IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received February 29, 2020, accepted April 12, 2020, date of publication May 6, 2020, date of current version May 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2991923

Predicting Lightning-Related Outages in Power Distribution Systems: A Statistical Approach

MILAD DOOSTAN[1] AND BADRUL CHOWDHURY[2], (Senior Member, IEEE)
[1]Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
[2]Energy Production and Infrastructure Center (EPIC), University of North Carolina at Charlotte, Charlotte, NC 28223, USA

Corresponding author: Milad Doostan (mdoostan@uncc.edu)

This work was funded by the Energy Production & Infrastructure Center (EPIC).

ABSTRACT This paper presents a novel data-driven approach for predicting lightning-related outages that occur in power distribution systems on a daily basis. In order to develop an approach that is able to successfully fulfill this objective, there are two main challenges that ought to be addressed. The first challenge is to define the extent of the target area. An unsupervised machine learning approach is proposed to overcome this difficulty. The second challenge is to adequately identify characteristics of lightning-related outages and to explore the relationship between these outages and weather-related variables (thunderstorm events). In this paper, these outages are clustered into a few manageable groups. Then, a probabilistic model is presented to estimate the likelihood of each group of outages. Finally, a machine learning classification algorithm that can handle the imbalanced problem is developed to predict what group will the outage belong to on a specific day in a specific area of the system under study. Actual outage data, obtained from a major utility in the U.S., in addition to radar weather forecast data are utilized to build the proposed approach. Also, three case studies are provided to show several issues associated with predicting lightning-related outages, and to demonstrate how the proposed approach can address those problems adequately.

INDEX TERMS Data analytics, lightning-related outage, machine learning classification, outage prediction, power distribution systems, statistical modeling.

I. INTRODUCTION

Lightning is a major cause of outages in power distribution systems [1]. Transient over-voltages caused by direct or indirect lightning strikes may inflict severe damages to the susceptible equipment and can produce detrimental effects on power quality and reliability of the system. With upward trends in extreme weather and climate events in recent years, the intensity and frequency of the lightning activities are expected to increase, leading power utilities to be confronted by a growing problem with regards to this weather-related phenomena [2]–[4].

In order to reduce the destructive effects of lightning on distribution systems, a common strategy is to implement a proper lightning protection design, i.e., installing surge protective devices and shielding wires [1]. While taking such preventive actions appear to be effective for protecting the system against severe damages, momentary outages caused by the activation of these protective devices can exert secondary effects on the system. On the other hand, should the lightning effect exceeds the designed levels of the protection system, permanent faults may occur, leading the system to experience a sustained interruption [5]. Therefore, besides implementing preventive measures, it is important for utilities to take an appropriate response to these outages either by identifying them immediately after they occur or by predicting them.

By looking at historical outage data, analyzing electrical characteristics of the system, and tracking lightning activity in an area immediately after an outage occur, the utility companies are able to identify and distinguish lightning-related outages from other causes. Moreover, multiple studies have been carried out to develop models for identifying this source of outages [6], [7]. Opposed to that, predicting lightning-related outages has been left relatively unattended. This might stem from the complex nature of this type of outages, lack of enough information pertaining to them, and shortcomings of classical mathematical models.

However, with the technological advancement in data gathering through the smart grid framework and due to

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

tremendous improvements in weather forecasting efforts, a massive amount of outage and weather data has become available in recent years. This could shed light on the different characteristics of lightning-related outages. Moreover, advanced data analytics techniques are now being developed and combined with mathematical methods, creating powerful tools that can noticeably enhance predictive abilities. Adopting these predictive approaches will enable effective and timely decision-making actions by operators as well as planners, ultimately improving operational integrity and resiliency of the system.

By this time, several studies have been carried out to explore different aspects of lightning-related outages and to ultimately predict some characteristics of these outages.

The authors in [8] propose an approach for estimating the number of wind and lightning-related outages combined together on a daily basis. In order to carry out this task, they develop a machine learning regression model based on an ensemble boosting algorithm. In [9], the authors propose a method for forecasting the cumulative number of outages during a storm condition. They first create empirical models for different types of storms and then develop an exponential model for the forecasting purpose. They employ the model to predict lightning-related outages that may have occurred during several summer storms.

Another study [10] presents a Monte Carlo simulation model to study the reliability indices under lightning storm condition. Their model is based on the storm parameters and the outage rate. Therefore, first, the authors build a statistical model to explain the storm intensity and duration and then utilize that model as well as a data-driven approach to calculate the lightning-related outage rate. The performance of the proposed model is evaluated by conducting a case study using data collected from lightning storm weather conditions that occurred in an area in the Midwest United States in the span of five years. In [1], the authors carry out an experimental study to investigate significant factors that influence the frequency of lightning strike flash-overs. Moreover, they develop a probabilistic model for estimating the number of lightning-related outages on an annual basis.

The authors in [11], [12], and [13] present comprehensive studies on predicting hurricane and storm-related outages by addressing important issues such as the presence of a high imbalance in the response variable, engineering informative predicting variables, and building multi-stage models. It is worth mentioning that the focus of the aforementioned studies is not necessarily on lightning-related outages but rather on storm-related that could include various causes of outages (and their combination) such as wind, vegetation, and lightning, to name a few.

Even though some approaches, including [1], [8], [9], [10], are proposed to predict the rate or trend of lightning-related outages, they could be identified with several shortcomings. As a matter of fact, a majority of these approaches are developed based on the combination of weather-related outages and therefore cannot to be used for predicting only lightning-related outages. Moreover, they mostly consider the entire distribution system under study for the prediction task; hence, do not provide the ability to make the prediction for a specific area within the system. Furthermore, a majority of these approaches make the prediction for a long-term horizon (i.e., yearly) and consequently cannot be utilized for making predictions on a short-term horizon (i.e., daily or weekly). Additionally, most of these approaches focus on estimating the outage rate, which is usually defined as $outages/100\ km/year$. While estimating this rate is extremely useful for planning purposes, it would not provide much helpful information for taking short-term proactive measures. Last but not least, approaches that attempt to predict the exact number of outages are expected to deliver a poor performance when the number of outages is large. This argument would be fully supported later on.

In this paper, we propose a novel approach to predict lightning-related outages that occur in power distribution systems. In particular, we build an approach that is able to predict whether zero outage, one outage, or two or more outages will occur on a specific day that experiences thunderstorm events in a particular area in the system. The proposed approach overcomes the aforementioned shortcomings and provides the ability to make the prediction on a short-term horizon (i.e., daily basis) for a specific area within the service territory.

The proposed approach seeks to provide a meaningful knowledge about risks and locations of lightning-related outage problems. Hence, it can present a succinct view of the current system status to the operators, which enables effective and timely decision-making actions with regards to lightning-related problems. By providing a preliminary but accurate prediction, the proposed approach allows operators to effectively utilize advanced satellite imagery or sophisticated lightning detection systems to find the exact locations in the system that could have a high risk of a lightning-related outage. Moreover, the proposed approach enables electric utility companies to more intelligently allocate and dispatch crew members, as opposed to simply putting crews on alert when thunderstorms are expected, and dispatching them to wherever permanent outages require attention.

The main contributions of the proposed approach are summarized below:

1) We offer a workable solution to address the challenges posed by the extent of the prediction's target area.
2) We demonstrate that to obtain the best possible predictive performance, lightning-related outages should be clustered into a few manageable groups, which in this study, are three main groups, namely zero outage, one outage, and two or more outages.
3) We provide a probabilistic model to calculate the likelihood of each group of outages and validate the model using statistical tests.
4) We develop a machine learning classification algorithm that can handle the imbalanced problem and utilize it to predict the specific group of outages that will likely occur using the likelihood values as input.

The rest of this paper is organized as follows. In Section II, the outage and weather data used in the analysis is described. In Section III, an approach is presented to address the challenges brought about by the extent of the prediction's target area. In Section IV, a statistical analysis is provided to cluster the outages. In Section V, a probabilistic model is provided for calculating the likelihood of each group of outages. In Section VI, a machine learning algorithm is presented for predicting the outages. In Section VII, three case studies are provided to show some practical issues associated with predicting lightning-related outages, and to demonstrate the effectiveness of the proposed approach to address those issues. Finally, in Section VIII, conclusions are provided.

II. DATA DESCRIPTION

The input data for the proposed approach is obtained from two main sources: 1) historically recorded outages, and 2) radar weather forecasts. Outage data is collected by a major investor-owned utility company serving the southeastern US. The data includes information on the time and locations of sustained lightning-related outages that occurred around approximately 85 substations located in the states of North Carolina and South Carolina between the years 2010 and 2014. The data is comprised of almost 800 samples of outages. The radar weather forecast data is collected from several external sources for weather stations located close to power substations over the span of the aforementioned years and includes the number of thunderstorm events that occurred on a daily basis for each weather station. In order to calculate the number of thunderstorm events, the hourly weather forecast for the entire 24 hours is considered and the summation value of the logical variable that shows whether or not each hour might experience a thunderstorm is calculated. The logical variable is available in almost any weather forecast platform.

III. AGGREGATING SUBSTATIONS

As mentioned, one challenge with developing the proposed approach is defining the extent of the prediction target area. In fact, if one intends to point-predict the number of lightning-related outages at the location of any given substation, one might encounter serious difficulties. These difficulties lie in the fact that 1) the lightning can occur at the location of a substation; however, it may travel and hit a location that is at a large distance from the substation, and vice versa, 2) accurate radar weather forecasts may not be available at the exact location of each substation, and 3) the degree of randomness for the number of outages that occur for each substation is relatively large, making it difficult to predict.

In order to provide a workable solution to this challenge, we propose to aggregate substations and to build larger areas, in which, each area includes multiple substations and local weather stations, where the weather forecast for each substation is obtained from its closest weather station. Such clustering can solve the aforementioned issues because by aggregation, instead of examining a single substation and a single weather station, we examine a broader area that contains multiple weather stations, and calculate the average number of thunderstorm events in the area. By doing this, we obtain a better weather forecast. Moreover, if the lightning is created within the area, even if it travels, it is highly likely that it ultimately hits a point within the area. Also, aggregation reduces the considerable randomness in the data and helps to see the patterns more clearly. In addition to the aforementioned reasons, clustering substations based on their location and proximity has other justifications. In fact, location data could be an indicator of several important variables. For example, substations close to each other might have similar geographical characteristics, weather patterns, and etc. These are factors that affect the exposure and vulnerability of the system to thunderstorms and, hence, it would be reasonable to cluster substations based on proximity.

In order to define the aforementioned areas, k-means clustering algorithm is utilized in this paper. This algorithm is a widely used unsupervised machine learning algorithm, which aims at clustering a given dataset into a certain number (k) of groups. In this paper, this algorithm is used to cluster substations into different groups where each group represents an area. The main idea of this algorithm is to define k centroids at random, one for each cluster, and then to minimize the squared error function represented in (1) [14].

$$J(r, \mu) := \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} r_{ij} \left\| x_i - \mu_j \right\|^2 \tag{1}$$

where m is the number of data points, k is the number of clusters, r_{ij} is an indicator, which is 1 if, and only if, x_i is assigned to cluster j, x_i is data point, μ_j is the centroid for cluster j, and $\|.\|^2$ denotes the Euclidean distance. In this study, data points are locations of substations (approximately 85 data points), which are represented by latitude and longitude in a two-dimensional space.

One major challenge with this algorithm is the need to specify the number of clusters. In fact, there is no global theoretical method to find the optimal value of this parameter; however, a few approaches are common to deal with this problem. One most used approach is to run k-means clustering for a range of different k values and to calculate the aforementioned squared error function for each value. In this case, the error tends to decrease toward zero as k increases; however, after a certain k value is reached, the decrease in error would be very gradual. Therefore, analyzing different values of k and finding the aforementioned threshold could help in deciding a reasonable number of clusters [15].

Applying k-means clustering algorithm and using the aforementioned method to calculate the proper number of clusters for grouping substations will result in 14 areas. These areas define the extent of the prediction. Fig. 1, illustrates these areas. It is worth mentioning that although there are other clustering approaches, we believe that the approach adopted here is the most suitable for this study. In fact, since the goal is to cluster a two-dimensional data (i.e., latitude and longitude) and to work with distances, the k-means approach

TABLE 1. A sample of the dataset under study.

Time	Area	Thunderstorm Events	Number of Outages
2011-08-13	5	0	0
2012-08-22	3	3	1
2012-05-22	12	30	2
2012-09-02	8	26	3
2013-07-17	8	50	4
2014-07-03	12	38	5

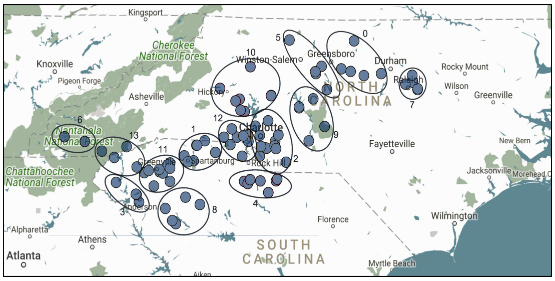

FIGURE 1. Demonstration of different areas.

makes the most sense. Moreover, considering the size of clustering data which is small, there would be no need for utilizing more sophisticated algorithms.

IV. CLUSTERING THE OUTAGES

By conducting various analyses we have reached the conclusion that predicting the exact number of outages in days that have thunderstorm events may not be possible. In fact, we argue that an increase in the number of thunderstorm events does not necessarily translate into an increase in the number of outages. In what follows we provide statistical support to our argument. In particular, we postulate that lightning-related outages may be clustered into a few manageable groups. In this study, these groups are 1) zero outage, 2) one outage, and 3) two or more outages.

Before providing the rationale behind our argument, we first look at a sample of the dataset under study. Table 1 shows a sample of the dataset that includes six observations. As seen, each observation is associated with four attributes of time, area, number of thunderstorm events, and the number of outages. The complete dataset contains the aforementioned information for all fourteen areas shown in Fig. 1 for the years 2010 to 2014 on a daily basis.

The research question that we are examining is the relationship between the number of thunderstorm events and the number of outages. The maximum number of outages which has been recorded in the available data is five. As a result, the number of outages has a limited number of outcomes and therefore could be considered as an ordered categorical variable. This consideration is reasonable because it would be highly unlikely that an area experiences a large number of lightning-related outages on a specific day.

In order to investigate the relationship between the aforementioned variables, and especially to realize whether or not there is a difference between the number of thunderstorm events among different values of outages (including days with zero outage), we carry out a one-way ANOVA test. The ANOVA may be used to determine whether there are any statistically significant differences between the means of two or more independent groups regarding a specific explanatory variable [16]. In this study, the null hypothesis in the test would be that the mean number of thunderstorm events is the same across different values of outages. On the other hand, the alternative hypothesis would be that at least one pair of mean values are different from each other.

TABLE 2. Results of ANOVA.

Source	DF	SS	MS	F	P
Regression	5	389018	77804	403	<2e-16
Residuals	20126	3885367	193		

This analysis is conducted and the results are provided in Table 2. It is worth mentioning that the necessary analysis is carried out to make sure that the ANOVA assumptions [16] hold for this study and the available data can be analyzed using this test. According to the table, the *p*-value is almost zero; hence, we can reject the null hypothesis in favor of the alternative hypothesis and conclude that the average number of thunderstorm events is not equal for different values of outages.

A follow-up question, which could shed more light on the differences between values of outages with regards to the number of thunderstorm events is to investigate the difference in a pairwise manner and to quantify it. To carry this out, we conduct a post-hoc test, Tukey's HSD. This test allows answering which means are different and by how much and whether or not the difference between outages in a pairwise manner is statistically significant. It is worth mentioning that since the ANOVA and Tukey's HSD are very well-established methods whose formulations are readily available [16], their details are not discussed in this paper.

The results of the Tukey's HSD are provided in Table 3. The pair column shows the combination of days with two different numbers of outages. The difference column represents the differentiation between the average number of thunderstorm events between pairs. The lower and upper columns demonstrate the limits of the 95% confidence interval for the difference in the average value, and finally, the *p*-value shows the results of the hypothesis that there is not any statistically significant difference in the population of the average number of thunderstorm events for each pair. A *p*-value of less than 0.05 shows that the results are significant.

According to the table, it can be inferred that for days with zero outage, compared to days with one or more outages, the average number of thunderstorm events is smaller and the difference is statistically significant. This is because the confidence interval does not include zero (i.e., equivalently, the *p*-value is zero). The same argument could be made for days with one outage compared to other days. However, for days with two or more outages, as seen in the table, the confidence interval ranges from negative to positive

TABLE 3. Results of Tukey's HSD test.

Pair	Difference	Lower	Upper	p-value
1-0	28.8	26.2	31.5	0.0
2-0	40.9	35.5	46.3	0.0
3-0	43.7	36.3	51.0	0.0
4-0	45.6	36.3	55.0	0.0
5-0	51.7	39.1	64.2	0.0
2-1	12.1	6.1	18.1	0.0
3-1	14.9	7.0	22.7	0.0
4-1	16.8	7.1	26.5	0.0
5-1	22.8	10.0	35.6	0.0
3-2	2.7	-6.4	11.9	0.9
4-2	4.7	-6.1	15.5	0.8
5-2	10.7	-2.9	24.4	0.2
4-3	1.9	-9.9	13.8	0.9
5-3	8.0	-6.5	22.5	0.6
5-4	6.0	-9.6	21.7	0.9

values (i.e., it includes zero), and the p-value is greater than 0.05. As a result, we fail to reject the hypothesis that there is no difference between these pairs with regards to the number of thunderstorm events and therefore can conclude that there is sufficient evidence that the average number of thunderstorm events is the same for days that have two or more outages.

The aforementioned analysis creates the foundation for clustering the outages. Considering the facts that 1) there is a causal relationship between thunderstorm events and lightning outages and that 2) there is a significant difference between the days with zero outage compared to other days, and days with one outage compared to other days, and 3) the observation that days with two or more outages do not show distinguishable characteristics with each other with regards to the number of thunderstorm events, we cluster the outages into three groups: 1) zero outage, 2) one outage, 3) two or more outages.

Considering the aforementioned clustering, the ultimate objective would be to predict which group of outages will occur on a certain day in a specific area. Such clustering is necessary to obtain the best possible predictive performance. Therefore, we believe models that attempt to predict the exact number of outages (some of which were mentioned in the literature review), are expected to deliver a low degree of accuracy especially when the number of outages is large. A quantitative result is provided in Section VII for this argument.

V. LIKELIHOOD OF OUTAGES

Based on the preceding methodology, we concluded that during days with thunderstorm events, there could be three possible outcomes with regards to the number of outages. The next step would then be to determine the likelihood of the occurrence of each group of outages on a given day with a given number of thunderstorm events at a specific area. For this purpose, we propose that the binomial distribution model would be an appropriate model to calculate the likelihood of lightning-related outages. The rationale behind our proposed model is three-fold:

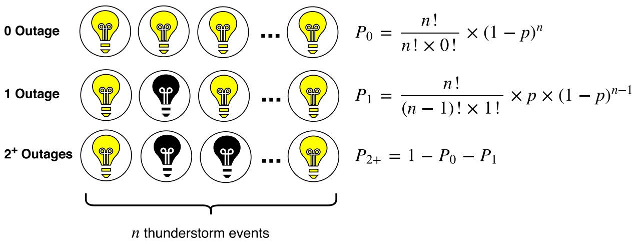
0 Outage
1 Outage
2+ Outages
$P_0 = \frac{n!}{n! \times 0!} \times (1-p)^n$
$P_1 = \frac{n!}{(n-1)! \times 1!} \times p \times (1-p)^{n-1}$
$P_{2+} = 1 - P_0 - P_1$
n thunderstorm events

FIGURE 2. Calculating the likelihood of groups of outages using binomial probability model.

1) Each thunderstorm event results in one of two possible outcomes (outage or no outage). This is confirmed by the data.
2) The probability of the outage is the same for each thunderstorm event (because it depends on the geographical characteristics of the area which is expected to more or less remain the same)
3) The thunderstorm events are independent, meaning that the fact that a thunderstorm event results in an outage does not impact the probability of an outage in another thunderstorm event. We assume that after an outage, the responsible dispatched crew is able to repair the protective devices (such as fuses and surge arresters) that were impacted and therefore they will operate as expected. It is worth mentioning in cases that the outage is adequately dealt with automatic reclosing devices, crews need not be dispatched and the system will continue operating normally.

The aforementioned properties indicate that the assumption of a binomial model holds; therefore, it is a valid candidate model for our purpose at hand.

Using the binomial model, one may calculate the likelihood of the occurrence of each group of outages. In order to clarify this, suppose that the weather forecast for the next day for a specific area demonstrates a total number of n thunderstorm events. Let's assume, by using the historical data, we have realized the probability that a thunderstorm event leads to an outage in that area is p. Considering this information, the likelihood of having no outage, having exactly one outage, and finally having two or more outages can be calculated as illustrated in Fig. 2.

In order to quantitatively examine whether or not the binomial model is an appropriate model, we devise a hypothesis test. The null hypothesis would be that the occurrence of outages arises from a binomial model with the probability of p. The alternative hypothesis would be that the data does not come from a binomial distribution. We could base the test on the differences between the observed and expected numbers of outcomes. This could be carried out by the Chi-square test [17]. It is worth mentioning that the critical value of the test is chosen based on 99% confidence interval.

The Chi-square values for all possible values of thunderstorm events (i.e., n) are calculated and plotted in Fig. 3. For a significant majority of the number of thunderstorm

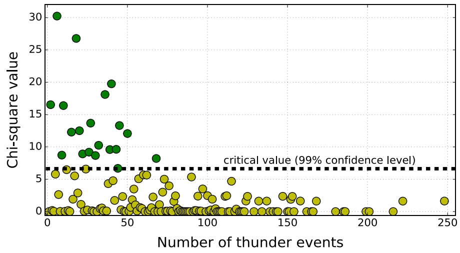

FIGURE 3. Results of the Chi-square test for goodness of fit for binomial probability model.

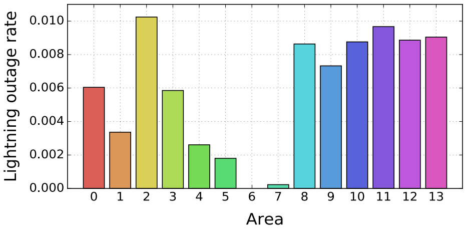

FIGURE 4. P_{outage} values for different areas within the system.

events, the Chi-square values fall below the critical value. This demonstrates that the assumption of the binomial probability model is reasonable. In fact, only 12.8% data points fall out of the range for the 99% confidence level and especially the cases with a high number of thunderstorm events are all represented accurately, with only one data point with over 50 thunderstorm events falling out of the 99% confidence level area. We can now, with great confidence, argue that the binomial distribution would be an adequate model to find the likelihood of groups of outages (i.e., 0, 1, 2+) occurring given a certain number of thunderstorm events.

In order to calculate the likelihood values, we follow the following procedure. First, for a given day and given area, we calculate the number of thunderstorm events, n, from the weather data. This was explained in Section II. Then, by using the historical data, we calculate the probability that a thunderstorm event leads to an outage, p, for each area. In order to compute this value, we calculate the total number of outages that occurred in each area divided by the total number of thunderstorm events experienced by that area. This probability value is called P_{outage} from this point and is demonstrated in Fig. 4. As seen in the figure, P_{outage} for some areas is considerably higher compared to others. This may be explained by the geographical characteristics of that area and its exposure to lightning strikes or the lightning protection level of each area. We believe P_{outage} can summarize such information into a single number. By knowing these values and employing the binomial probability model as illustrated in Fig. 2, we can calculate the likelihood of each group of outages.

VI. PREDICTING OUTAGES

By calculating the likelihood of groups of outages for a given day, weather condition, and a given area in the system using the aforementioned methodology, one may make a final prediction on what group of outages will occur. It should be noted that due to small P_{outage} values, the likelihood values obtained for two or more outages group would be generally smaller than one outage group, and for one outage would be smaller than zero outage group (albeit small, it could be critical). Therefore, simply using the largest likelihood value for making the prediction would be impractical.

In order to provide an appropriate means of predicting which outage group will occur, we define the problem as a machine learning multiclass classification problem. In fact, by using the binomial model, we can calculate three likelihood values for zero outage, one outage, and two or more outages. The actual group of outages is also known from the historical data. As a result, we would have a supervised machine learning problem, in which the variables are the likelihood values, and the label is the class of outages (i.e., 0, 1, 2+).

With the aforementioned context, the main objective here would be to predict different classes of outages correctly while the minimum number of alarms is issued. An alarm is issued when the model predicts either one outage or two or more outages. Since the occurrence of lightning-related outages is not very frequent, the majority of the alarms turn out to be false. Therefore, it is crucial to build a classifier that minimizes the false alarm ratio while enabling the outage instances, especially, two or more outages to be detected correctly as much as possible. Hence, the metric that we use to build and evaluate our classifier would be the outage detection rate. This metric, which is also known as recall value, is defined as $\frac{tp}{tp+fn}$. For each class of outage, the tp is the number of true positives (i.e., outages detected correctly) and fn is the number of false negatives (i.e., misclassified outages).

Therefore, there is a trade-off between the outage detection rate score for different classes. In other words, if one intends to predict all of the two or more outages correctly (i.e., maximizes the outage detection rate for that class), one might get less accurate results on one outage class and one needs to issue a great number of false alarms (i.e., outage detection rate for two other classes decreases). On the other hand, one can obtain very good results for one outages; however, one might see an increase in false alarms for zero outages and two or more outages. In order to deal with this trade-off problem, we suggest setting different threshold values for the outage detection rate of different classes. For example, we could assume that the classifier should be able to deliver an outage detection rate of greater than 0.7 for two or more outages and a detection rate of greater than 0.5 for one outage. These values could be customized by the user.

One challenge with regards to the classification problem defined here is the presence of imbalanced classes. The majority class of outages is zero outages. The occurrence of one outage class is significantly smaller and the occurrence of two or more outages is rather infrequent. While the occurrence of one or two or more outages is considerably small, they are of interest to the utility company, and therefore an appropriate model should be able to identify them correctly as much as possible. Such a difference between the occurrence of classes is known as an imbalanced problem. Imbalanced class distribution of a data set is problematic as it can result in biased predictions and misleading accuracy for most classification learners [18]. A quantitative result is provided in Section VII to demonstrate the impact of the imbalanced problem.

In order to address the imbalanced problem, a variety of methods has been proposed. These methods could be categorized under three well-established approaches of data-level, algorithm-level, and cost-sensitive learning [18]. In this study, we employ the algorithm-level approach to tackle the imbalanced problem. In this approach, a bias is introduced in the objective function of classifiers to give different weights to the majority and minority classes. In fact, the level of imbalance is very significant; as a result, data-level approach, especially generating synthetic data, won't be practical. Moreover, we would like to develop a classifier that has the ability to be customized by the user. In other words, we would like the user to have the ability to give customized importance to different classes with a flexible degree and desired outage detection rate. This is perfectly possible in the algorithm-level approach.

In order to carry out the classification task, we use the logistic regression as our baseline model. Logistic regression is among the most well-established classifier algorithms. While there could be several other choices; considering the size of the data set and the number and type of features, logistic regression would suffice for this problem. Especially, the loss function of the logistic regression could be easily modified to tackle the imbalanced problem. It is worth mentioning that some other models such as neural networks have the same ability and therefore the analyst should examine which algorithm suits the problem the best.

In logistic regression, the assumption is that all classes (i.e., primarily two classes) are equally important and hence have the same weight (i.e., importance) and the objective is to minimize a log loss function (formulation available in [15]). However, in a weighted logistic regression, the importance of the classes is different and therefore different classes have different weights associated with them. Considering the weight, w, the generic log loss function of the logistic regression can be re-written as (2):

$$logLoss = -\sum_{i=0}^{n-1}[w \cdot y_i \cdot log(f(x_i)) + (1-w) \cdot (1-y_i) \cdot log(1-f(x_i))] \tag{2}$$

where, y_i is the actual class, $f(x_i)$ is the predicted class, and n is the number of observations.

It should be noted that logistic regression is a binary classifier, meaning that it cannot handle target vectors with more than two classes. To make the multi-class classification possible, the logistic regression should be used in a procedure known as one-vs-all [19]. Moreover, in order to avoid the over-fitting problem, the $L2$ regularization terms, with the rate of λ [14], should be added to the aforementioned loss function.

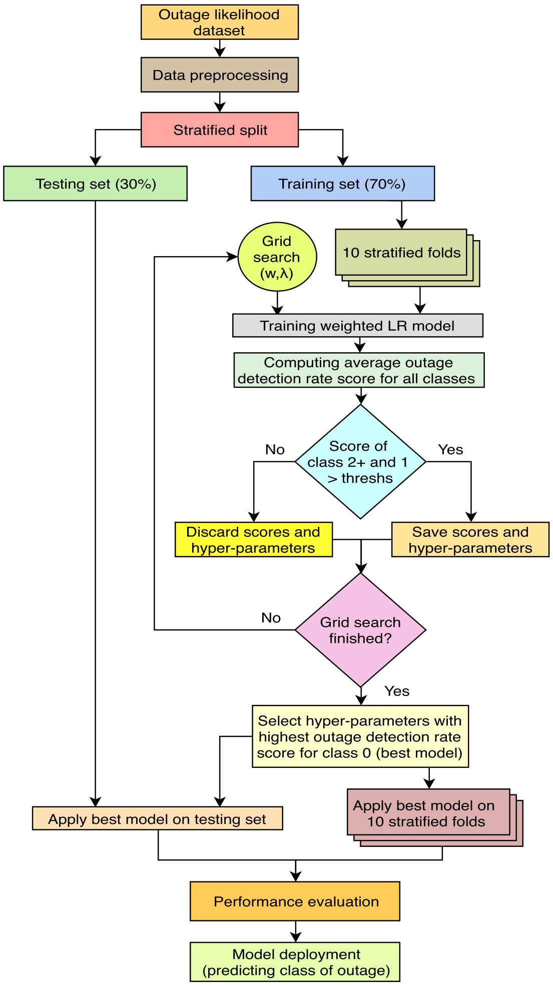
Outage likelihood dataset
Data preprocessing
Stratified split
Testing set (30%)
Training set (70%)
Grid search (w,λ)
10 stratified folds
Training weighted LR model
Computing average outage detection rate score for all classes
No
Score of class 2+ and 1 > threshs
Yes
Discard scores and hyper-parameters
Save scores and hyper-parameters
No
Grid search finished?
Yes
Select hyper-parameters with highest outage detection rate score for class 0 (best model)
Apply best model on testing set
Apply best model on 10 stratified folds
Performance evaluation
Model deployment (predicting class of outage)

FIGURE 5. Procedural flowchart of the proposed classifier.

The weighted logistic regression would be the cornerstone of our classifier. However, to build a robust model, general steps such as data pre-processing, creating training and testing sets, tuning hyper-parameters through cross-validation, etc. have to be performed as well. The complete procedure for building the classifier is demonstrated in the flowchart shown in Fig. 5. Several points should be made regarding the procedure as follows:

1) Data pre-processing includes handling missing data and outliers and normalizing the data, if necessary.
2) The data is split into training and testing sets in a way that the proportion of classes in both sets is similar (A.K.A., stratified splitting)
3) The hyper-parameters of the logistic regression model include three weights (w) for classes as well as regularization rate (λ), and are tuned through cross-validation.

TABLE 4. Results of case study 1.

Model / Outage	0	1	2	3	4	5
RF	0.99	0.03	0.2	0.25	0.2	0
NB	0.97	0.2	0.26	0.25	0	0
LR	0.94	0.39	0.27	0.25	0	0

4) The threshold value for the outage detection rate for different classes could be customized by the user to satisfy the desired outage detection rates. If the outage detection rate for the critical classes (i.e., two or more outages and one outage) is greater than the desired thresholds, the hyper-parameters would be stored.
5) The best model is the one which satisfies the desired thresholds for the critical classes and has the highest outage detection rate for class 0.
6) The ability of the model to generalize is evaluated using k-fold cross-validation as well as its performance on the testing set.

The proposed algorithm allows predicting outages with a desired detection rate for critical classes. The effectiveness of the proposed approach is demonstrated through a case study (third case study) in the next section.

VII. CASE STUDIES

In order to quantitatively show the practical issues that were discussed with regards to predicting lightning-related outages and to demonstrate the effectiveness of the proposed approach, we will provide three case studies as follows.

A. CASE STUDY 1 (BENCHMARK RESULTS)

As explained earlier, clustering outages into a few manageable groups (three groups in this study) seems necessary to obtain the best possible predictive performance. In fact, we showed that attempts to predict the exact number of outages could lead to a low degree of performance especially when the number of outages is large. We, also, argued that this problem exacerbates because of the imbalanced problem.

In order to show the impact of the aforementioned issues, we will carry out a case study. In case study 1, we do not cluster the outages; additionally, we skip the proposed probability model for calculating the likelihood of outages. We utilize three well-known machine learning classifiers: Random Forest (RF), Naive Bayes (NB), and Logistic Regression (LR) (not weighted) and feed them two main inputs of area number (categorical variable) and the number of thunderstorm events on a daily basis. We, then, tune necessary hyper-parameters through 10-fold cross-validation. The objective is to predict the exact number of outages (i.e., zero to five). In order to explore the performance of the model, we again utilize a 10-fold cross-validation procedure and obtain the average outage detection rate for each class (i.e., zero to five in this case study), in which the results are provided in Table 4.

The results clearly highlight the impact of the aforementioned issues. In fact, as seen, the outage detection rate for large values of outages is very low, even in some cases is zero. Moreover, due to the imbalanced problem, the models are biased toward the majority class (i.e., zero outage) and therefore deliver very low outage detection rate for minority classes (i.e., outage instances).

TABLE 5. Results of case study 2.

Model / Outage	0	1	2+
RF	0.99	0.15	0.36
NB	0.96	0.31	0.40
LR	0.96	0.17	0.53

B. CASE STUDY 2 (IMBALANCED PROBLEM)

In this case study, we cluster outages to three groups (i.e., 0, 1, 2+) and we calculate the likelihood values for each group of outages. We, then, utilize the likelihood values and feed them to the three aforementioned classifiers to predict the class of outages. However, in order to show the impact of the imbalanced problem, we do not take any action to deal with that problem. Tuning the hyper-parameters and assessing the performance of the model are carried out through the 10-fold cross-validation again. The results (i.e., outage detection rates for three classes) are provided in Table 5.

As seen in the table, clustering the outages improves the outage detection rates compared to the first case study. However, still, the models are biased toward the majority class (i.e., zero outage) and therefore deliver poor results for one outage and two or more outage classes.

C. CASE STUDY 3 (PROPOSED APPROACH)

In this case study, we implement the proposed approach and demonstrate its success in addressing the aforementioned issues. We again cluster outages to three groups (i.e., 0, 1, 2+) and we calculate the likelihood values for each group of outages. We, then, utilize the likelihood values (i.e., outage likelihood dataset) and feed those to the proposed classification algorithm illustrated in Fig. 5. The threshold values that we consider for our most important class (i.e., two or more outages) is 0.85 and for our second important class (i.e., one outage) is 0.55. This means that we tune the weights such that we make sure to obtain those outage detection rate values on the cross-validation. The performance of the model (outage detection rates) is also evaluated on 10-fold cross-validation as well as on the testing set (30% of the whole data), where the results are provided in Table 6.

As seen in the table, by optimally tuning the weight values (w) for different classes, we are able to obtain outage detection rates that satisfy defined threshold values. As mentioned, there is a trade-off between outage detection rate values of different classes. In this case study, we placed the highest importance to two or more outage class (i.e., outage detection rate of 0.85) and lower importance to one outage class. As a result, some of the one outage instances are misclassified in favor of two or more outages, as the model is intentionally biased towards two or more outages, which represent the severest of outages.

TABLE 6. Results of case study 3.

Set / Outage	0	1	2+
CV	0.82	0.57	0.85
Test	0.81	0.55	0.86

One important observation that demonstrates the remarkable performance of the proposed approach is the outage detection rate obtained for zero outage instances. In fact, even though the classifier is intentionally biased to outage instances, it is able to detect zero outage observations with a high score of 0.82. This means that the number of false alarms issued by the model is significantly small. Another observation that proves the superior performance of the proposed approach is high outage detection rates that are obtained on the testing set (unseen data while developing the model). The values obtained on the testing set are significant and very similar to those obtained on cross-validation, demonstrating that the model is tuned properly. This indicates that the model is not over-fitting or under-fitting and is able to generalize very well.

VIII. CONCLUSION

A data-driven approach was proposed for predicting lightning-related outages in power distribution systems on a daily basis. Based on this study, the following conclusions can be drawn.

1) In order to develop a practical approach, records of outages and weather-related factors (thunderstorm events) should be obtained and processed.
2) A key step in building a realistic approach is to adequately define the extent of the predictions' target area. Aggregating substations and creating broader geographical areas by using clustering algorithms seems a workable solution for this purpose.
3) In order to obtain the best possible predictive performance, lightning-related outages should be categorized into a few manageable groups.
   These groups exhibit distinguishable characteristics with regards to the number of thunderstorm events.
4) To find the likelihood of groups of outages (i.e., 0, 1, 2+) given a certain number of thunderstorm events and a specific area in the system, the binomial probability model is an adequate model.
5) An important issue that should be addressed to build a successful predictive model is the imbalanced problem. The weighted logistic regression model can handle this problem and can deliver an appropriate classification of different groups of outages.

Although many different pieces of information pertaining to lightning-related outages were examined in this study, we did not account for all possible factors due to lack of access to related data. Therefore, the performance of the proposed approach may be improved by the inclusion of additional climatological and geographical information (e.g., satellite data for more accurate identification on thunderstorm events, or some electrical characteristics of lightning and its intensity, or protection level of the system). In fact, all the advantages of the proposed approach are built upon generic outage data collected by utilities, and typical daily weather forecast data, which is publicly available. This fact makes the implementation of the approach easily attainable within a great level of performance. It should be noted that the results presented in this study are system dependent. Different distribution systems may experience different patterns of lightning-related outages. Nevertheless, the proposed approach can be applied to any distribution system.

REFERENCES

[1] T. Miyazaki and S. Okabe, "Experimental investigation to calculate the lightning outage rate of a distribution system," *IEEE Trans. Power Del.*, vol. 25, no. 4, pp. 2913–2922, Oct. 2010.

[2] P.-C. Chen and M. Kezunovic, "Fuzzy logic approach to predictive risk analysis in distribution outage management," *IEEE Trans. Smart Grid*, vol. 7, no. 6, pp. 2827–2836, Nov. 2016.

[3] D. M. Ward, "The effect of weather on grid systems and the reliability of electricity supply," *Climatic Change*, vol. 121, no. 1, pp. 103–113, Nov. 2013.

[4] M. Panteli and P. Mancarella, "Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies," *Electr. Power Syst. Res.*, vol. 127, pp. 259–270, Oct. 2015.

[5] A. Piantini and J. M. Janiszewski, "Lightning-induced voltages on overhead lines—Application of the extended rusck model," *IEEE Trans. Electromagn. Compat.*, vol. 51, no. 3, pp. 548–558, Aug. 2009.

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

[7] L. Xu, M.-Y. Chow, J. Timmis, and L. S. Taylor, "Power distribution outage cause identification with imbalanced data using artificial immune recognition system (AIRS) algorithm," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 198–204, Feb. 2007.

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

[9] D. Zhu, D. Cheng, R. P. Broadwater, and C. Scirbona, "Storm modeling for prediction of power distribution system outages," *Electr. Power Syst. Res.*, vol. 77, no. 8, pp. 973–979, Jun. 2007.

[10] N. Balijepalli, S. S. Venkata, C. W. Richter, R. D. Christie, and V. J. Longo, "Distribution system reliability assessment due to lightning storms," *IEEE Trans. Power Del.*, vol. 20, no. 3, pp. 2153–2159, Jul. 2005.

[11] D. Cerrai, D. W. Wanik, M. A. E. Bhuiyan, X. Zhang, J. Yang, M. E. B. Frediani, and E. N. Anagnostou, "Predicting storm outages through new representations of weather and vegetation," *IEEE Access*, vol. 7, pp. 29639–29654, 2019.

[12] S. Shashaani, S. D. Guikema, C. Zhai, J. V. Pino, and S. M. Quiring, "Multi-stage prediction for zero-inflated hurricane induced power outages," *IEEE Access*, vol. 6, pp. 62432–62449, 2018.

[13] 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.

[14] A. Smola and S. Vishwanathan, *Introduction to Machine Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[15] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed. Cambridge, MA, USA: MIT Press, 2009.

[16] B. Shahbaba, *Biostatistics With R: An Introduction to Statistics Through Biological Data*. New York, NY, USA: Springer, 2012.

[17] N. Balakrishnan, V. Voinov, and M. Nikulin, *Chi-Squared Goodness of Fit Tests With Applications*. Boston, MA, USA: Academic, 2013.

[18] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognit.*, vol. 40, no. 12, pp. 3358–3378, Dec. 2007.

[19] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, Dec. 2004.

MILAD DOOSTAN received the B.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2013, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of North Carolina at Charlotte, Charlotte, NC, USA, in 2015 and 2019, respectively. His research interests include statistics, forecasting, and machine learning.

BADRUL CHOWDHURY (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 1987. He is currently a Professor with the Department of Electrical and Computer Engineering with a joint appointment with the Department of Systems Engineering and Engineering Management, University of North Carolina at Charlotte (UNC Charlotte), Charlotte, NC, USA. He is also the Assistant Director of the strategic initiatives and faculty liaison with the Energy Production and Infrastructure Center, UNC Charlotte. Prior to joining UNC Charlotte, he spent 14 years as a Professor of electrical and computer engineering at the Missouri University of Science and Technology. His current research interests include power system modeling, analysis and control, and renewable and distributed energy resource modeling and integration in smart grids. He is the Chair of the Charlotte Chapter of the IEEE Power and Energy Society.