Sensitivity Analysis of Machine Learning Algorithms for Outage Risk Prediction

Rashid Baembitov, Mladen Kezunovic
Texas A&M University
bae_rashid@tamu.edu,
kezunov@ece.tamu.edu

Daniel Saranovic, Zoran Obradovic
Temple University
daniel.saranovic@temple.edu,
zoran.obradovic@temple.edu

Abstract

Severe weather conditions are known for causing forced outages in the electric distribution grid. Recent research efforts were aimed at predicting outages using weather and historical outage data. This paper studies the sensitivity of different Machine Learning (ML) algorithms to the inclusion of weather parameters from adjacent geographic areas and data availability. We analyzed the ability of different ML algorithms to predict electric grid outage State of Risk (SoR). The selected algorithms are trained and tested on actual utility company data. The findings indicate that a bigger size of the training dataset improves the performance of all models, which is measured by the Receiver Operating Curve, Average Precision, and F1 Score. Conducted experiments suggest that at least two years of training data is required to achieve satisfactory performance. Also, we investigate a statistical significance in models' performance with the inclusion of weather in adjacent geographic areas.

Keywords: ML, State of Risk, Outage Prediction.

1. Introduction

The occurrence of forced outages in power systems, resulting from short circuits caused by faults or equipment failure, can pose a considerable safety hazard and economic burden for utilities, their customers, and society as a whole. The rise in severe weather conditions due to climate change has become one of the major concerns since it causes more frequent impacts of inclement weather on overhead feeders and other exposed components of the electric grid (Panteli et al., 2015). A new approach of predicting outages in the system allows a proactive mitigation approach to reducing or avoiding detrimental impact (M. Kezunovic et al., 2022), (Khoshjahan et al., 2021).

Several data model aspects such as historical weather and forecasts, GIS representation of utilities' assets, machine learning (ML) methods, and digitalization of utility operations to assess the potential risk to the power grid have been reported so far (Kezunovic et al., 2020), (Kezunovic et al., Nov., 2019). The selection of the best ML algorithm is vital in accurately predicting the State of Risk (SoR) for outages in the network, which reflects the probability of an outage occurrence in each place and time. Different algorithms have various strengths and weaknesses, which can impact the accuracy of the predictions. For instance, the ensemble algorithms such as Random Forest (RF) and Gradient Boosting are known for their robustness and accuracy, but they are less interpretable than decision trees (Shi et al., 2018), (Leistner et al., 2009).

In the past, RF algorithm was used along with dimensionality reduction techniques to predict the probability of transmission line outage during severe weather storms (Taylor et al., 2023). The Neural Network (NN) was utilized to predict the time of repair and restoration in distribution networks (Arif et al., 2018). Logistic Regression (LR) was implemented to predict the likelihood of power grid elements failure from an approaching hurricane (Eskandarpour et al., 2017). Support Vector Machine (SVM) that considers the deterioration level of the equipment was suggested in (Eskandarpour et al., 2018) to estimate the operability of the grid's components during extreme weather events. Graph Convolutional NN were used to process weather parameters and anticipate outages in the system (Owerko et al., 2018). Ensemble learning approaches were also utilized for outage prediction (Shashaani et al., 2018), (Zhang et al., 2018). Most of the studies analyze the outages during extreme storms or hurricanes that damage the grid infrastructure.

Forced outages in the power systems caused by environmental factors pose a threat to the safe and economical operation of distribution grids. The industry aims at reducing the number of outages as well as their impact (duration, number of customers affected, monetary loss, etc.) The cause of the outage may be attributed to human error or wear and tear, but most outages are caused by weather-related events such wind, lightning, or overgrown trees. We utilize the approach of gathering data that is relevant to

outage causes, training an ML algorithm with the general goal of predicting SoR of outages, and then taking mitigation actions to reduce or eliminate outage impact (M. Kezunovic et al., 2022), (Nematirad et al., 2023). In the context of this broader goal, this paper contributes by analyzing the outage prediction capability of diverse ML algorithms under conditions of data scarcity. Additionally, we explore the influence of incorporating weather conditions from a broader area on the performance. The paper evaluates how different models respond to data limitations and the integration of weather factors from adjacent regions, providing practical insight on the most effective approaches for outage prediction in distribution grids.

The comparison of ML algorithms for SoR predictions in the case of normal day-to-day operations during severe weather received less attention in the literature than SoR prediction during extreme weather events. Moreover, it is unclear what is the minimal amount of data needed for training an ML algorithm to achieve optimal performance. Our contribution is in analyzing algorithm sensitivity to the data scarcity conditions when estimating the size of the training dataset needed for an effective SoR model implementation for forced outages. We also evaluate the statistical significance of including weather parameters from a wider region for improving model accuracy. Our work provides guidance for the implementation of practical solutions under field data constraints*.

The rest of the paper is organized as follows. Section II gives the evaluation background. Section III describes data extraction, correlation, and preparation procedures used in this study. Section IV discusses the training of different algorithms. Section V summarizes the analysis of algorithm performance, while section VI draws conclusions. References are given at the end.

2. Background

Prior Work

This section provides an overview of outage prediction, illustrating the challenges of prior research. Correlation between outage frequency and environmental conditions, topographic exposure, and tree parameters were analyzed (Hirata, 2011). A statistical approach was utilized to explore 6 years of outage data based on transmission inventory in a utility located in Kazakhstan (Bapin et al., 2020). An analysis of outages in a Canadian utility shows that adverse weather conditions are the primary cause of outages (Bin et al., 1998). An outage risk-based approach to tree trimming scheduling to lower the impact of outages is proposed (Dokic et al., 2019). The literature review shows that significant benefits can be achieved by adopting an outage prediction approach in improving customer satisfaction and power grid operations (customer notifications, tree trimmings etc.), but it is left unclear which prediction algorithm is more accurate under specific data conditions. The metrics to compare the algorithms and evaluate their performance were also not used consistently in the prior studies.

Prior research also did not differentiate between the geographical scope (spatial aspect) of input variables. Including additional data from adjacent regions in the analysis may have positive effects on the accuracy and comprehensiveness of the assessment. By incorporating data from neighboring areas, the model can capture a broader range of factors that may impact the power grid's risk profile, such as weather patterns or infrastructure conditions. However, the use of more data needs to be justified because it is associated with increased computation times, storage requirements, download times, etc.

The data scarcity (temporal aspect) for the task of predicting outages has not been extensively addressed in the existing literature. The availability of sufficient and high-quality data is vital for developing accurate and reliable predictive models. However, in many cases, there may be limited historical data or incomplete datasets specifically tailored to the task of outage prediction in networks.

Our Contribution

In this paper, we focus on the sensitivity study of different ML algorithms to the available data temporally and spatially. The findings of this paper are built upon our prior work by delving into the specific issue of data availability and the geographical scope of inputs. While our previous research focused on various aspects of risk prediction when anticipating power grid outages, this study a) assesses the temporal sensitivity to data quantity and b) analyzes spatial extent of features and their impact on performance when incorporating data from adjacent regions.

We explore how the temporal quantity of available data influences the accuracy of different models and establish the minimum data quantity required to achieve acceptable outage prediction using a nested cross validation testing technique, which produces robust estimates. By demonstrating the dependency between data availability during the training step and the ability to predict outages, this paper provides insights into the performance of various ML algorithms for outage prediction. The experiments also reveal a high correlation between the number of historical outages in test dataset and model performance once the minimum data quality is satisfied.

*Field constraints include, but are not limited to, data acquisition, input processing and cleaning, data storage, outputs generation, etc., in real-time.

Neglecting the data availability and its impact can have detrimental effects on the accuracy and reliability of outage prediction models. Models may lack robustness and generalizability without determining the minimum training data quantity.

To address the problem of incorporating more data from adjacent geographical regions, we investigate two approaches for the spatial scope of input data. Using paired one-tailed Student's t-test in our experiment, we prove the statistically significant advantageous impact on model performance resulting from the incorporation of weather data from neighboring areas. The use of statistical testing provides a rigorous and quantitative assessment of the performance improvement achieved through the inclusion of spatial data. By comparing the model's performance with and without the incorporation of data from adjacent regions, the analysis demonstrates that the observed improvement is not due to random chance but is statistically significant.

3. Data Preparation

Cluster feeders in GIS

To investigate the impact of including weather parameters from adjacent areas, we divide the entire service area of a utility into segments where each segment of the grid consists of several feeders. We refer to these segments as *feeder clusters*. In our study, each feeder cluster represents a distinct segment of the grid comprising multiple feeders. We break down the system into feeder clusters so that we can train and test ML algorithms separately for each cluster. This enables us to compare the performance of the models when including or excluding weather data from adjacent feeder clusters.

The clusters can be formed in different ways: from having a single feeder in each cluster to having all feeders in a single cluster. The decision about the clustering technique can be based on a number of factors, such as the size of the service territory, number of feeders, feeder length, number of substations, or geographical conditions of the region (e.g., valleys, mountains, plains, forests). Clustering can be done manually, or it can be automated based on selected feeder characteristics. The final outage SoR reflects the outage probability in each of the feeder clusters created in this step. Options for clustering could be by voltage level, substations, closest weather station, etc. Optimization of clustering for performance maximization is outside the scope of this paper and is left for future research.

In our experiment, we segmented the distribution network that consists of 192 feeders into 3 clusters. The original map of the network (90x90 miles) with each feeder represented by distinct color and ASOS weather stations available in the area, is shown in Fig. 1. For each location in the network weather information is represented as min, max, and mean statistics from stations within a 25 km proximity. Clustering was performed using Arc GIS Pro (Esri). Arc GIS Pro has an automated way of clustering spatial features (feeders) included in the Spatial Statistics toolbox: *Build Balanced Zones*. We have used the option that sets the number of clusters and accounts for their total area, which allowed us to form clusters of approximately equal area. We have also used "compactness" criteria, which resulted in clusters being geographically compact. We created clusters comparable to each other in size but at the same time different regarding geographic locations. The difference in geographic location accounts for weather parameters difference, which allows better differentiation between events in each cluster. The map with the clustered network is shown in Fig. 2.

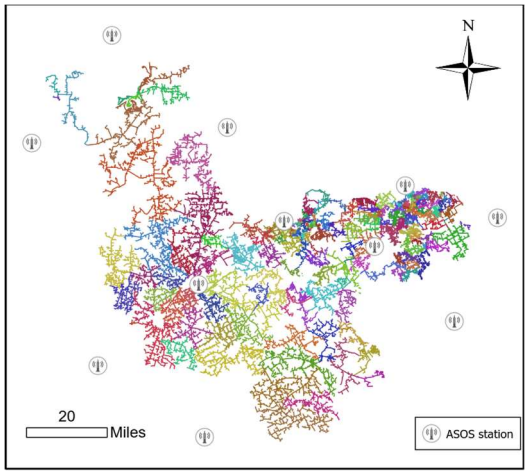

Figure. 1. Original network map.

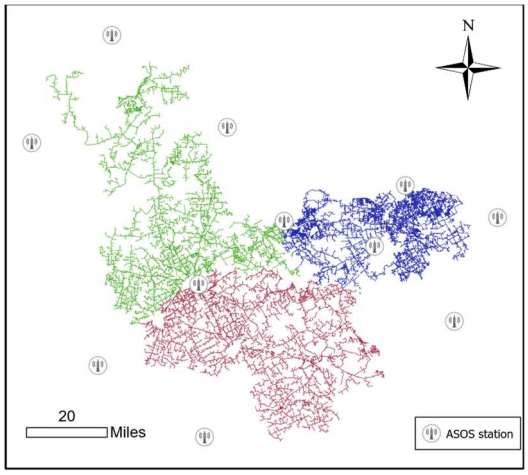

Figure. 2. Clustered electric grid map.

Spatiotemporal correlation of outages

To train the model, we utilize a historical outage dataset obtained from a utility company in Texas. In general, the historical outage dataset needs to reflect when and where the outages took place in the past. The underlying hypothesis is that the data model can identify specific patterns in the conditions that lead to an outage.

To use the dataset, one needs to spatially correlate each occurrence of an outage to the corresponding feeder cluster. The dataset provided by a utility company includes fields that represent outage start and end time, failed equipment ID, closest feeder, outage cause code, and comments of the repair crew. We have utilized the equipment ID and closest feeder fields to map each outage to the corresponding feeder cluster.

Our historical outage dataset was logged using the local time zone. Because most public datasets use UTC to report their timestamps, we have converted the outage timestamps from the local time zone to UTC and associated the events with the feeder cluster ID. Our study uses 6 years of historical outage data from January 2015 to December 2020.

Weather data preprocessing

The predominant causes of forced outages in the distribution system are weather conditions and vegetation intrusion. One needs to extract surrounding weather parameters to capture the environmental conditions at the time and location of outage occurrence. For this study, we have used a historical weather dataset collected by ASOS (*Automated Surface Observing Systems*, 2016). Iowa State University has created a convenient interface to fetch ASOS data (*Iowa Environmental Mesonet: ASOS-AWOS-METAR Data Download*). Their portal also supports data download by using an open-source Python script that we modified to our needs and used in our study (Herzmann). ASOS weather data has varying time resolutions, but most recent data has a 1-minute time resolution.

The obtained weather dataset contains the following weather parameters: weather station ID, weather station location, wind speed, direction, gust speed, air temperature, air pressure, dew point temperature, relative humidity, and one-hour precipitation.

For each weather station and each weather parameter, we: a) substitute string trace report filler value (T) for hourly precipitation with a small value (0.0001), b) substitute missing data with None values, c) convert all the values to floating-point type. Then we proceed with substituting None values by mean values for all parameters except wind direction and wind gust speed. For wind direction, we use median values to fill in missing data, and for wind gust speed, we use corresponding values from wind speed. We also drop duplicate values, keeping the first occurrence only.

The temporal correlation analysis of weather parameters is performed by first creating 1-hour timestamps for the full length of the period (2015-2020). For each hour and each weather station, the dataset is sliced temporarily, taking all values that were reported in the last hour. These values are aggregated using mean, min, and max functions. In such a manner, for each weather parameter, we obtained mean, min, and max values for each timestamp. Since the aggregation was performed on the temporal window, we refer to these values as temporal statistics. This aggregation allows one to characterize the weather in the last hour by just three values; however, one also loses information due to aggregation. Other techniques may be used to describe the data, for example, wavelet or Fourier transform. At the end of this step, weather parameter statistics are calculated for each hour for each weather station and saved.

Correlation of weather parameters to outages

Once weather data is preprocessed, we correlate it to the feeder clusters. We first define which weather stations are closest to each cluster. Operation is performed in ArcGIS using *a spatial join* tool: weather stations (points) are spatially joined with clusters (polygons) within a 25 km radius. As a result, weather stations get feeder cluster IDs assigned to them, creating cluster-weather station mapping. Based on this mapping, the weather data is pulled to clusters on each timestamp and then spatially aggregated for each cluster regardless of number of stations using min, max, and mean statistics. The hyperparameter of a 25 km radius was chosen empirically and may need to be tuned for each application considering the size of the clusters and the separation between them.

The last step calculates min, max, and mean spatially for each temporal statistic within a 25 km radius of each feeder cluster yielding 9 statistics for each weather parameter for each timestep. Here we aggregate the data again, but spatially as opposed to temporally in the previous step. The result characterizes weather conditions within the specified feeder cluster in the last hour. To decrease the number of features, we only keep the spatial minimum of temporal minimums, spatial maximum of temporal maximums, and spatial mean of temporal means, which gives us 3 values per weather parameter.

Now, the weather data is joined with outage data based on the corresponding timestamp and feeder cluster ID. We take weather parameters from the past hour as inputs and outages from the next hour as targets for our outage prediction models.

4. ML algorithm Selection and training

Algorithm types

We have implemented the prediction framework using five ML algorithms: Random Forest (RF), Catboost (Cat), Logistic Regression (LR), Support Vector Machine (SVM), and Neural Network (NN). These algorithms were chosen as representatives of different "families" of ML algorithms, aiming to explore potential variations in their performance and identify whether one type may outperform the others in the context of outage prediction.

Random Forest is a decision tree-based ensemble algorithm that combines multiple decision trees to make a more accurate and stable prediction (Breiman, 2001; Scornet, 2016). Each of the decision trees is trained on a random subset of the data and a random subset of the features. Such an approach helps to avoid overfitting and improves the accuracy of the model. The output of the RF is the majority vote of all the trees in the forest.

Catboost is an open-source gradient boosting type of algorithm based on decision trees and is capable of handling categorical data with almost no preprocessing (Baembitov et al., 2021; Dorogush et al., 2018). As opposed to RF, where the trees are used independently, Catboost uses gradient boosting to iteratively improve the predictive accuracy of a set of decision trees. The algorithm also possesses a built-in feature for importance estimation, which allows for determining the most impactful input parameters (Baembitov et al., 2023).

Logistic Regression is primarily used for classification tasks. It uses the logistic function to model the probability of the output prediction. It takes a linear combination of input features to create a log odds ratio and then puts it through the logit function to estimate the probability (Sperandei, 2014). LR is a simple and powerful algorithm that performs well on linearly separable data (Ezukwoke et al., 2019).

In the Support Vector Machine algorithm, a hyperplane is found that separates the classes by maximizing the margin between them. SVM is based on mapping the input data into a high-dimensional feature space using a kernel function, which allows for non-linear decision boundaries. The algorithm then finds the hyperplane that maximizes the margin between the classes in this feature space. The hyperplane is essentially a decision boundary that divides the input data into two or more classes. The margin is the distance between the hyperplane and the closest data points from each class (Cervantes et al., 2020), (Mohammadi et al., 2021).

A Neural Network is comprised of an input, output layer, and one or more hidden layers containing artificial neurons (Chasiotis et al., 2020). There are a few classes of NNs, in this study, we are using the Deep Feed Forward Neural Network, also referred to as Multi-Layer Perceptron (Hemeida et al., 2020). Neurons in NN have a non-linear activation function, which allows them to capture non-linear dependencies in data. We are using an Exponential Linear Unit (ELU) as an activation function, which produces negative outputs for negative inputs allowing to balance the mean activation of the neurons in the network (Qiumei et al., 2019).

For all the algorithms, we have used default hyperparameters as defined in related Python packages. For the NN, we used empirically determined 2 hidden layers with 30 artificial neurons in each layer. Kernel initializers are set to HE Uniform, and the activation is set to Exponential Linear Unit (ELU). The optimizer used for training is Adam, while the loss function is set to a binary focal cross-entropy from logits. The batch size was set to 16, and the number of epochs was set to 30.

Creating the training and test datasets

To address the algorithm sensitivity to temporal data availability and to obtain robust estimates of the model's performance, we employed nested cross validation (nCV) for forming temporal training and test datasets (Vu et al., 2022). As discussed in (Varma et al., 2006), nested cross-validation provides a nearly unbiased method for assessing performance.

To provide a more comprehensive understanding of nCV for the sensitivity study, we offer additional details below. The process of nCV is illustrated in Fig 3.

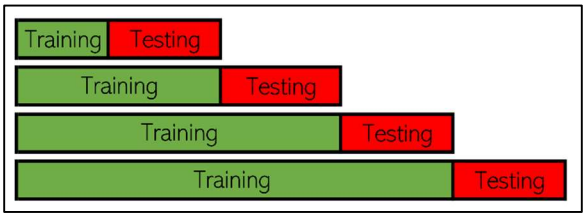

Figure 3. Nested Cross Validation

We start with three months of data to train the models, and we test them in the next three months. In the next fold, we expand the training dataset with the predefined time (Δ = 3 months in our experiments) and simultaneously shift the borders of the test dataset for the same value. In this manner, we repetitively increase the size of the training dataset and obtain test results in the next 3 months. nCV approach imitates the addition of new data once the solution is deployed into operation. The folds do not contain the same number of examples as in classical nCV, as the number of examples depends on the number of actual outages in the system. By using nCV, we ensure robust model assessment that accounts

for variations in data distribution and simulates real-world operational conditions. It enhances the reliability of our results and demonstrates the applicability of our approach to practical scenarios.

An outage in the distribution system is a rare event when looking from an hourly perspective at multiple years of data. That makes our dataset highly unbalanced. There are a total of 7,023 reported outages during the 6 years of available data. In this study, we will refer to outages as a positive class and refer to timestamps when no outage occurred as a negative class or normal operation (NO). In our study we do many experiments to ensure robustness of results.

To train the model, we subsampled the training dataset, which allowed us to decrease the number of negative class examples and use a more balanced dataset. We also applied a temporal exclusion window of 5 hours around the positive class (faults) to select NO examples. Such a technique provides the model with more distinctive features for each class. The downside is that model is not trained on examples in the temporal vicinity of the faults, making it harder for the model to classify such examples. When applying the exclusion window, we only consider the target cluster, that is, the cluster that the specific model is trained to predict outages. Such an approach leaves the "noisy" training examples, where the fault has occurred in a different cluster from the target cluster. For instance, when training the model for cluster 1, a "noisy" example would be a case when cluster 2 experiences an outage, but cluster 1 operates normally. We consider "noisy" examples as a negative class (NO). The inclusion of "noisy" examples helps to improve the models' ability for spatial differentiation. We apply the same approach to form the test datasets. To obtain metrics for several combinations of positive and negative classes, we create 10 test datasets for the same time period with a random sampling of negative classes. The varying composition of the test datasets allows one to acquire robust results for sensitivity study. Also, it creates additional cases for statistical testing to validate our hypothesis.

Spatial extent of features

In our study, we are using two sets of features in the training and test datasets. The first is comprised of features specific to the target cluster, where we limit the input weather parameters to the spatial extent of the target cluster plus the buffer radius (25 km). This approach prioritizes the weather around the specific cluster where the model is predicting the outage risk. The hypothesis behind such an approach is that outage prediction would be more accurate if in ML models influence of weather conditions is limited to the area of interest. We refer to this method as the "*cluster-specific*" method.

In an alternative setting, ML models are accounting for the weather parameters from all the available clusters in the network. That allows us to capture the weather not only over the target cluster but also the weather over the non-target clusters. The underlying hypothesis is that the consideration of the weather in the far-off territories from the target cluster would yield better performance for the model. The basic concept behind the second approach is that the current weather patterns or conditions observed over a certain distance can be a reliable indicator or predictor of the likelihood of power outages occurring in the future.

To differentiate between the two feature structures, we use a target-source notation. The target field indicates the target cluster as described earlier. The source field reflects the cluster(s) from which the weather parameters are taken. An integer in the source field indicates a specific cluster and corresponds to the "*specific cluster*" method. Letter "A" corresponds to the second approach and stands for "[A]ll of the available weather parameters."

All the algorithms are trained for cluster-specific outage prediction on the balanced dataset using both "cluster-specific" and "A" methods. Given 3 clusters in the network, we obtain 6 trained models for each algorithm type.

Data Scarcity Sensitivity

When deploying our approach, data availability for the initial training process is a problem that needs to be addressed. It is desirable to have enough data which would represent the latent relationship between outages and environmental conditions. In this study, we analyze the model performance depending on the amount of relevant training data. More data may provide valuable examples for the model making it more selective and powerful, but in practice, there is usually a saturation point, and further increase of the training data is not significantly improving the accuracy of the model. We perform a sensitivity analysis to study the effect of the size of the training dataset on the model's performance. A nested CV provides a natural way for such an experiment. As we break down the six years of data into chunks of 3 months, the models' performance is obtained for each testing fold and then can be averaged. The Nested CV helped us to obtain unbiased and robust estimates of the model's performance by preventing overfitting and potential data leakage. Also, through the repeated splitting of the data into different training and test sets, we were able to observe how the algorithms performed under varying data availability conditions.

5. Performance Comparison

Metrics

To compare the performance of different ML algorithms, we utilize 3 metrics: F-1 Score (F1), Average Precision (AP), and Area Under the Receiver Operator Curve (ROC) (Powers, 2011).

F-1 Score is a harmonic mean between Precision and Recall. Precision indicates how accurate the positive predictions of the model are, and Recall measures how well the model can identify all positive cases in the dataset. F1 is a metric that allows one to estimate the performance at a glance and is useful in cases where both false positives and false negatives are equally important (Hennessy et al., 2021).

AP is calculated by computing the area under the Precision-Recall curve, which plots the precision against the Recall at different classification thresholds. It represents the balance between Precision and Recall, where one needs to be sacrificed for the improvement of the other. AP is closely related to the Area under the Precision-Recall Curve (AUPRC); however, it is shown that AP does not yield overoptimistic results as compared to AUPRC (Davis et al., 2006). AP is useful in cases of heavily imbalanced datasets as it accounts more for a positive class (Saito et al., 2015).

The Receiver Operator Curve is formed by plotting the Recall against the false positive rate (FPR) at different classification thresholds showing the trade-off between them. Measuring the area under the Receiver Operator Curve summarizes the model performance in a single value. ROC metric reflects the likelihood of the model to rank a positive instance higher than a negative instance when both are chosen at random. It is less useful in unbalanced datasets (Fawcett, 2004), (Fawcett, 2006).

Any metric used to evaluate the performance of the model can reflect only a part of the model's capabilities, no single metric can provide a full image of the model's performance. It is essential to analyze several metrics to understand why one model performs better or worse than the other. It is also strongly advised to use business metrics in addition to mathematical metrics. Examples of such are an average reduction of repair time of an outage, yearly reduction of costs for equipment repair, impacts on grid reliability indices, change of revenue due to implementation of optimized mitigation measures, etc.

Business metrics reflect the ultimate impact of the model decisions on different stakeholders. These can also help in threshold selection, reflecting the penalty/award balance of the decisions for a stakeholder. Such thresholds need to be investigated separately for each individual case. The task of defining thresholds is outside the scope of this paper and is left for future research.

Student's t-test for the spatial extent of features

We have performed the paired one-tailed Student's t-test to verify the statistical significance of metrics improvement due to the inclusion of weather parameters from adjacent clusters (Hsu et al., 2014). We ran the test for each combination of model type (5), target cluster (3), and metric (3), totaling 45 tests. The null hypothesis is that for a given cluster, the performance metrics (AP, ROC, F1) are *equal* when including adjacent cluster features and when only using target cluster features. The alternative hypothesis is that the performance metrics for a model with adjacent cluster features are *greater* than the metrics of the model with target features only. The selected significance level is 0.05. The results have shown that in 36 tests, we can reject the null hypothesis and accept the alternative hypothesis, and in 9 tests, we cannot reject the null hypothesis. The results for the tests where the null hypothesis could not be rejected are presented in Table I. As can be seen from the table few cases of AP and ROC metrics did not pass the t-test. F1 metric, on the other hand, has passed the test in all cases.

Table I. T-test results for p-value greater than the significance level

Metric	Model Type	Target	Tstat	Pvalue
AP	NN	0	0.866	0.194
		1	0.912	0.181
		2	0.446	0.328
	RF	0	0.462	0.322
ROC	LR	0	1.275	0.102
		1	-0.951	0.829
	NN	0	-0.790	0.785
		2	-3.283	0.999
	SVM	1	0.415	0.339

One of the paper's contributions stems from comparing these results. We conclude that incorporating features from neighboring clusters leads to a consistent enhancement in the model's performance across most scenarios. It leads to a significant implication: the model's ability to forecast forced outages in the upcoming hour is not solely reliant on data within the focal region. Rather, it also depends on the weather conditions surrounding the area of interest. Since the weather conditions are "moving" across the region mainly by winds, information about the weather in adjacent regions reflects what would happen in the next hour. This finding also points to the necessity of including weather parameters from a wider area around the target cluster. Creating a dynamic buffer that is positively correlated to the wind speeds can further

improve the temporal predictive capabilities of the model. These topics are left for future research.

Based on the findings of this section, we have removed the models of the "*cluster-specific*" method from further analysis and only focus on the second method, which includes weather parameters from all clusters.

Performance analysis

Every model was evaluated on each of the test datasets of nested CV. The test was performed 10 times using the same positive class examples but different randomly selected negative class examples from the same CV fold. The aggregated results of average scores per fold are presented in Fig. 4, where the target dimension is omitted, so the results reflect the metrics for the entire network. The X-axis denotes temporal nCV folds, while the Y-axis displays metric values and fault counts per fold.

As can be seen from the figure, the model's performance is highly correlated to the number of outages in the given period (fold). The more outages the network experiences, the better the performance a model archives. We can conclude that the outage prediction approach is most effective when there is a large amount of positive class present, which usually happens under severe weather conditions.

Examination of metrics behavior illustrates another contribution of the paper. The AP metric increases for all models as more training data is used. The F1 score appears to have a strong positive correlation with the number of outages: it improves as the number of outages grows and drops when there are fewer outages. The ROC metric is very low in the case of a small-size training dataset, and it increases sharply when more training data is added. ROC metric is initially minimal with a limited training dataset, yet it shows a rapid rise as more training data is incorporated. This growth of the ROC occurs in the first 7-8 folds, which translates into 21-24 months of training data. Based on these findings, we can conclude that for practical applications, utilities need to accumulate at least around 2 years of relevant data. While addition of temporal data beyond 2 years marginally enhances performance, the incorporation of more diverse features, innovative feature engineering, and varied datasets could potentially lead to significant performance improvements.

With no substantial implementation advantage of any algorithm over the rest, we infer that tested algorithms possess comparable performance on the task of outage prediction using the weather parameters described in Section 3. We note that NN requires more data than the other algorithms to output comparable results, but at the same time, it tends to have lower fluctuations of the metrics when more training data is available, as can be seen from the F1 score. Both RF and SVM show higher initial F1 score values and both possess quicker training times. The practical takeaway is that utilities have the flexibility to select any of these algorithms for deployment purposes unless they wish to explore further feature engineering approaches which may benefit from diverse or enhance datasets.

The dependence of models' performance from the target cluster is shown in Table II, which averages the

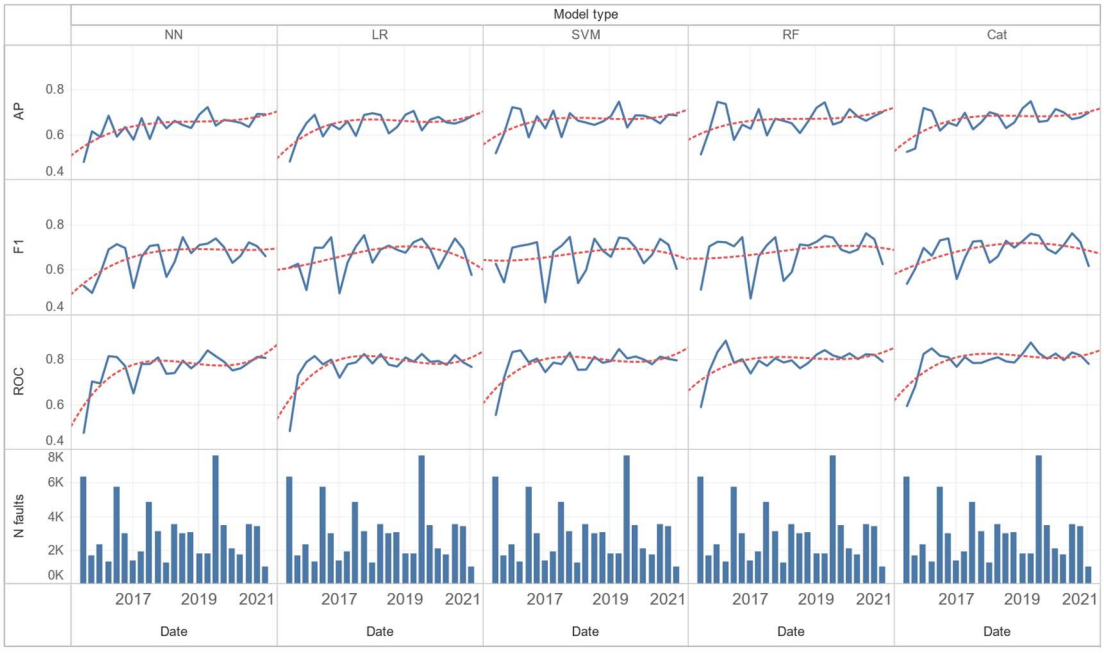

Figure 4. Nested CV results.

metrics for all algorithms and presents them on a per-target cluster basis. The results indicate that all metrics improve for cluster 2, which has the greatest number of outages. That corroborates our previous conclusion about the correlation between performance and the number of outages. Cluster 2 covers urban areas and is characterized by dense feeder locations. That explains the increased number of outages when compared to the other two clusters, which are in more rural areas.

Table II. Per target metrics.

Target	N faults	AP	F1	ROC	Precision	Recall
0	90,650	0.6489	0.6590	0.7741	0.7515	0.6325
1	99,600	0.6426	0.6559	0.7697	0.7500	0.6314
2	156,300	0.6809	0.7071	0.8097	0.7809	0.6766

6. Conclusions

In this paper, we have presented a sensitivity study of different ML algorithms used for predicting outage risks in the distribution system. Based on our findings, we make the following conclusions.

1. Models are sensitive to the geographical extent of input features. The inclusion of weather parameters from the adjacent territories of a feeder's cluster improves the model performance. This is a result of considering weather from a wider region that may be moving towards the target cluster.
2. All the considered models exhibited a similar level of sensitivity to the availability of data. To obtain satisfactory performance, at least 2 years of training data are needed to prepare a prediction model for deployment. A lower data quantity does not provide models with enough training examples.
3. The performance of the model is highly correlated with the number of forced outages in the system. When there is an abundance of forced outages, models detect outages more accurately, which may lead to faster restoration times.
4. Models for outage prediction are most accurate during harsh weather conditions. Inclement weather induces an increased amount of outage occurrences, making prediction models more applicable.

Acknowledgment

Data used in the project was provided by United Cooperative Services. We thank the United staff for their guidance and technical support. Funding is made available through the Texas A&M Engineering Experiment Station (TEES) Smart Grid Center Membership Agreement and National Science Foundation Award # 2125985.

7. References

Arif, A., & Wang, Z. (2018). Distribution Network Outage Data Analysis and Repair Time Prediction Using Deep Learning. *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, *00*, 1-6.

Automated Surface Observing Systems. (2016). NOAA's National Weather Service. Retrieved Mar. 2023 from https://www.weather.gov/asos/asostech

Baembitov, R., Kezunovic, M., . . . Obradovic, Z. (2023). Incorporating Wind Modeling Into Electric Grid Outage Risk Prediction and Mitigation Solution. *IEEE Access*, *11*, 4373-4380.

Baembitov, R., Kezunovic, M., . . . Obradovic, Z. (2021). Graph Embeddings for Outage Prediction. 53rd North American Power Symposium (NAPS 2021), College Station, TX, USA.

Bapin, Y., Ekisheva, S., . . . Zarikas, V. (2020, 18-21 Aug. 2020). Outage Data Analysis of the Overhead Transmission Lines in Kazakhstan Power System. 2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS),

Bin, S., Koval, D., . . . Shen, S. (1998, 25-28 May 1998). An analysis of extreme-weather-related transmission line outages. Conference Proceedings. IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No.98TH8341),

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.

Cervantes, J., Garcia-Lamont, F., . . . Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189-215.

Chasiotis, V. K., Tzempelikos, D. A., . . . Moustris, K. P. (2020). Artificial neural network modelling of moisture content evolution for convective drying of cylindrical quince slices. *Computers and Electronics in Agriculture*, *172*, 105074.

Davis, J., & Goadrich, M. (2006). *The relationship between Precision-Recall and ROC curves* Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA.

Dokic, T., & Kezunovic, M. (2019). Predictive Risk Management for Dynamic Tree Trimming Scheduling for Distribution Networks. *IEEE Transactions on Smart Grid*, *10*(5), 4776-4785.

Dorogush, A. V., Ershov, V., . . . Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *ArXiv e-prints*.

Eskandarpour, R., & Khodaei, A. (2017). Machine Learning Based Power Grid Outage Prediction in Response to Extreme Events. *IEEE Transactions on Power Systems*, *32*(4), 3315-3316.

Eskandarpour, R., & Khodaei, A. (2018). Leveraging Accuracy-Uncertainty Tradeoff in SVM to Achieve Highly Accurate Outage Predictions. *IEEE Transactions on Power Systems*, *33*(1), 1139-1141.

Esri. *ArcGIS Pro*. In Esri Inc.;. arcgis.com

Ezukwoke, K., & Zareian, S. (2019). Logistic Regression and Kernel Logistic Regression—A Comparative Study

of Logistic Regression and Kernel Logistic Regression for Binary Classification. *University Jean Monnet: Saint-Etienne, France*.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, *31*(1), 1-38.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Hemeida, A. M., Hassan, S. A., . . . El-Din, A. B. (2020). Nature-inspired algorithms for feed-forward neural network classifiers: A survey of one decade of research. *Ain Shams Engineering Journal*, *11*(3), 659-675.

Hennessy, P. J., Esau, T. J., . . . Corscadden, K. W. (2021). Hair Fescue and Sheep Sorrel Identification Using Deep Learning in Wild Blueberry Production. *Remote Sensing*, *13*(5), 943.

Herzmann, D. *IEM ASOS download service script*. In https://github.com/akrherz/iem/blob/main/scripts/asos/iem_scraper_example.py

Hirata, F. G. (2011). *Mapping and Modelling the Probability of Tree-Related Power Outages using Topographic, Climate, and Stand Data* University of British Columbia].

Hsu, H., & Lachenbruch, P. A. (2014). Paired t test. *Wiley StatsRef: statistics reference online*.

Iowa Environmental Mesonet: ASOS-AWOS-METAR Data Download. ASOS.

Kezunovic, M., & Dokic, T. (Nov., 2019). Big Data Framework for Predictive Risk Assessment of Weather Impacts on Electric Power Systems. Grid of the Future, CIGRE US Nat. Committee, Atlanta.

Kezunovic, M., Pinson, P., . . . Bessa, R. (2020). Big data analytics for future electricity grids. *Electric Power Systems Research*, *189*, 106788.

Khoshjahan, M., Baembitov, R., . . . Kezunovic, M. (2021, Oct.). Impacts of weather-related outages on DER participation in the wholesale market energy and ancillary services. CIGRE Grid of the Future Symposium, Providence, RI, USA.

Leistner, C., Saffari, A., . . . Bischof, H. (2009, 27 Sept.-4 Oct. 2009). On robustness of on-line boosting - a competitive study. 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops,

M. Kezunovic, R. Baembitov, . . . M. Khoshjahan. (2022). Data-driven State of Risk Prediction and Mitigation in Support of the Net-zero Carbon Electric Grid. 11th Bulk Power Systems Dynamics and Control Symposium – IREP'2022, Banff, Canada.

Mohammadi, M., Rashid, T. A., . . . Hosseinzadeh, M. (2021). A comprehensive survey and taxonomy of the SVM-based intrusion detection systems. *Journal of Network and Computer Applications*, *178*, 102983.

Nematirad, R., Ardehali, M. M., . . . Khorsandi, A. (2023). Optimization of Residential Demand Response Program Cost with Consideration for Occupants Thermal Comfort and Privacy. *ArXiv e-prints*.

Owerko, D., Gama, F., . . . Ribeiro, A. (2018, 26-29 Nov. 2018). Predicting Power Outages Using Graph Neural Networks. 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP),

Panteli, M., & Mancarella, P. (2015). Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies. *Electric Power Systems Research*, *127*, 259-270.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37--63.

Qiumei, Z., Dan, T., . . . Fenghua, W. (2019). Improved Convolutional Neural Network Based on Fast Exponentially Linear Unit Activation Function. *IEEE Access*, *7*, 151359-151367.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, *10*(3), e0118432.

Scornet, E. (2016). Random Forests and Kernel Methods. *IEEE Transactions on Information Theory*, *62*(3), 1485-1500.

Shashaani, S., Guikema, S. D., . . . Quiring, S. M. (2018). Multi-Stage Prediction for Zero-Inflated Hurricane Induced Power Outages. *IEEE Access*, *6*, 62432-62449.

Shi, K., Qiao, Y., . . . Lu, Z. (2018). An improved random forest model of short-term wind-power forecasting to enhance accuracy, efficiency, and robustness. *Wind Energy*, *21*(12), 1383-1394.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12-18.

Taylor, W. O., Nyame, S., . . . Anagnostou, E. N. (2023). Machine learning evaluation of storm-related transmission outage factors and risk. *Sustainable Energy, Grids and Networks*, *34*, 101016.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91.

Vu, H. L., Ng, K. T. W., . . . Hosseinipooya, S. A. (2022). Impacts of nested forward validation techniques on machine learning and regression waste disposal time series models. *Ecological Informatics*, *72*, 101897.

Zhang, W., Sheng, W., . . . Hu, L. (2018, 20-22 July 2018). Fault Prediction Method for Distribution Network Outage Based on Feature Selection and Ensemble Learning. 2018 5th International Conference on Information Science and Control Engineering (ICISCE),