decision tree-based modeling algorithms are trained and tested. Results show the ability of the established models to diagnose the root cause of an outage fairly well. In addition, an ensemble of the decision tree-based models is built, which outperforms the other two models in almost all cases. Finally, applications of such models in decreasing outage duration and improving the reliability of the power distribution network are discussed.

Keywords- Power system reliability; Distribution system outages; Data mining; Random forest; Cost matrix; Ensemble.
1. Introduction

Data mining has been attracting the attention of most researchers in almost every science category as a way to turn the problem of large accumulated data into an opportunity. Data mining methods like classification, clustering, outlier detection, and visualization give the researchers a chance to gain a better understanding of the category. The power system is one of the research areas that has received contribution from data mining methods. Meter and power consumption data have been used in several types of research to classify consumers or to identify distinctive ones, hence detecting electricity theft [1]–[3]. Visualization techniques have been used to ease the monitoring operations [4], [5] and to realize sudden changes in power systems’ parameters[6].

In recent years with the availability of outage data, the amount of literature on reliability of power distribution networks has been increased. This is due to the fact that most outages in power system are rooted in distribution section of the network [7]. Since many types of research show that value of the lost load is extremely higher than its actual cost [8], studying reliability in the power systems especially distribution system seems inevitable. Reliability of distribution system depends on many parameters, such as number of outage occurrences, duration of the outages, number of affected consumers. Therefore, investigating the impacts of these parameters and roots, which endanger reliability of distribution system, has recently gained a lot of attention. Since various and frequent data are available in this context, data mining methods have been utilized for diagnosis or prognosis of failure occurrences in the system. For example, component outages caused by hurricane are predicted in [9] using logistic regression. Reference [10] investigates the effect of storm on outages using high-resolution weather simulation data using decision tree, random forest and boosted gradient tree. Significant correlated factors in outage
occurrence have been studied in [11] using association rule mining, and in [12] using fuzzy inference system. In [13], animal-related outages are investigated, several related factors are introduced and a prevention method is proposed. Other works have been done to estimate fault location, hence shorten the investigation part of restoration. For example, Discrete Fourier transform and artificial neural networks have been used in [14] to determine faulty section in teed transmission lines. Fault causes are often classified as vegetation-, animal-, weather-related [15] and equipment failure [16] groups. Many types of research have tackled the problem of identifying the root cause of an outage. This problem has been addressed in [15] using fuzzy classification to identify vegetation, animal and lightening related outages and in [16], [17] to identify equipment failure and vegetation related outages, respectively. All these researches used historic data from past outages and use multi-label or binary classification methods to identify the main outage cause upon occurring.

In this paper, data mining-based methods are used to find a predictive model for analyzing and deducing the root cause of the power distribution system outages upon occurrence. To this end, four-year outage data from local distribution company and local weather data are integrated. Then, outage cause labels from outage data are reclassified into vegetation-, animal-, equipment-, and weather-related classes. Other fault causes that do not belong to these classes are labeled as ‘other’ class. After reclassifying the outage labels, data are partitioned into train and test partitions. A customized cost matrix is configured based on the importance of each class and the balance ratio between classes. In this case, using a cost matrix is crucial for two reasons. First, the cost of incorrect classification of an outage class should not be the same for all cases. Therefore, some misclassifications should receive higher penalty. Another important reason to use the cost matrix is to account for imbalance ratio of classes, which may result in in valuable
realization. After building the cost matrix, several data mining classification models, such as random forest [18] and C5.0 [19]-[20], are considered upon training partition. Then, generalizability of these two models is verified against test partition. In addition, another model is built as an ensemble [21] of the previous ones, which inherits the best characteristics of each model and outperforms both models in most cases. The models are tested and compared to show the ability of these models to predict the cause of outages upon occurrence. The results are presented, and the effects of such predictors on the reliability of distribution system are discussed.

2. Material and methods

Data sources used in this work are non-simulated real-world data. Distribution outage data has been collected from a local distribution network, which its characteristics are presented in table I, which consists of substation name, outage cause, date, and time of the occurrence, and a coded feeder number among other features. Also, weather data for the area is retrieved from reliable prognosis weather database [22] for the same period of time. Time and date, temperature, humidity, pressure, precipitation, and wind speed are some of notable features of this data.

2.1. Preprocessing

The described data sources have different temporal granularity. In other words, weather data is collected regularly every three hours, whereas each row of outage data is appended to the data source upon outage occurrence. This fact leads to the generation of two data sources with different timestamps. To resolve this issue, integration between the two data sources is done.
More preprocessing steps to make the resulting dataset ready are taken as follows. In the aforementioned distribution company, 96 outage cause labels are used by technicians to classify outage causes. Some of these labels are related and could be combined together and some others could be removed due to being either less frequent or less important than others. Therefore, first scheduled outages are removed from this data since they are inevitable and could be planned to be less effective on reliability. Outages that are self-resolving or the ones whose duration is less than five minutes are also removed due to low effect on reliability. Then the remaining outage causes are reclassified and labeled as vegetation-related, animal-related, equipment-related, and weather-related. Other fault causes that could not be classified in the aforementioned classes are classified into ‘other’ class. With further inspection in data, and some features with more than 95% null values are removed. Moreover since online root cause analysis models should be able to work as soon as outage occurrence, some other features like ENS or restoration time are removed as they are not available within five minutes from outage manifestation.

2.2. Visualization

To get a better understanding of the data, several illustrations of the relations between features are created and investigated. Figure 1 shows the imbalance ratio of class members. As it is illustrated, most of the records in the data are from the equipment class. This would be problematic at modeling phase, as most of the modeling algorithms tend to favor the majority class to improve accuracy. This problem is addressed before training the models.

To investigate the effect of region on each root cause of outages, the number of outages is illustrated in fig.2, separately for each substation, divided by root cause. It is obvious that outages occur less frequently in some substations like substations 16, 23, 41, 44, 50 and 52. On the other hand, substations 11, 12, 21, 46, 58 and 63 have more outage count than the others.
Moreover, the ratio of the root causes in each substation is different. For example, animal caused outages are responsible for a higher ratio of outages in substation 12 than substation 58. This shows that to predict the root cause of an outage, a model should consider substation number as an input.

### 2.3. Feature selection

With the growth of computer power, and advancement of data collection technologies, there are a lot of data generated in every field. Even using data mining and machine learning algorithms, this amount of data is hard to analyze. Feature selection is one of the methods that deals with this problem. Feature selection is the operation to choose a smaller subset from the main set of features without sacrificing accuracy by removing irrelevant or redundant features. This operation allows us to achieve similar or even improve predictor’s performance and at the same time increase model’s comprehensibility[23]. In this paper, the feature selection is performed in two phases. First features are ranked based on their correlation to the outage cause. This type of feature selection is called filter method[24]. This correlation could be measured by likelihood ratio or Pearson metrics among others. In this paper, since the sample size is relatively small, likelihood ratio is preferred. This phase is done regardless of the chosen predictor model and other inputs. Based on quality of the feature and diversity of values, 25 features are selected in total.

In the next phase of feature selection, which is similar to wrapper[24] methods, predictor model and correlations between input fields are also considered. In this phase, the subset of features chosen in the filter phase is used as input. Since different models are built, in each one different number of features are selected and used in the training phase. Figures 3 and 4 illustrate the most important features based on random forest and C5.0 algorithms respectively.
2.4. Cost matrix

Since misclassification of records is costly, in many data mining algorithms, there is a cost matrix that inputs cost of incorrectly classified record based on observed label and predicted label. In many use cases, there is not much difference between the costs of misclassifications. In other words, cost is not dependent on observed label and predicted label. Rather every misclassification costs are the same. Table II shows a common cost matrix, used in many classifiers.

In cases that the cost of misclassification varies for each observed and predicted pair, the cost matrix can be configured to reflect this difference. For example, in this research, misclassifying the ‘equipment’ class as ‘other’ would cost significantly more than misclassifying ‘other’ as ‘equipment’. The reason is that in the first case, the repair crew might be dispatched without being properly prepared to deal with equipment failure and it would prolong the restoration process. On the other hand, being prepared for equipment failure would not significantly influence the repair crew’s readiness to resolve other types of failure.

As illustrated in the visualization section, records are not equally distributed in all classes. Rather, more than half of the records are from equipment class. In such imbalanced cases, in order to improve the overall accuracy, data mining algorithms would favor the major class over other classes. This would result in accurate models, which may be impractical in real-world situations. Another important use of a customized cost matrix is to deal with this kind of situations by increasing the cost of misclassifying other classes as the major class. Table III shows the cost matrix, which is configured to deal with above issues. As the least important class, ‘other’ row has been altered to be zero so that there is no penalty for misclassifying members of this class.
In table III using the same cost values in equipment failure column as ‘other’ columns would result in a model that predicts most root causes as ‘equipment’ class. To prevent this issue and to build a fair model, costs on this column were changed so that predicting animal, vegetation, and weather classes as ‘equipment’ class would be more costly. In this way the model will adapt to predict classes in a more trustworthy manner.

2.5. Model training

The goal of this paper is to build a predictive model to analyze and deduce the root cause of outages upon occurrence. This would greatly help the network operator choose and dispatch the most suited and best-equipped repair crew available to deal with the failure. To this end, two data mining-based methods and an ensemble method are used, as described in the following.

2.5.1. Random forest results and discussion

Random forest algorithm was introduced by Breiman in 2001[18]. In the random forest algorithm, a model is built in several iterations. In each iteration, a sample of data is chosen which is called a bootstrap sample. Then a single decision tree is built based on this sample. To this end, at the root node of the tree, the sample data are divided into two subgroups. Subgroups are chosen based on an impurity index. Then, for each new node, its respective subgroup is split into two more subgroups based on the same index. The algorithm continues until a stopping condition, and a single decision tree is constructed. To build a random forest model, hundreds of decision trees are built in the same way using different bootstrap samples. To predict the class of a new instance, every tree makes a decision between available classes. The final decision of this model is made by voting between all of the trees in the forest. This algorithm is more robust against overfitting [18] in comparison with other decision tree-based algorithms, as a result of
this randomness. So, the predictions on the test part could be expected to have the same accuracy as the predictions on the train part. Table IV presents the result of applying this algorithm to integrated outage and weather data. Each row shows observed or real root cause of outages while each column indicates the predicted label by random forest algorithm. The number on each cell is the percentage that the members of the representative row class, labeled by the predictor as the respective column class. Bolded values show the accuracy of the predictor the respective class. This table shows that random forest algorithm is able to predict animal, equipment, vegetation, and weather classes fairly well. Because of using customized cost matrix on this predictor, none of the records of these four important classes is falsely predicted as ‘other’, as predicting ‘other’ would not help the repair crew to prepare and diagnose the problem. Furthermore, although the equipment class is by far the most populated class, the predictor does not significantly prioritize this class over other classes.

Table V shows the result of the model built by the random forest algorithm, considering the default cost matrix. As expected, the algorithm optimizes the model for the highest accuracy, and as a result, most of the instances are predicted as the majority classes, i.e., equipment and weather. While this model improves the accuracy of the equipment class, it immensely reduces other classes’ accuracy. Thus, this model may not be applicable in a real-world power distribution network.

2.5.2. C5.0 results and discussion

C5.0 algorithm, proposed in 1997 by JR Quinlan, is based on basic decision tree and C4.5[20]. Unlike the random forest algorithm, C5.0 provides a single decision tree and does not use bootstrap sampling. The advantage of building a single tree is that the results could be interpreted and used as new knowledge in the field. On the other hand, C5.0 model suffers from overfitting
problem, which occurs when a model has high accuracy on the train partition and has low accuracy on the test partition. Hence, the model cannot be generalized well to new instances.

Table VI shows the results of the built model based on the C5.0 algorithm. Like the random forest algorithm, C5.0 algorithm shows promising results in predicting the root cause of outages. The C5.0 model is outperformed by random forest in two important classes, but it has the advantage of building simpler to understand model, compared to the random forest model. The reason of C5.0 model’s simplicity is that it builds a single tree, unlike the random forest-based model, which operates as a black box predictor and does not reveal the logic behind each prediction.

Table VII shows the results of the model built by using C5.0 algorithm without the customized cost matrix. Results show that similar to the application of random forest algorithm, this model predicts most of the instances as equipment and weather, which are the majority classes while reducing the other important classes like animal and vegetation, drastically.

2.5.3. Ensemble results and discussion

Ensemble method was proposed in 1993 by Hinton and has been used to combine models to get accurate results. Since none of the presented models is completely dominant over the other, to capture the most confident prediction of each model, using the ensemble of both models is proposed. As it is observed in tables IV to VII, while the random forest algorithm predicts vegetation and weather causes more accurately, in the case of equipment failure, it is outperformed by the C5.0 model. To combine these models, two conditions are considered based on the class predictions for each record. In the case of an agreement between the labels suggested by both algorithms, suggested label would be produced as the ensemble model’s output as well.
In the case of a disagreement between models’ output label, the classifier with highest confidence would be the selected. Based on these circumstances, ensemble results are generated and presented in table VIII. It is observed that ensemble model outperforms C5.0 model in all cases and it is rather better than the random forest model in all but vegetation related outages.

To compare the results of C5.0, random forest, and the proposed ensemble algorithms, the accuracy of all presented models for each class is presented in fig.5. Based on this illustration, models, which use default cost matrices, fail to predict animal and vegetation classes and have the lowest accuracy level for the weather class. Furthermore, these models are not useful in practice as they are just able to predict one class correctly. Among the other three models, the proposed ensemble algorithm can accurately predict animal, equipment, and weather classes, while the random forest algorithm has the highest accuracy in predicting vegetation class.

It is also observed in all three models that vegetation related outages have the lowest percentage of accuracy. This fact might be related to two factors. First, the low number of samples of this class, and second, the unpredictable nature of this kind of outage. Figure 6 shows the accuracy of each model, with different train and test sets. It is observed that the proposed ensemble has the best performance, and random forest is the least desirable one in terms of accuracy.

3. Conclusion

In this paper, animal, equipment, vegetation, and weather-related outages in a power distribution system are discussed. First, four models are trained to analyze the root cause of outages upon occurrence. In the experiments, it was shown that unbalanced nature of data could be overcome using the appropriate cost matrix. It was also proved that the degree of importance of a given root cause could be reflected in the same matrix, and these models could be adapted to the
requirement of the distribution system operator. For example, if animal-related failures are the most important, the cost matrix could be customized so that this kind of outage has the most priority in the resulting model. To increase the predictor performance, it was proposed to combine random forest and C5.0 based models to form an ensemble model. The results showed that accuracy of prediction may be increased by 5% and 2% in the ensemble model than random forest and C5.0 models, respectively. The proposed data mining-based model may shorten the repair times and improve the distribution power system reliability.

4. References

[1] Zheng, K. , Chen, Q. , Wang, Y. , et al., “A Novel Combined Data-Driven Approach for Electricity Theft Detection,” *IEEE Trans. Ind. Informatics*, 15(3), pp. 1809–1819 (2018).

[2] Jokar, P. , Arianpoo, N. , and Leung, V. C. M., “Electricity theft detection in AMI using customers consumption patterns,” *IEEE Trans. Smart Grid*, 7(1), pp. 216–226 (2016).

[3] Singh, S. K. , Bose, R. , and Joshi, A., “PCA based electricity theft detection in advanced metering infrastructure,” *7th Int. Conf. Pow. Syst. (ICPS)*, pp. 441–445 (2017).

[4] Venkatesh, A. , Cokkinides, G. , and Meliopoulos, A. P. S., “3D-visualization of power system data using triangulation and subdivision techniques,” *Proc. 42nd Annu. Hawaii Int. Conf. Syst. Sci. HICSS*, pp. 1–8 (2009).

[5] McMorran, A. W. , Ault, G. W. , and McDonald, J. R., “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).

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

[7] Brown, R. E., “Electric power distribution reliability,” 2nd Edn., CRC press (2008).

[8] Schroder, T. and Kuckshinrichs, W., “Value of lost load: An efficient economic indicator for power supply security? A literature review,” *Front. energy Res.*, 3, p. 55 (2015).

[9] Eskandarpour, R. and Khodaei, A., “Machine Learning Based Power Grid Outage Prediction in Response to Extreme Events,” *IEEE Trans. Power Syst.*, 32(4), pp. 3315–3316 (2017).

[10] Wanik, D. W., Anagnostou, E. N., Hartman, B. M., et al., “Storm outage modeling for an electric distribution network in Northeastern USA,” *Nat. Hazards*, 79(2), pp. 1359–1384 (2015).

[11] Doostan, M. and Chowdhury, B. H., “Power distribution system fault cause analysis by using association rule mining,” *Electr. Power Syst. Res.*, 152pp. 140–147 (2017).

[12] Cai, Y. and Chow, M. Y., “Cause-effect modeling and spatial-temporal simulation of power distribution fault events,” *IEEE Trans. Power Syst.*, 26(2), pp. 794–801 (2011).

[13] Chow, M. and Taylor, L. S., “Analysis and prevention of animal-caused faults in power distribution systems,” *IEEE Trans. Power Deliv.*, 10(2), pp. 995–1001 (1995).

[14] Warlyani, P., Jain, A., Thoke, A. S., et al., “Fault Classification and Faulty Section Identification in Teed Transmission Circuits Using ANN,” *Int. J. Comput. Electr. Eng.*, 3(6), pp. 807–811 (2012).

[15] Xu, L., Chow, M.-Y., and Taylor, L. S., “Power distribution fault cause identification with
imbalanced data using the data mining-based fuzzy classification E-Algorithm,” *IEEE Trans. Power Syst.*, 22(1), pp. 164–171 (2007).

[16] Doostan, M. and Chowdhury, B. H., “Power distribution system equipment failure identification using machine learning algorithms,” *IEEE Pow. Energy Soc. Gen. Meet.*, pp. 1–5 (2018).

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

[18] Breiman, L., “Random forests,” *Mach. Learn.*, 45(1), pp. 5–32 (2001).

[19] Quinlan, J. R., “C5. 0 Data Mining Tool,” *RuleQuest Research, 63*, p.64. (1997).

[20] Quinlan, J. R., “C 4.5: Programs for machine learning,” *Morgan Kaufmann Ser. Mach. Learn. San Mateo, CA Morgan Kaufmann, c1993*, (1993).

[21] Hinton, G. and Van Camp, D., “Keeping neural networks simple by minimizing the description length of the weights,” *Proc. 6th Ann. ACM Conf. on Computational Learning Theory*, (1993).

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

[23] Liu, H. and Motoda, H., “Feature selection for knowledge discovery and data mining,” 454. Springer Science & Business Media, (2012).

[24] Guyon, I. and Elisseeff, A., “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, 3(Mar), pp. 1157–1182 (2003).
List of table captions

Table I. characteristics of distribution network

Table II. A common cost matrix

Table III. Customized cost matrix used in this paper

Table IV. Random forest-based predictor with customized cost matrix. Observed vs. predicted

Table V. Random forest-based predictor with default cost matrix. Observed vs. predicted

Table VI. C5.0 based predictor with customized cost matrix. Observed vs. predicted

Table VII. C5.0 based predictor with default cost matrix. Observed vs. predicted

Table VIII. Ensemble model output. Observed vs. predicted

List of figure captions

Fig.1. Distribution of root cause classes in dataset

Fig.2. Number of root causes of outages in different substations

Fig.3. Feature importance for random forest algorithm

Fig.4. Feature importance for C5.0 algorithm

Fig.5. Comparing predictors’ accuracy

Fig.6. Comparing predictors’ accuracy
List of tables

Table I. characteristics of distribution network

|                  |       |
|------------------|-------|
| Substations      | 64    |
| Feeders          | 491   |
| Feeder Length (km)| 10981 |
| Distribution Transformers | 26155 |
| Underground Posts | 1593  |
| Overhead Posts    | 23207 |

Table II. A common cost matrix

|                  | Class 1 | Class 2 | Class 3 |
|------------------|---------|---------|---------|
| Observed Class 1 | 0       | 1       | 1       |
| Observed Class 2 | 1       | 0       | 1       |
| Observed Class 3 | 1       | 1       | 0       |

Table III. Customized cost matrix used in this paper

|                  | Animal | Equipment | Other | Vegetation | Weather |
|------------------|--------|-----------|-------|------------|---------|
| Observed Animal  | 0      | 2         | 1     | 1          | 1       |
| Observed Equipment| 1     | 0         | 1     | 1          | 1       |
| Observed Other   | 0      | 0         | 0     | 0          | 0       |
| Observed Vegetation| 1    | 2         | 1     | 0          | 1       |
| Observed Weather | 1      | 2         | 1     | 1          | 0       |
Table IV. Random forest-based predictor with customized cost matrix. Observed vs. predicted

| Predicted          | Animal | Equipment | Other | Vegetation | Weather |
|-------------------|--------|-----------|-------|------------|---------|
| Observed Animal   | 72.73  | 25.76     | 0.00  | 1.52       | 0.00    |
| Equipment         | 7.45   | 83.28     | 0.00  | 2.29       | 6.99    |
| Other             | 7.26   | 79.49     | **1.28** | 3.85       | 8.12    |
| Vegetation        | 5.63   | 26.76     | 0.00  | **47.89**  | 19.72   |
| Weather           | 1.93   | 9.18      | 0.00  | 7.25       | **81.64** |

Table V. Random forest-based predictor with default cost matrix. Observed vs. predicted

| Predicted          | Animal | Equipment | Other | Vegetation | Weather |
|-------------------|--------|-----------|-------|------------|---------|
| Observed Animal   | **1.25** | 96.25     | 0.00  | 0.00       | 2.50    |
| Equipment         | 0.36   | **96.79** | 0.24  | 0.12       | 2.49    |
| Other             | 0.00   | 87.18     | **5.98** | 1.28       | 5.56    |
| Vegetation        | 0.00   | 61.40     | 5.26  | **10.53**  | 22.81   |
| Weather           | 0.00   | 30.10     | 1.02  | 1.02       | **67.86** |

Table VI. C5.0 based predictor with customized cost matrix. Observed vs. predicted

| Predicted          | Animal | Equipment | Vegetation | Weather |
|-------------------|--------|-----------|------------|---------|
| Observed Animal   | **72.73** | 27.27     | 0.00       | 0.00    |
| Equipment         | 3.09   | **90.95** | 0.69       | 5.27    |
| Other             | 4.27   | 81.20     | 2.14       | 12.39   |
| Vegetation        | 9.86   | 43.66     | **23.94**  | 22.54   |
| Weather           | 1.45   | 17.39     | 2.42       | **78.74** |
Table VII. C5.0 based predictor with default cost matrix. Observed vs. predicted

| Predicted | Animal | Equipment | Other | Vegetation | Weather |
|-----------|--------|-----------|-------|------------|---------|
| Animal    | 2.44   | 95.12     | 0.00  | 0.00       | 2.44    |
| Equipment | 0.22   | 95.73     | 0.66  | 0.11       | 3.29    |
| Other     | 0.00   | 92.48     | 3.54  | 1.33       | 2.65    |
| Vegetation| 0.00   | 70.00     | 6.67  | 15.00      | 8.33    |
| Weather   | 1.82   | 32.73     | 1.36  | 0.91       | 63.18   |

Table VIII. Ensemble model output. Observed vs. predicted

| Predicted | Animal | Equipment | Vegetation | Weather |
|-----------|--------|-----------|------------|---------|
| Animal    | 74.24  | 25.76     | 0.00       | 0.00    |
| Equipment | 2.29   | 91.07     | 1.03       | 5.61    |
| Other     | 2.56   | 83.33     | 2.14       | 11.97   |
| Vegetation| 5.63   | 39.44     | 39.44      | 15.49   |
| Weather   | 0.00   | 14.01     | 2.42       | 83.57   |
List of figures

Fig. 1. Distribution of root cause classes in dataset

Fig. 2. Number of root causes of outages in different substations
Fig. 3. Feature importance for random forest algorithm

Fig. 4. Feature importance for C5.0 algorithm
Fig. 5. Comparing predictors’ accuracy

Fig. 6. Comparing predictors’ accuracy
Biographies:

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.

Mohammad Javad Bordbari received the B.Sc. and M.Sc. degrees in electrical engineering from the Shiraz University of Technology and Shiraz University in 2015 and 2018, respectively. His current research interests include modeling home energy management systems, energy efficiency analysis in buildings, and power system reliability and resiliency studies.