IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 September 2023, accepted 8 October 2023, date of publication 13 October 2023, date of current version 19 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324609

RESEARCH ARTICLE

State of Risk Prediction for Management and Mitigation of Vegetation and Weather Caused Outages in Distribution Networks

RASHID BAEMBITOV, (Graduate Student Member, IEEE), AND MLADEN KEZUNOVIC, (Life Fellow, IEEE)
Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA

Corresponding author: Rashid Baembitov (bae_rashid@tamu.edu)

This work was supported in part by the Texas A&M Engineering Experiment Station Smart Grid Center.

ABSTRACT The paper proposes a novel approach for the outage State of Risk (SoR) assessment caused by weather and vegetation in the distribution grid. Machine Learning prediction algorithm is used in conjunction with GIS application for mapping the SoR for the entire network. The proposed optimization approach leads to the specification of the mitigation strategies that utility staff and customers can coordinate to minimize the impact of outages. The resulting SoR assessment enables the implementation of an innovative decision-making solution for utility operators, represented in the form of risk maps. Additionally, utilizing the SoR assessments, a Customer Notification System (CNS) is introduced to enhance customer awareness and facilitate the adoption of mitigation measures. This holistic approach shifts outage management from a reactive process to a proactive initiative, promoting grid resilience and reliability through planned outage mitigation.

INDEX TERMS Customer notification, machine learning, outage mitigation, outage prediction, state of risk.

I. INTRODUCTION

The outages in the electric system impose significant losses to the economy as well as a major non-monetary detrimental societal impact. It has been reported that the population of the United States experiences more blackouts than in any other developed nation [1]. Based on the Electric Emergency Incident and Disturbance Reports (Form DOE-417) from The US Department of Energy (DoE), the annually affected load loss has increased more than 10-fold from 3247.6 MW/year to 39411 MW/year, and the number of affected customers has soared from 6 524 651 customers/year to 8 603 823 customers/year [2]. DoE gives an estimate of $150 to $164 billion per year as the annual cost of outages to the US economy [3], [4].

The research in outage loss estimation shows that a notification of the customers about the upcoming possible outage can reduce the outage costs by 25-70% [5], [6], [7]. The notification of the customers about the possible outage transforms the experience from an unexpected, forced event into a planned event with some assigned probability. Incorporation of the customer notification systems (CNS) into the operations of a utility company offers a unique way of limiting the losses from outages by allowing utility staff and customers to coordinate mitigation strategies ahead of time to reduce the outage impact. Additionally, if the operators receive a timely estimation of the current State of Risk (SoR) of the system, this may lead to better decision-making, in turn leading to improved resilience and power quality [8], [9]. Thus, the timely and precise prediction of the State of Risk (SoR) of outages is of utmost importance for limiting economic and societal losses and ensuring public safety.

Latest advances in Machine Learning (ML), and developments in remote sensing and weather forecasts, bundled with Geographic Information Systems (GIS), have paved the way for the proposed SoR prediction approach [10]. Incorporating data from diverse sources and combining it with historical data about the causes and location of

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

outages from a utility company allows the creation of datasets for SoR ML algorithm training and testing. The resulting SoR prediction, if integrated into utility daily operations, allows planning of a variety of mitigation actions (equipment replacement and repair, customer notification, network topology switching, back-up generator startup, etc.) aimed at reducing overall impact and curtailing losses [11], [12], [13], [14], [15], [16], [17].

ML ensemble models presented in [18], [19], [20], and [21] propose predicting outages in distribution networks resulting from catastrophic weather events. Analysis of network resilience is performed in [18] by employing predicted risk levels produced by the Naïve Bayes model [22]. An approach to distribution transformers (DT) outage prediction using Logistic Regression is proposed in [23]. Optimization of tree trimming scheduling based on predicted SoR is analyzed in [12]. The above-mentioned applications use short and long-time horizon weather forecasts as input to the prediction model. However, the uncertainty level in long-term weather forecasts increases with time [24]. This adversely affects the accuracy of predicted SoR.

Our contributions are: (a) identifying key data sources for outage SoR predictions, (b) introducing a new spatiotemporal approach for GIS data processing for predicting outage SoR using ML on historical data, and (c) formulating a novel mitigation optimization method that uses SoR predictions for determining best mitigation strategy for reducing the outage impacts on utility and enhancing customer satisfaction levels. We focus on day-to-day operations during severe weather that doesn't cause catastrophic damage to infrastructure. With the introduction of SoR prediction, outage management is transferred from reactive to proactive, allowing utility staff and customers to anticipate and prepare for a possible outage.

After Introduction, the network data import to GIS is explained in Section II. Section III focuses on data preprocessing and Section IV on ML model development. The optimized mitigation approach to minimizing the impact is given in Section V, and evaluation results are presented in Section VI. The conclusions and references are given at the end.

II. NETWORK DATA IMPORT TO GIS

A. ORGANIZING DATA IN GEODATABASE

The SoR prediction has two inherent dimensions: spatial and temporal. We use GIS ArcGIS Pro software to work with spatial aspects of the problem by utilizing tools from the Graphical User Interface [25]. We prioritized usage of python and arcpy [26] library in the proposed framework, whenever possible, for the following reasons: a) the history and order of data manipulation and the utilized tools are automatically preserved and logged, it can be readily changed, updated and re-run on the same types of datasets; b) computer code (as opposed to manual processing) has inherent scalability, so it is applicable for processing significant amounts of data in parallel; c) code is also more structured, which leaves significantly less room for human error when developing and documenting it.

Organizing Geodatabases (GDB) in ArcGIS is critical for efficient use. We created several databases connected to our project and defined feature datasets (FD) within geodatabases to logically place initial data and intermediate results [27].

The location of distribution feeders is encoded in utility-provided shapefiles. Once imported into GDB, the merge operation combines them into one feature class (FC). The advantage of organizing data into FD is that it ensures that the same coordinate system is used for all FC within the FD. All lines are placed into a single FD.

In the next step, we import locations of the outages into GDB based on the latitude and the longitude from the utility-provided data. As seen from a part of the network in Fig. 1, the initial outage locations do not intersect the feeders in some parts of the system (blue dots). That might occur for several reasons, such as insufficient accuracy during data acquisition. Nevertheless, we need to associate all the outages with a corresponding feeder segment. The Snap tool in ArcGIS is used: it moves each outage point to the closest feeder. The result is presented by red crosses.

The service area of interest is located near a major city in the US and spreads across 4 counties. We import shapefiles with counties' boundaries into GDB as a separate FD. After importing, we also merge them to have a FC representing all the counties. Counties FCs are usually used to clip bigger datasets or to cluster the processing of bigger datasets (for example, imagery datasets).

We used seven weather stations closest to the network. Since the number of weather stations (WS) in the area can change throughout the years, one needs to account for the possible addition of data in a new location near the area of interest or some of the data becomes available after being unavailable for certain periods. We used the WSs that had data for the entire period of interest and discarded the rest. Each row of data in the weather dataset has a timestamp, the name of WS where the measurement was taken, and

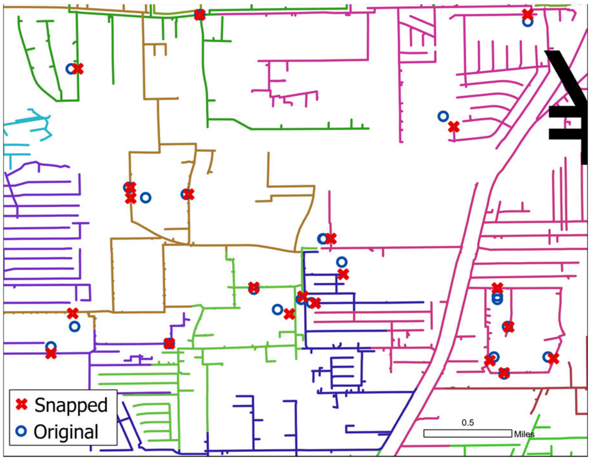

FIGURE 1. Adjusting fault location.

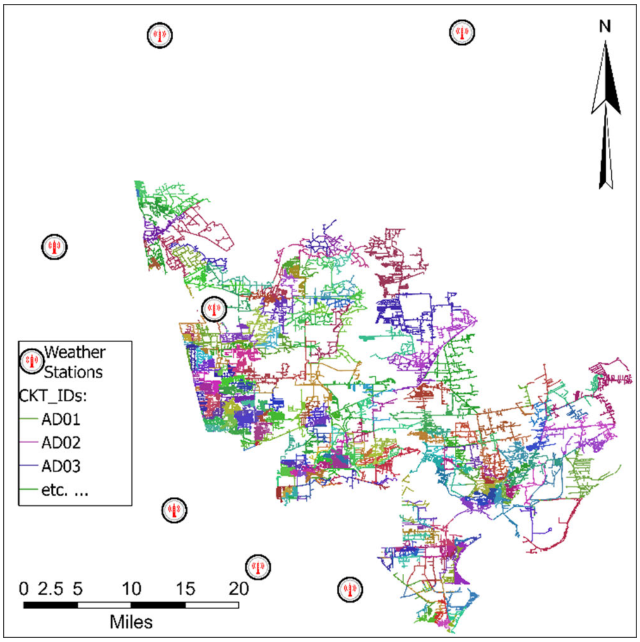

FIGURE 2. Location of weather stations.

WS coordinates. We extracted the coordinates of WS into separate files and then imported them into GDB as points. Fig. 2 shows the location of WS around the network.

B. CREATING BUFFERS FOR FEATURE EXTRACTION

The prediction object or the prediction entity needs to be defined. We are using several distribution feeder segments, grouped into clusters with the same circuit name identifier (CKTID) as a prediction entity. We are estimating and predicting risk levels for each CKTID in the system. The framework can be used on arbitrary prediction entity: from a single feeder segment to an area served by a single substation to even an entire distribution system.

One of the ways to spatially aggregate information about an object is the use of buffers, which are polygons created around geographical objects with a predefined distance from the object. They allow the extraction of the necessary information that belongs to a particular object from various datasets. In our study, we used different buffers around distribution feeders. An example of 20-, 100- and 500-meter buffers is shown in Fig. 3. We also generate buffers that are grouped by county. These buffers are used to process imagery datasets in several steps to decrease processing time and lower possible errors. Processing imagery in a single step is very computationally intense and unstable and prone to software errors.

C. ASSIGNING CKTIDS TO NULL FEEDER SEGMENTS

In the GIS data provided by the utility, not all the feeder segments had CKTID assigned to them, so we refer to these as NULL feeder segments. The probable reason for this is that there may have been new construction, and recently built feeder segments were imported into the database but were not assigned CKTIDs. We have created a procedure that allows us to assign missing CKTIDs, while ensuring the feeder connectivity. Specifically, CKTIDs are assigned to these segments based on their geographical proximity and their connection to feeders, particularly at the points where they touch. We note that an obvious approach of using spatial self-join with parameter "boundary touches" [28] yields inadequate results.

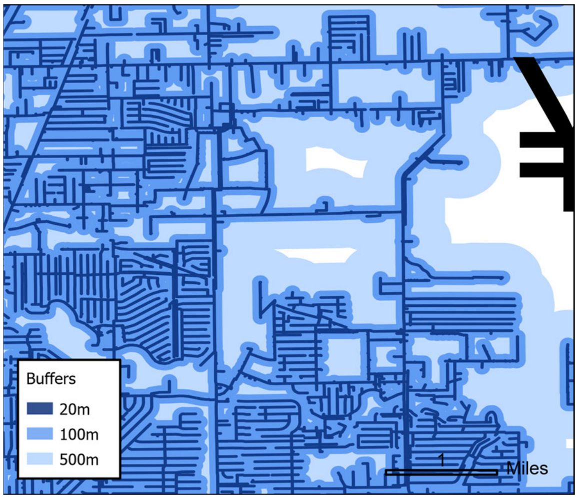

FIGURE 3. 20-, 100- and 500-meter buffers around power lines.

The suggested method is an iterative process of dissolving (or fusing together) feeder segments into bigger clusters, when they meet each other. First, the *NULL segments* touching known feeders are joined. Then, *NULL segments* that do not touch any known feeders are combined into *NULL clusters* based on their connectivity to each other. Next, these *NULL clusters* are assigned CKITIDs based on proximity to known feeders. In such a manner, the feeder segments "snowball" into clusters, growing from single elements to fully connected parts. The segments with the CKTIDs known are separated in the beginning and later used as a reference for assigning CKTIDs to NULL clusters. After all the segments have the CKTID assigned, the datasets are merged (combined) into a single dataset.

The disadvantage is that in some cases identical NULL clusters can be formed because each NULL cluster has several initial starting points. Since we do not have prior information on where and which segments can form a NULL cluster, we assume that any *NULL segment* can be a starting point for a *NULL cluster*. To overcome this problem, the identical clusters are removed at the end of the process. The proposed method can be enhanced by using some method for cluster center initialization, reducing the number of iterations and computational burden. In our case, 5 iterations were enough to cluster all the NULL segments.

We note the importance of the subpar GIS data on the quality of the SoR predictions. Small variations in the GIS placement of feeders, if they remain in proximity to their actual locations, might not have a drastic impact on the framework's performance. An exception would be mountainous regions where environmental conditions can vary significantly between peaks and valleys. In such territories, a higher degree of precision in mapping the power lines' locations is desired. Another issue is mislabeling of GIS objects. It leads to using incorrect geographical information by ML algorithms, which can result in the learning of inaccurate patterns, ultimately producing less reliable predictions. In our study, the method for NULL segments yields adequate results, since the number of NULL segments is low, and they are in the vicinity of known feeders.

III. DATA PREPROCESSING

A. DATA SOURCES

While the opportunities of incorporating datasets in the proposed framework are broad, the principal limitation is the time required for the following steps: 1) searching and identifying new data source, 2) learning how to deal with new dataset, 3) creating an automated process for data incorporation, 4) providing computational resources for data processing, 5) retraining and recalibrating the data model. We use the diverse dataset from the National Oceanic and Atmospheric Administration (NOOA) has recently launched Big Data Initiative Program, offering public access to its open data by accreting its uniquely generated data (satellites, radars, ships, weather models) to public and private partnerships through commercial cloud platforms [29]. In this study, we have fused together datasets from several sources: a) Utility provided data, b) National Agriculture Imagery Program (NAIP) imagery [30], c) Automated Surface Observing Systems) historical weather from NOAA [31], d) Historical lightning data from Vaisala [32], and e) County borders disposition from Esri [33].

An iterative approach for adding new datasets (or altering any step in the framework) may be developed following the CI/CD concept [34]. The approach offers several benefits in SoR prediction applications:

- Process is standardized and streamlined.
- New features can constantly be added to the framework.
- New dataset effects can be readily evaluated against previous implementations.
- Testing and quality control procedures ensure a smooth transition to the production stage.

B. HISTORICAL OUTAGE DATA

We analyze the impact of different data parameters on the outage SoR predictions. Utility-provided historical outage logs contain information about outages. First, we separate planned outages from all other types of outages. To better understand how faults are distributed throughout time, we aggregate outages by the quarter of occurrence and cause. Environmental conditions (vegetation or weather) constitute a substantial portion of known causes of outages, as shown in Fig. 4.

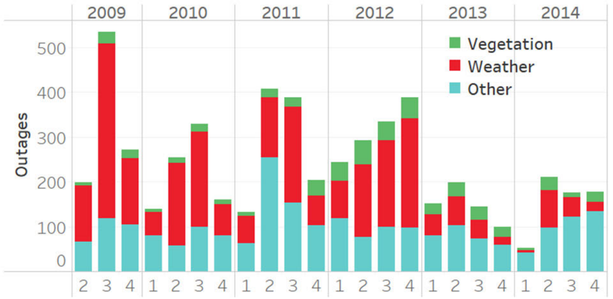

FIGURE 4. Outage distribution in time.

The utility outage data we used is in Central Time Zone (CDT and CST). Other datasets used in this study are in UTC (Coordinated Universal Time), which never switches to daylight saving time. To ensure all the data are in the same time zone, we converted outage time stamps to UTC. The *pytz* library offers a convenient way for such conversion.

C. WEATHER DATA

We obtained the historical weather data from National Oceanic and Atmospheric Administration (NOAA) Automated Surface Observing System (ASOS) [31]. Historical weather dataset comes in a variety of temporal resolutions: from 5-min to 1 hour. The same dataset can be obtained through the user-friendly website of Iowa Environmental Mesonet (IEM) [35], which allows one to select weather station locations, types of weather parameters, and timespan directly from the website's interface. The IEM also provides a script for the automated download of data [36]. The resulting downloaded file is in .csv format, which can be conveniently ingested by *pandas* and further manipulated.

The next step is to select the weather parameters of interest. We used the following parameters with least of missing values:

- Air Temperature,
- Dew Point Temperature,
- Wind Direction in degrees from true north,
- Wind Speed,
- Wind Gust,
- One-hour precipitation for the period from the observation time to the time of the previous hourly precipitation reset,
- Relative Humidity,
- Present Weather Codes.

The missing data is detected and discarded for each WS and parameter.

We have also accounted for the duration of high wind speed by summing number of hours with wind speed higher than 7, 10, 13 knots in last 3, 6 and 12 hours. We note that our analysis focuses on severe weather, and not catastrophic weather with very high winds, and the infrastructure remains intact.

The weather parameters need to be spatially correlated to CKTIDs. We use centroids of CKTIDs as a point where the weather parameters need to be calculated. The distances between each centroid of CKTID and WS are calculated and stored in a table. These distances are then used to calculate weather parameters for each CKTID.

To correlate weather parameters spatially and temporally to the outages, we get an average of the available parameters for each event weighted by distance and time (1):

$$P_{CKTID} = \frac{\sum_{i=1}^{N} Wgeog_i \cdot Wtime_i \cdot P_i}{\sum_{i=1}^{N} Wgeog_i \cdot Wtime_i}, \qquad (1)$$

where

$$Wgeog_i = \frac{1}{Geographic\ dist.(CKTID; WS)},$$

$$Wtime_i = \frac{1}{Time\ dist.(Weather\ measurement\ time; Event\ time)}.$$

In our study, we used the Euclidean distance between the centroids of CKTIDs and the WSs as a geographical distance in (1). The kernel (2) is used for the time distance:

$$Time\ dist.\ (t1; t2) \begin{cases} 1, & if\ t2 - t1 < 60\ min \\ \infty, & otherwise \end{cases} \qquad (2)$$

The time distance kernel only considers the measurements available in a 1-hour window before the event, discarding the measurements outside this window. That approach assumes that the weather preceding a fault has a major effect on it. However, different time kernels can be used to give more weight to measurements that are closer to the event timestamp.

For each hour of the study, the weather parameters are calculated for corresponding CKTIDs using distances between CKTIDs and WSs. Afterwards, the dataset is labeled into two classes: faults and normal operation (NO) based on the timestamps of fault occurrence. Then, to address the imbalance of the dataset, NO is randomly sampled to be of the same order as faults resulting in 517 faults and 581 NO. Only vegetation and weather caused outages are considered.

D. IMAGERY DATA

National Agriculture Imagery Program (NAIP) imagery data [37] used in this study can be accessed from Texas Natural Resources Information System (TNRIS) [30]. NAIP imagery consists of 3 bands: R, G, B. Each band is captured by a separate sensor during the imagery acquisition, which is performed by means of aerial photography [38]. NAIP imagery is clustered by county and is updated every two years.

For our study, we focused on extracting specific parts of the imagery data that are near the feeders. These features characterize the amount of vegetation around the feeders and how close vegetation is to a feeder. The underlying hypothesis is that the more vegetation around a feeder and the smaller its proximity to the feeder, the bigger the risk of an outage due to an increased probability of: a) tree branches touching conductors during strong wind, b) trees and/or branches falling onto the feeder during severe weather conditions, and c) trees growing into the conductors from underneath the feeder [39].

In total, there are 12 original NAIP imagery datasets: 4 datasets corresponding to the counties for 2010, 2012, and 2014. The computer used for the processing has 16 cores of Intel ®Core ™i9-9900 CPU with 3.1 GHz and 64 GB of RAM. One NAIP raster consumes around 30% of the computer resources during clipping. So, we can run 3 parallel computing nodes simultaneously on one machine. Buffers of 20 meters around lines grouped by county are used as the clipping boundaries for imagery.

After the clipping is finished, we are left with raster files in 20 meters vicinity around feeders in each county. The next step is to separate tree locations from the rest of the dataset. For that purpose, unsupervised clustering is used, followed by the reclassification of raster cells into two categories: vegetation (1) and no vegetation (0). ArcGIS Pro has an unsupervised clustering tool: IsoCluster. To run the tool, one needs to specify the number of clusters. The optimal number of clusters is determined empirically. Usually, the optimal number is around three times the number of bands in a raster dataset. In our case, we used 30 clusters.

After running unsupervised clustering, we get a raster with cells classified into 30 clusters. The tool uses information embedded in all the bands to assign the cluster to each cell. At this point, one needs to decide whether each cluster corresponds to an area with vegetation (1) or without (0). That process is manual, and it helps to have the original raster underneath the clustered raster. The resulting reclassified raster represents the location of vegetation around the power lines. An example of such a raster is shown in Fig. 5, where green areas represent vegetation in the lines' vicinity. Once all the raster files for counties are reclassified, same-year files are merged. The reclassified raster files are then converted into vector representation for easier use with other datasets.

E. LIGHTNING DATA

The lightning data comes from the National Lightning Detection Network operated by Vaisala [32], [40]. For each

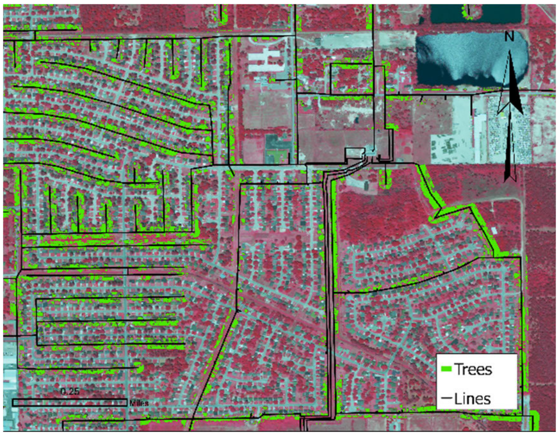

FIGURE 5. Reclassified raster.

lightning strike, the following information is collected: the location of the lightning strike, timestamp, lightning current, and type of lightning (cloud to cloud or cloud to ground).

We import lighting data into GDB as points, only the cloud-to-ground lightning strikes are extracted and sundered into different FC for each year. The hypothesis is that only the cloud-to-ground lightning strikes affect the feeders and cause outages. Then, the absolute value of the current is calculated from the original current, which accounts for polarity.

The hypothesis is also that certain parts of the network may be more susceptible to lightning strikes than others, which results in a higher risk for such parts of the network. The reason might be in its geographical location and its surroundings: if the feeder pole as a relatively tall structure is standing out in a particular area, then the lightning is more likely to strike it since lightning is "attracted" to taller objects on the ground [41], [42], [43]. To get the statistics of lightning strikes over different CKTIDs, we are using the buffers described in section III-B. The distances used are 5, 10, 15, 20, 30, 50, 100, 500, and 1000 meters. So, there are 9 buffers around each CKTID of the network. We count how many lightning strikes hit inside each buffer over a predefined time interval and calculate their average current. These are used as features for the ML classifier.

IV. DEVELOPMENT OF SOR PREDICTION MODEL

A. MODEL TRAINING AND EVALUATION

After all the features are prepared, we use them to train the ML classifier. The performance of 3 classifiers is compared: Random Forest (RF), Logistic Regression (LR), and Catboost (CB) [44]. Performance is measured by the following metrics: F1 Score, Area Under the Precision-Recall Curve (PRC AUC), and Area Under the Receiver Operating Characteristic (ROC AUC). Descriptions of the algorithms and metrics used for performance evaluation can be found in [45], [46], [47], [48], [49], and [50]. Classifiers are trained and tested using Stratified K-Folds cross-validator with 5 folds. The average performance metrics scores of the algorithms are presented in Table 1. The highest achievable score for each metric is 1.0. Our data indicates that while ML algorithms show strong predictive abilities, they are not flawless. Specifically, both CB and RF demonstrate similar performance, surpassing LR. A direct and unbiased performance comparison with existing methods is challenging, given the variations in spatiotemporal focus among studies and the unique regional weather patterns they consider.

B. CALCULATING RISK MAPS

The ML classifier outputs the probability of an outage under given weather conditions for each individual CKTID in the network for a given timestamp. The SoR values for several timestamps are combined and exported as a .csv table and then imported into ArcGIS. The tables need to have the prediction timestamp as a separate column to use the time-series visualization capabilities of ArcGIS. After importing, the risk values from the algorithm are joined with lines FC. Predefined layer symbology parameters are applied to the imported dataset to standardize the color scheme.

TABLE 1. Performance metrics scores.

	RF	LR	CB
ROC AUC	0.91	0.761	*0.926*
PRC AUC	0.916	0.748	*0.93*
F1	0.823	0.655	*0.836*
Precision	0.857	0.662	*0.872*
Recall	0.793	0.65	*0.803*

To illustrate the risk map usage, the 12-hour windows before the known outages are selected to build risk maps for that period. One can see how the risk changes in the system as the outage approaches in Fig. 6. As can be seen, the risk is low in the beginning and increases with time, eventually leading to an outage.

The spatial differentiation of the risk maps is not as accurate as differentiation timewise. The improvement of spatial awareness of the framework is left for future work. The risk maps can be used by utilities to improve real-time awareness of network vulnerabilities and support predictive decision-making practices. These risk maps can be used to establish various proactive measures that will help mitigate future high risks in the system. The information may also be used by the customers to prepare for times of elevated SoR levels.

V. OPTIMIZATION OF MITIGATION ACTIONS

In this paper, we are focused only on the application of SoRs to deploy a CNS by a utility, which improves the overall satisfaction of the customers. We introduce mitigation optimization based on SoR predictions. Our approach differs from the current reactive approach, where the assessment of the impacts is performed after the event (postmortem) [51], [52]. We are taking a proactive approach where customers are notified in advance and have time to prepare for a possible power interruption.

A. PROBLEM STATEMENT AND HYPOTHESIS

Our hypothesis is that given the SoR for the future timesteps for different parts of the network, it becomes feasible to devise and select appropriate mitigation measures that would reduce or eliminate the losses resulting from outages. We propose an optimization approach based on the acquired SoR levels, which outputs a set of mitigation actions from a predefined set for a given situation.

B. OBJECTIVE FUNCTION AND CONSTRAINTS

The objective function $\mathcal{F}(X)$ of optimization is maximized by selecting the mitigation actions (MAs) from the set Θ.

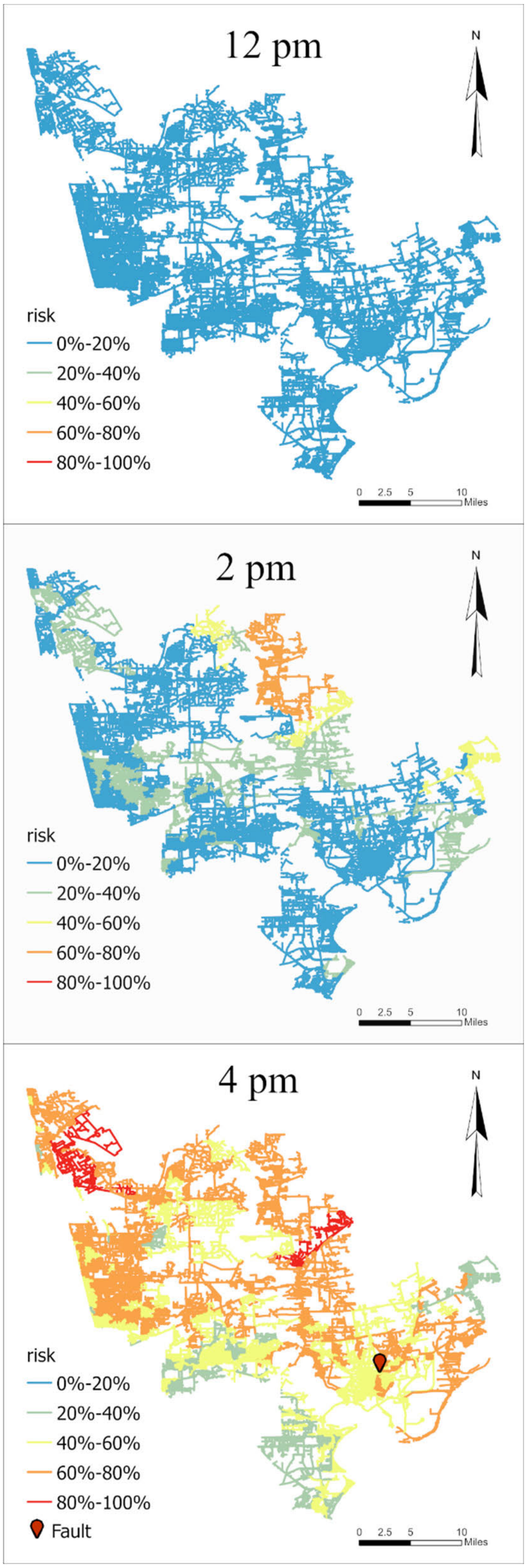

FIGURE 6. Risk maps for 12 pm – 4 pm.

The choice of a specific $\mathcal{F}(X)$ is made to best suit the interests and priorities of utility companies and end consumers. The objective function should be designed to optimize the overall outcome and ensure that the selected MAs effectively address the needs and concerns of the stakeholders involved.

The set of mitigation actions Θ (by utility and/or customers) is determined by the availability of the resources, system topology, cost of action, market conditions, level of flexibility of consumers and prosumers, time of the day, societal expectations, etc. Certain utility actions may necessitate longer time frames and require more resources, such as replacing old equipment or executing tree trimming. Some customer actions can be taken immediately, such as canceling a family event or moving to a warming/cooling center. We refer to these attributes as the inertia of an agent towards a specific mitigation action, reflecting their inclination and readiness to undertake it.

One also must account for the constraints $g_i(X)$ and $h_j(X)$ that may be present in the system at the time of MA scheduling and execution. Some MA can be infeasible at the time of high risk, while other parameters may need to remain unchanged. Accounting for these constraints would ensure that the selected MAs align with the current system conditions and limitations.

The proposed approach for optimization can be summarized as follows (3), (4):

$$argmax_{\Theta}\mathcal{F}(X) \tag{3}$$

$$s.t.: \begin{cases} g_i(X) \leq 0, i = 1, \ldots, l \\ h_j(X) = 0, j = 1, \ldots, k \end{cases} \tag{4}$$

where X represents a vector of parameters on which the objective function and constraints depend.

VI. CUSTOMER SATISFACTION

We introduced the customer satisfaction index (CSI) as a quantitative measure of customers' satisfaction with utility services. We demonstrate how the CSI may be improved by sending notifications to customers about potential outages in the system that can affect them.

A. UTILITY FUNCTIONS

At every moment in time, each customer is assigned a utility function (UF) denoted by $r_j(t1,t2)$ (5). This function represents the customer's perceived value of being correctly or falsely notified about an outage that will eventually happen in a predefined time interval. One can also think of this function in terms of the cost of false positive (FP) and false negative (FN) signals and the reward of true positive (TP) and true negative (TN) signals provided by the prediction model. The UF reflects Customer Interruption Cost since it aggregates both direct and indirect impacts [53], [54]. The utility function is dependent on time because it may change throughout the day/month/year and is subject to personal preference:

$$r_j\left(t1, t2\right) = \begin{cases} a_j - \int \varphi_j(Tout)dTout, \ \textit{if} \ O_j\left(t1, t2\right) = +1, N_j\left(t1, t2\right) = +1 \\ b_j - \int \gamma_j\left(Tout\right) dTout, \ \textit{if} \ O_j\left(t1, t2\right) = +1, N_j\left(t1, t2\right) = -1 \\ c_j, \ \textit{if} \ O_j\left(t1, t2\right) = -1, N_j\left(t1, t2\right) = +1 \\ d_j, \ \textit{if} \ O_j\left(t1, t2\right) = -1, N_j\left(t1, t2\right) = -1 \end{cases} \quad (5)$$

where

- a is a reward for a correct notification about an outage that has occurred,
- b is a penalty for a missed notification about an outage that has occurred,
- c is a penalty for an incorrect notification about an outage that did not happen (disturbance cost),
- d is a reward for not notifying when there is no outage,
- φ_j and γ_j are dissatisfaction rate functions,
- *Tout* is the duration of an outage.

We also utilize indicator functions $O_j(t1,t2)$ and $N_j(t1,t2)$ that take a value of +1 in case of an outage or notification taking place, respectively, and a value of -1 in case of an outage or notification not taking place in the time interval [t1, t2) (6), (7):

$$O_j\left(t1, t2\right) = \begin{cases} +1, & \textit{if outage occurred during}[t1, t2) \\ -1, & \textit{otherwise}, \end{cases} \quad (6)$$

$$N_j\left(t1, t2\right) = \begin{cases} +1, & \textit{if notification sent during}[t1, t2) \\ -1, & \textit{otherwise} \end{cases} \quad (7)$$

During actual outages ($O_j\left(t1,t2\right) = +1$), the UF includes an integral of the dissatisfaction rate functions φ_j and γ_j with respect to *Tout*. The longer it takes to return the power supply to a customer once the outage happens, the more dissatisfied customer becomes with the utility. The dissatisfaction rate from the outage duration is different in case of being notified in advance and not being notified in advance, as can be seen from the example of such functions, which is shown in Fig. 7. The form of functions is not limited to exponential functions and can be an arbitrary function, perhaps determined through behavioral experiments.

The contents of the notification message to a customer can consist of a set of items: outage probability, expected outage duration, possible mitigation actions, recommendations, etc. An effective message structure would yield better satisfaction and a more considerable impact reduction. Formulation of the message structure and estimation of how efficiently a customer acts, given the notification, lies beyond the scope of this paper and is left for future research. In this paper, we limit the message to a warning about a potential outage in the customer's area in the next hour. Each customer is assumed to act according to his/her personal circumstances to reduce outage impact.

$Tres_{cust}$ is the expected time of restoration of power the customer anticipates, and when the power outage lasts more than $Tres_{cust}$, dissatisfaction grows at an accelerated rate. The $Tres_{cust}$ is customer dependent and, in general, is affected by the personal background and experience of a particular customer. For example, a customer may base his/her estimation based on previous outages. The time of restoration expected by a customer can be purposely influenced by a utility sending a notification of the expected restoration time for a particular outage, thus, making the customer's expectation more specific. The functions will be revisited in Section VIII.

A utility has its own expected restoration time for each outage: $Tres_{util}$. It can be predicted by a separate ML model or can be assessed by means of statistical analysis for various parts of the system, for example, using the historical mean. The repair crews' allocation during outages can be optimized by reducing actual restoration time for customers with higher dissatisfaction rates.

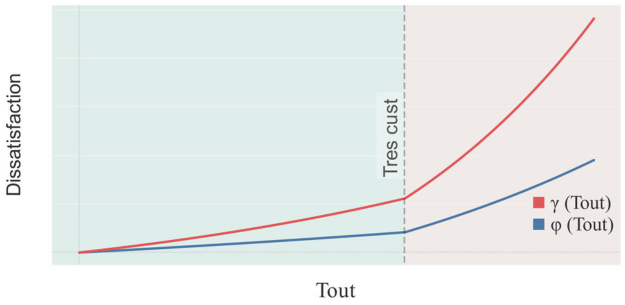

FIGURE 7. Dissatisfaction rate from the power outage.

While the previous two restoration times are expected values by different parties, after the restoration is completed, the actual restoration time is known. We denote this as $Tres_{actual}$. The dissatisfaction of a customer would be calculated by comparing the actual and predicted restoration time for each outage. However, the decision on MA must be made based on the expected values since the actual restoration time is not available at the time of making the decision.

The coefficients of the utility function and dissatisfaction rate functions for individual customers are subject to behavioral economics assessment because the perceived cost/value of an outage by a customer is different from the monetary cost/value. Surveys of customer opinions are necessary to address the issue. In this paper, we assume that the utility functions are known.

B. CUSTOMER SATISFACTION INDEX

By informing customers about potential disruptions, we aim to improve their overall satisfaction and reduce any inconvenience caused by unexpected service interruptions. The notifications serve as a proactive measure to keep customers informed and engaged, enhancing their perception of the utility service provider's responsiveness and reliability.

Satisfaction Index CSI_j for a given customer j is a sum of all rewards/penalties increments from the UF up to a given moment in time $t0$, discounted by a discounting factor E (8):

$$CSI_j\left(t0\right) = \sum_{t=-\infty}^{t0-dt} \left[r_j(t, t+dt) \cdot O_j(t, t+dt) \cdot N_j(t, t+dt)\right] e^{t \cdot E}, \tag{8}$$

where dt is a discretization time step.

VII. STATE OF RISK INCORPORATION INTO OPTIMIZATION

A. STATE OF RISK

The likelihood of an outage is reflected by the State of Risk (SoR). SoR represents the conditional probability p of the system element i failure in the time interval [t1, t2) given the set of operation conditions Ω, which includes historical and forecasted weather conditions, system topology, loading and generation conditions, etc. (9):

$$SoR_i\left(t1, t2 | \Omega\right) = p\Omega)p\left(element\ i\, fails\ in[t1, t2,) | \Omega\right) \tag{9}$$

Each customer j in the network at current time t0 is described by the following parameters, including SoR:

- Geographical location of the customer in the network,
- Customer location in the grid topology (electrical location),
- SoRs for the next time intervals: $SoR_j\left(t0{,}t1\right)$, $SoR_j\left(t1{,}t2\right) \ldots$,
- History of experienced outages in the past: $HO_j\left(t0\right)$
- History of the notifications sent: $HN_j\left(t0\right)$,
- UFs for the next time intervals: $r_j\left(t0, \mathrm{t}1\right)$, $r_j\left(t1, \mathrm{t}2\right) \ldots$,
- Current Customer Satisfaction Index: $CSI_j\left(t0\right)$.

The relation between element i and customer j can be formulated in several ways. In this paper, element i is a feeder to which customer j is connected.

B. SOR BASED ACTIONS

Given the uncertainty of the outage in the future period, one can define a random variable that represents the possible gain or loss of the Customer Satisfaction Index ΔCSI_j using the predicted SoR levels for that period. The gain/loss of the next period depends on whether the notification will be issued and whether an outage will take place. The PMF of such random variable is presented in (10) (time periods are omitted to simplify the notation):

$$\begin{aligned}
&P\left(\Delta CSI_j = a_j + \int \varphi_j\left(Tout\right)d\ Tout \,|\, O = 1, N = 1\right) = SoR \\
&P\left(\Delta CSI_j = -b_j - \int \gamma_j\left(Tout\right)d\ Tout \,|\, O = 1, N = -1\right) \\
&\quad = SoR \\
&P\left(\Delta CSI_j = -c_j \,|\, O = -1, N = 1\right) = 1 - SoR \\
&P\left(\Delta CSI_j = d_j \,|\, O = -1, N = -1\right) = 1 - SoR
\end{aligned} \tag{10}$$

The action vector θ_j represents a mitigation action for each customer (11):

$$\theta_j = \begin{cases} \left(1\ 0\right), & if\ N = +1 \\ \left(0\ 1\right), & if\ N = -1 \end{cases} \tag{11}$$

The objective function $\mathcal{F}\left(X\right)$ at time $t0$ for the optimization is to maximize the Customer Satisfaction Index in the next time period across the entire grid with consideration of Satisfaction Index change ΔCSI_j, which is an expected value of future reward/penalty of the utility function, given the SoR (12):

$$\mathcal{F}\left(X\right) = \sum_{j=1}^{M} CSI_j\left(t0\right) + \theta_j \cdot \begin{pmatrix} E\left(\Delta CSI_j \,|\, N = 1\right) \\ E\left(\Delta CSI_j \,|\, N = -1\right) \end{pmatrix}, \tag{12}$$

where M is the total number of customers in the grid.

Using SoR, we can calculate the expected values of the Satisfaction Index change for both cases: notification will be sent ($N = 1$) and will not be sent ($N = -1$) (13):

$$\begin{aligned}
&E\left(\Delta CSI_j \,|\, N = 1\right) = \\
&SoR \cdot \left(a_j - \int \varphi_j\left(Tout\right)dTout\right) + \left(1 - SoR\right) \cdot \left(-c_j\right) \\
&E\left(\Delta CSI_j \,|\, N = -1\right) = \\
&SoR \cdot \left(-b_j - \int \gamma_j\left(Tout\right)dTout\right) + \left(1 - SoR\right) \cdot \left(d_j\right)
\end{aligned} \tag{13}$$

The optimization problem is then to choose such mitigation vectors θ_j, so that the objective function is maximized or, in other words, to choose which customers should be notified about possible outages.

There are several constraints to the optimization problem. First, we would not notify a customer if there is already an outage at its feeder. Second, each customer can have a "do not disturb" mode when notifications are not accepted. We also consider a third constraint, namely the total number of notifications in the system. Even though the cost of sending a single notification is minuscule (given that the notifications are sent by means of the Internet), in an exceptionally large system sending frequent notifications may require more processing power in the hardware and faster Internet connections. The constraints are summarized in (14):

$$\begin{aligned}
&\theta_j = \left(0\ 1\right), \quad if\ experiencing\ an\ outage \\
&\theta_j = \left(0\ 1\right), \quad if\,\text{do``not disturb}''\text{ mode} \\
&\sum_{j=1}^{M} \theta_{j1} - N_{max} \leq 0,
\end{aligned} \tag{14}$$

where N_{max} is the maximum number of notifications in each period.

C. HYPERPARAMETER OPTIMIZATION

To further improve the CSI of the customers, we introduce an additional step in optimization: find and set minimum SoR threshold $SoR_{min}(s,t)$ as a function of time t and locations s, below which the notifications will not be issued (15):

$$\theta_j(s,t) = \begin{pmatrix} 0 & 1 \end{pmatrix}, if \ \ SoR_j < SOR_{min}(s,t) \tag{15}$$

This hyperparameter helps to fine-tune message notifications and tie thresholds to a current situation in the service area. We suggest that the minimum threshold is updated on a periodic basis (which can be chosen arbitrarily) based on the performance of the CNS in the last period(s). Value of the $SoR_{min}(s,t)$ is the threshold to maximize the SI at location s during the previous period(s).

VIII. MITIGATION EVALUATION

We have evaluated the impact of Customer Notification System implementation on 1 year of real-life data.

We have randomly generated a utility function for each customer in the network. The forms of dissatisfaction rate functions are assumed to be linear (16):

$$\varphi_j(Tout) = \begin{cases} \lambda 1_j Tout, & if \ Tout < Tres \\ \lambda 2_j Tout + w_j, & otherwise \end{cases}$$

$$\gamma_j(Tout) = \begin{cases} \mu 1_j Tout, & if \ Tout < Tres \\ \mu 2_j Tout + z_j, & otherwise \end{cases} \tag{16}$$

where $0 < \lambda 1 < \lambda 2$, $0 < \mu 1 < \mu 2$, $\lambda 1 < \mu 1$, $\lambda 2 < \mu 2$, and w_j and z_j are such that the functions are continuous in $Tres_{cust}$.

The system has a total of 698 313 customers located at different feeders. The time of restoration $Tres_{actual}$ of an outage obtained from utility provided data. $Tres_{util}$ is set to 2 hours for all outages, as the current utility practices suggest. For each customer, $Tres_{cust}$ is modeled as a sample from a lognormal distribution with a mean of 0.7 hours and a standard deviation of 0.2 hours. Coefficients $\lambda 1, \lambda 2, \mu 1, \mu 2$ are modeled by uniform distribution with bounds (0,1) respecting the conditions above. Hence the utility function coefficients have only 2 degrees of freedom [55], we model $dj= 0$, $bj= 0.1$, aj as a lognormal distribution with a mean of 1 and standard deviation of 1, cj as a lognormal distribution with a mean of 1 and standard deviation of 1, multiplied by 0.01.

In general, the coefficients aj, bj, cj, dj can be of any sign because there possibly might exist a customer who, for example, likes being falsely notified. However, we deem it to be viable to assume that such customers are rare, for that reason, all the coefficients are modeled as a positive number, which is in accordance with the "reasonableness" conditions presented in [55]. The initial CSI_j for each customer is assumed to be zero.

To ensure the robustness of the results, we have repeated the test for the entire system on 1 year of data for a total of 150 times, which is considered to be a sufficient number of samples for the Central Limit Theorem to be applicable [56], [57], and recorded the results of each run. The results of the optimization runs are shown in Fig. 8 in the form of the end-of-the-year percentage difference between the SI of the entire system with and without CNS implementation. As can be seen from the figure, the usage of CNS based on SoR predictions improves the Satisfaction Index by 54.3% on average.

IX. CONCLUSION

By summarizing our findings, we arrive to the following conclusions:

- Employing an iterative approach for incorporating new relevant datasets is essential to the performance evaluation of the SoR prediction models.
- The practical value of SoR maps for utilities is in ability to plan and anticipate potential outages, enhancing their operational preparedness for inclement weather events.
- Proactive outage management facilitated by CNS allows utilities to effectively communicate with customers, raising overall satisfaction levels and minimizing detrimental outage impacts.

Our future work will focus on enhancing our outage SoR prediction by incorporating new datasets and deploying advanced feature engineering. We will also conduct a detailed analysis of vegetation-related features and introduce outage duration prediction.

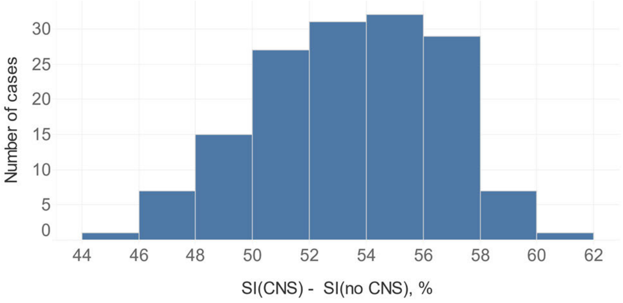

FIGURE 8. Distribution of percentage difference between CSI with CSN and SI without CNS.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of CenterPoint Energy staff by offering advice and historical outage data. Special thanks are due to Dr. Tatjana Dokic for her original work, which is further enhanced in this paper. They also thank Dr. Alex Brown for his valuable insights and guidance.

REFERENCES

[1] M. Clark. (Jul. 14, 2014). *Aging US Power Grid Blacks Out More Than Any Other Developed Nation*. [Online]. Available: https://www.ibtimes.com/aging-us-power-grid-blacks-out-more-any-other-developed-nation-1631086

[2] *Electric Disturbance Events (OE-417) Annual Summaries*. US Department of Energy. Accessed: Dec. 2022. [Online]. Available: https://www.oe.netl.doe.gov/OE417_annual_summary.aspx

[3] D. C. Lineweber and S. McNulty. (Jun. 2001). *The Cost of Power Disturbances to Industrial & Digital Economy Companies*. EPRI's Consortium for Electric Infrastructure for a Digital Society (CEIDS). Accessed: Dec. 2022. [Online]. Available: https://www.epri.com/research/products/3002000476

[4] *The Smart Grid: An Introduction*. US Department of Energy. Accessed: Dec. 2022. [Online]. Available: http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/DOE_SG_Book_Single_Pages%281%29.pdf

[5] M. Lehtonen and B. Lemstrom, "Comparison of the methods for assessing the customers' outage costs," in *Proc. Int. Conf. Energy Manag. Power Del.*, 1995, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/500691

[6] R. R. Richwine, "CASOM 18: The relationship between scheduled maintenance and forced outages and its economic impact on selecting availability goals," *World Energy Council, Section*, vol. 6, pp. 1–3, Jan. 2004.

[7] F. Carlsson and P. Martinsson, "Willingness to pay among Swedish households to avoid power outages: A random parameter tobit model approach," *Energy J.*, vol. 28, no. 1, Jan. 2007, doi: 10.5547/ISSN0195-6574-EJ-Vol28-No1-4.

[8] R. Diao, V. Vittal, and N. Logic, "Design of a real-time security assessment tool for situational awareness enhancement in modern power systems," *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 957–965, May 2010, doi: 10.1109/TPWRS.2009.2035507.

[9] M. Panteli and D. S. Kirschen, "Situation awareness in power systems: Theory, challenges and applications," *Electric Power Syst. Res.*, vol. 122, pp. 140–151, May 2015, doi: 10.1016/j.epsr.2015.01.008.

[10] M. Kezunovic, P. Pinson, Z. Obradovic, S. Grijalva, T. Hong, and R. Bessa, "Big data analytics for future electricity grids," *Electric Power Syst. Res.*, vol. 189, Dec. 2020, Art. no. 106788, doi: 10.1016/j.epsr.2020.106788.

[11] M. Kezunovic, Z. Obradovic, T. Djokic, and S. Roychoudhury, "Systematic framework for integration of weather data into prediction models for the electric grid outage and asset management applications," in *Proc. 51st Hawaii Int. Conf. Syst. Sci.*, 2018, pp. 2737–2746, doi: 10.24251/HICSS.2018.346.

[12] T. Dokic and M. Kezunovic, "Predictive risk management for dynamic tree trimming scheduling for distribution networks," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 4776–4785, Sep. 2019, doi: 10.1109/TSG.2018.2868457.

[13] M. Kezunovic and T. Dokic, "Big data framework for predictive risk assessment of weather impacts on electric power systems," in *Proc. Grid Future, CIGRE US Nat. Committee*, Atlanta, GA, USA, Nov. 2019.

[14] A. Ghasemi, A. Shojaeighadikolaei, K. Jones, M. Hashemi, A. G. Bardas, and R. Ahmadi, "A multi-agent deep reinforcement learning approach for a distributed energy marketplace in smart grids," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids (SmartGridComm)*, Nov. 2020, pp. 1–6, doi: 10.1109/SmartGridComm47815.2020.9302981.

[15] J. B. Leite, J. R. S. Mantovani, T. Dokic, Q. Yan, P.-C. Chen, and M. Kezunovic, "Resiliency assessment in distribution networks using GIS-based predictive risk analytics," *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 4249–4257, Nov. 2019, doi: 10.1109/TPWRS.2019.2913090.

[16] A. Shojaeighadikolaei, A. Ghasemi, A. G. Bardas, R. Ahmadi, and M. Hashemi, "Weather-aware data-driven microgrid energy management using deep reinforcement learning," in *Proc. North American Power Symp. (NAPS)*, Nov. 2021, pp. 1–6, doi: 10.1109/NAPS52732.2021.9654550.

[17] A. F. Soofi, R. Bayani, and S. D. Manshadi, "Analyzing power quality implications of high level charging rates of electric vehicle within distribution networks," 2021, *arXiv:2106.14819*.

[18] F. Yang, D. Cerrai, and E. N. Anagnostou, "The effect of lead-time weather forecast uncertainty on outage prediction modeling," *Forecasting*, vol. 3, no. 3, pp. 501–516, Jul. 2021, doi: 10.3390/forecast3030031.

[19] 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, doi: 10.1109/ACCESS.2019.2902558.

[20] D. Cerrai, M. Koukoula, P. Watson, and E. N. Anagnostou, "Outage prediction models for snow and ice storms," *Sustain. Energy, Grids Netw.*, vol. 21, Mar. 2020, Art. no. 100294, doi: 10.1016/j.segan.2019.100294.

[21] F. Yang, D. W. Wanik, D. Cerrai, M. A. E. Bhuiyan, and E. N. Anagnostou, "Quantifying uncertainty in machine learning-based power outage prediction model training: A tool for sustainable storm restoration," *Sustainability*, vol. 12, no. 4, p. 1525, Feb. 2020, doi: 10.3390/su12041525.

[22] P. Dehghanian, B. Zhang, T. Dokic, and M. Kezunovic, "Predictive risk analytics for weather-resilient operation of electric power systems," *IEEE Trans. Sustain. Energy*, vol. 10, no. 1, pp. 3–15, Jan. 2019, doi: 10.1109/TSTE.2018.2825780.

[23] E. H. Ko, T. Dokic, and M. Kezunovic, "Prediction model for the distribution transformer failure using correlation of weather data," in *Proc. 5th Int. Colloq. Transformer Res. Asset Manag.*, B. Trkulja and Ž. Štih, Eds. Singapore: Springer, 2020, pp. 135–144, doi: 10.1007/978-981-15-5600-5_12. [Online]. Available: https://www.researchgate.net/publication/342993702_Prediction_Model_for_the_Distribution_Transformer_Failure_Using_Correlation_of_Weather_Data

[24] D. Zastrau, M. Schlaak, T. Bruns, R. Elsner, and O. Herzog, "Differences in wind forecast accuracy in the German north and Baltic seas," *Int. J. Environ. Sci. Develop.*, vol. 5, no. 6, pp. 575–580, Dec. 2014, doi: 10.7763/IJESD.2014.V5.549.

[25] *ArcGIS Pro*, ESRI, Redlands, CA, USA, 2021.

[26] (2021). *ArcPy*. ESRI. [Online]. Available: https://www.esri.com/en-us/arcgis/products/arcgis-python-libraries/libraries/arcpy

[27] (2021). *Feature Datasets in ArcGIS Pro-ArcGIS Pro | Documentation*. [Online]. Available: https://pro.arcgis.com/en/pro-app/latest/help/data/feature-datasets/feature-datasets-in-arcgis-pro.htm

[28] ESRI. *Spatial Join (Analysis)*. Accessed: Sep. 2023. [Online]. Available: https://pro.arcgis.com/en/pro-app/latest/tool-reference/analysis/spatial-join.htm

[29] (2021). *Big Data Program*. [Online]. Available: https://www.noaa.gov/information-technology/big-data

[30] *TNRIS DataHub*. Accessed: Feb. 2023. [Online]. Available: https://data.tnris.org/?pg=1&inc=24&s=naip#2.51/31.46/-100.09

[31] (2021). *Automated Surface Observing Systems*. NOAA's National Weather Service. [Online]. Available: https://www.weather.gov/asos/asostech

[32] Vaisala. *Lightning Detection Networks*. Accessed: Feb. 2023. [Online]. Available: https://www.vaisala.com/en/products/systems/lightning-detection-networks

[33] Esri. *USA County Boundaries*. Accessed: Feb. 2023. [Online]. Available: https://www.arcgis.com/home/item.html?id=f16090f6d3da48ec8f144a0771c8fec4

[34] B. El Khalyly, A. Belangour, M. Banane, and A. Erraissi, "A new metamodel approach of CI/CD applied to Internet of Things ecosystem," in *Proc. IEEE 2nd Int. Conf. Electron., Control, Optim. Comput. Sci. (ICECOCS)*, Dec. 2020, pp. 1–6, doi: 10.1109/ICECOCS50124.2020.9314485.

[35] *Iowa Environmental Mesonet: ASOS One Minute Data Download*. Accessed: Feb. 2023. [Online]. Available: https://mesonet.agron.iastate.edu/request/asos/1min.phtml

[36] IEM ASOS Download Service Script. (2021). *GitHub Repository*. [Online]. Available: https://github.com/akrherz/iem/blob/main/scripts/asos/iem_scraper_example.py

[37] K. W. Davies, S. L. Petersen, D. D. Johnson, D. B. Davis, M. D. Madsen, D. L. Zvirzdin, and J. D. Bates, "Estimating juniper cover from national agriculture imagery program (NAIP) imagery and evaluating relationships between potential cover and environmental variables," *Rangeland Ecology Manag.*, vol. 63, no. 6, pp. 630–637, Nov. 2010, doi: 10.2111/REM-D-09-00129.1.

[38] USDA. (Oct. 2021). *NAIP Imagery*. [Online]. Available: https://naip-usdaonline.hub.arcgis.com/

[39] D. W. Wanik, J. R. Parent, E. N. Anagnostou, and B. M. Hartman, "Using vegetation management and LiDAR-derived tree height data to improve outage predictions for electric utilities," *Electric Power Syst. Res.*, vol. 146, pp. 236–245, May 2017, doi: 10.1016/j.epsr.2017.01.039.

[40] A. T. Pessi, S. Businger, K. L. Cummins, N. W. S. Demetriades, M. Murphy, and B. Pifer, "Development of a long-range lightning detection network for the pacific: Construction, calibration, and performance," *J. Atmos. Ocean. Technol.*, vol. 26, no. 2, pp. 145–166, Feb. 2009, doi: 10.1175/2008jtecha1132.1.

[41] R. H. Golde, "Lightning and tall structures," *Proc. IEE*, vol. 125, no. 4, pp. 347–351, Apr. 1978, doi: 10.1049/piee.1978.0084.

[42] M. A. Uman, *The Art and Science of Lightning Protection*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[43] V. Cooray, *Lightning Protection* (Energy Engineering). London, U.K.: The Institution of Engineering and Technology, 2010.

[44] P. S. Kumar, S. Mohapatra, B. Naik, J. Nayak, and M. Mishra, "CatBoost ensemble approach for diabetes risk prediction at early stages," in *Proc. 1st Odisha Int. Conf. Electr. Power Eng., Commun. Comput. Technology(ODICON)*, Jan. 2021, pp. 1–6.

[45] R. Baembitov, T. Dokic, M. Kezunovic, Y. Hu, and Z. Obradovic, "Fast extraction and characterization of fundamental frequency events from a large PMU dataset using big data analytics," in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2021, pp. 3195–3204, doi: 10.24251/HICSS.2021.389.

[46] R. Baembitov, M. Kezunovic, and Z. Obradovic, "Graph embeddings for outage prediction," in *Proc. North Amer. Power Symp. (NAPS)*, College Station, TX, USA, Nov. 2021, pp. 1–6.
[47] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
[48] M. Kezunovic, R. Baembitov, and M. Khoshjahan, "Data-driven state of risk prediction and mitigation in support of the net-zero carbon electric grid," in *Proc. 11th Bulk Power Syst. Dyn. Control Symp.*, Waterloo, ONT, Canada, 2022, pp. 1–10.
[49] T. Dokic, R. Baembitov, A. A. Hai, Z. Cheng, Y. Hu, M. Kezunovic, and Z. Obradovic, "Machine learning using a simple feature for detecting multiple types of events from PMU data," in *Proc. Int. Conf. Smart Grid Synchronized Meas. Analytics (SGSMA)*, Split, Croatia, May 2022, pp. 1–6.
[50] M. Khoshjahan, R. Baembitov, and M. Kezunovic, "Impacts of weather-related outages on DER participation in the wholesale market energy and ancillary services," in *Proc. CIGRE Grid Future Symp.*, Providence, RI, USA, Oct. 2021, pp. 1–8.
[51] B. Vajgel, P. L. P. Corrêa, T. Tóssoli De Sousa, R. V. E. Quille, J. A. R. Bedoya, G. M. D. Almeida, L. V. L. Filgueiras, V. R. S. Demuner, and D. Mollica, "Development of intelligent robotic process automation: A utility case study in Brazil," *IEEE Access*, vol. 9, pp. 71222–71235, 2021, doi: 10.1109/ACCESS.2021.3075693.
[52] S. Pathak, "Leveraging GIS mapping and smart metering for improved OMS and SAIDI for smart city," in *Proc. Saudi Arabia Smart Grid (SASG)*, Dec. 2016, pp. 6–8, doi: 10.1109/SASG.2016.7849663.
[53] S. Kufeoglu, "Economic impacts of electric power outages and evaluation of customer interruption costs," Ph.D. thesis, Dept. Elect. Eng. Automat., Aalto Univ., Helsinki, Finland, 2015.
[54] A. Brown and J. Liu, "Forecasts and mitigation: Theory and experiment," Texas A&M Univ., College Station, TX, USA, Work. Paper, 2023.
[55] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2001, pp. 973–978.
[56] L. Kish, *Survey Sampling*. New York, NY, USA: Wiley, 1965.
[57] S. M. Ross, *Introductory Statistics*. Cambridge, MA, USA: Academic, 2010. [Online]. Available: http://www.sciencedirect.com/science/book/9780123743886

RASHID BAEMBITOV (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical power engineering and economics and business administration from the National Research University "Moscow Power Engineering Institute," Moscow, Russia, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree with Texas A&M University. He is a Graduate Research Assistant with Texas A&M University. His main research interests include big data and machine learning for power system applications, power system risk assessment, and smart grids.

MLADEN KEZUNOVIC (Life Fellow, IEEE) received the Dipl.-Ing. degree from the University of Sarajevo, Sarajevo, Bosnia, in 1974, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Kansas, Lawrence, KS, USA, in 1977 and 1980, respectively. He has been with Texas A&M University, College Station, TX, USA, for 33 years, where he is currently a Regents Professor, a Eugene E. Webb Professor, and the Site Director of the "Power Engineering Research Center" Consortium. He acted for over 30 years as the Principal Consultant of XpertPower Associates, a consulting firm specializing in power systems data analytics. His expertise is in protective relaying, automated power system disturbance analysis, computational intelligence, data analytics, and smart grids. He has authored over 600 articles, given over 120 seminars, invited lectures, and short courses, and consulted for over 50 companies worldwide. He is a CIGRE Fellow, an Honorary and Distinguished Member, a registered Professional Engineer in TX, and a member of the National Academy of Engineering.