4th URSI AT-RASC, Gran Canaria, 19-24 May 2024

An Efficient Machine Learning Model for Lightning Localization via Lightning-Induced Voltages on Transmission Lines

Mostafa Asadi[(1)], Hamidreza Karami* [(2)], Siavash Rajabi[(3)], Marcos Rubinstein[(2)], and Farhad Rachidi[(4)]
(1) Department of Electrical Engineering, Bu-Ali Sina University, Hamedan, Iran
(2) Institute for Information and Communication Technologies, University of Applied Sciences of Western Switzerland (HES-SO), Yverdon-les-Bains, Switzerland; e-mail: hamidreza.karami@heig-vd.ch
(3) Department of Electrical Engineering, Hamedan University of Technology, Hamedan 65155, Iran
(4) Electromagnetic Compatibility Laboratory, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

Abstract

In this paper, three machine learning (ML)-based approaches—XGBoost, Artificial Neural Network (ANN), and Random Forest algorithms—are compared for the localization of lightning through the analysis of lightning-induced voltages on power transmission lines. Across all methods, two sensors are employed to capture lightning-induced voltages on power transmission lines. Numerical simulations demonstrate that the XGBoost algorithm exhibits higher efficiency in terms of location accuracy and computational time compared to the other algorithms. Additionally, the Principal Component Analysis (PCA) algorithm is applied to reduce the dimensionality of XGBoost by 50 times without compromising accuracy, thereby accelerating calculation time and reducing computational resource usage. The R2 score obtained from the model on test data, with a Signal-to-Noise Ratio (SNR) of 30 dB, exceeded 99%, and for data with an SNR of 10 dB, it reached approximately 98%. Various configurations of transmission lines and sensor locations were tested, revealing that the accuracy of the model is dependent upon the transmission line configuration and sensor positions.

1. Introduction

Knowing the exact geolocation of a lightning strike is important in a wide range of research and application domains, including geophysical research, lightning warning, aviation/air traffic, weather services, insurance claims, power transmission and distribution, etc. [1]. The location of the lightning strike is generally obtained using the so-called lightning location system (LLS). LLSs detect cloud-to-ground (CG) discharge signals using electromagnetic VLF/LF range sensors [2].

Traditional methods of lightning localization are divided into magnetic direction finding (MDF), time of arrival (TOA), time difference of arrival (TDOA), and interferometer (ITF) [3]. TOA is a technique that has been used as a method for 2D/3D localization. This method requires at least three sensors to work properly [3], [4]. The MDF technique requires at least two sensors and is used for 2D localization [5].

More recently, methods based on Electromagnetic Time Reversal (EMTR) have been proposed as a means of locating lightning [6]. EMTR has been proven to have high accuracy in identifying the lightning impact point, but it requires at least three sensors to be accurate enough [7], [8].

In [1], a knowledge-based method is used to find the two-dimensional geo-locations of lightning impact points, which requires two sensors, and uses induced-voltage on power lines. In [9], a combination of the TDOA technique and artificial neural networks is proposed to locate lightning. Some studies have used deep learning to locate the lightning source [6], [10].

In [11], 3D radar data and machine learning algorithms such as k-nearest neighbors (KNN), random forest (RF), and convolutional neural networks (CNN) are used to identify lightning strike locations.

In this study, lightning strike detection is achieved through the implementation of three machine learning algorithms: XGBoost, Artificial Neural Network (ANN), and Random Forest. Due to page limitations, we present results exclusively for the XGBoost method. The algorithms are applied to lightning-induced voltage time series data obtained from only two sensors on two transmission lines. To improve computational efficiency, principal component analysis (PCA) is utilized for dimensionality reduction, leading to a significant reduction in calculation time while preserving model accuracy at an acceptable level. Furthermore, an optimization of the model is achieved by modifying the problem's geometry configuration, leading to the selection of the best sensor arrangement. All code implementations are executed in Google Colab for accessibility and convenience.

The structure of the article is outlined as follows. Section 2 presents the methodology, including the creation of the database, feature selection, and model generation. Section

3 delves into the training and testing of the generated model, utilizing numerical simulation results to estimate the lightning impact point Concluding the article, Section 4 presents final discussions and conclusions.

2. Methodology

2.1 Problem geometry and data acquisition

In this section, to estimate the geolocation of lightning impact point, we trained a machine learning model. The geometry of the problem is shown in Figure 1.

The problem geometry is considered as a 50×50 km^2 rectangular area with two different transmission lines. Two voltage sensors are considered on each transmission line. We defined 10000 uniformly random positions for the source within the considered area shown in Fig. 1. The simulation of the lightning-induced voltages has been done using Rusck's formula [12]. It should be noted that the generated data are characterized by a signal-to-noise ratio (SNR) of 30, 20, or 10 dB.

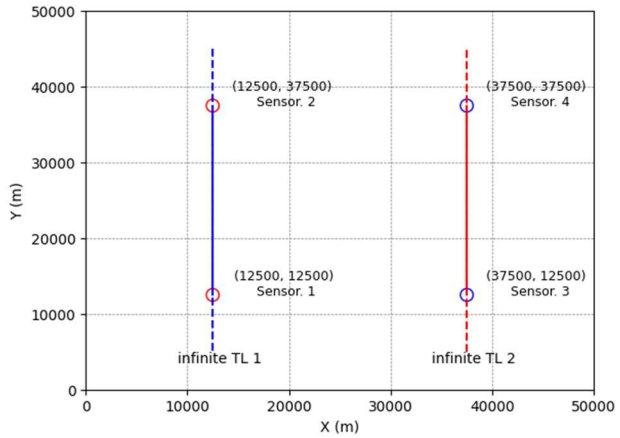

Figure 1. Geometry of the problem. Two transmission lines and four voltage sensors.

2.2 Data prepossessing

Cut-off: To make the simulated data representative of data obtained by real data recorders, a 10-V threshold is considered for the transient voltage.

Feature selection: It must be considered that the most valuable features in the lightning-induced waveform are in its early time. Therefore, to reduce the number of features to maintain system resources, the time window was selected to include only 2000 primary features of the induced lightning voltage wave to train the model. Furthermore, due to the cost reduction caused by sensor installation, finally used one sensor per transmission line. It should be noted that the data of two sensors were merged according to the following equation after extracting the features.

$$v_T(t) = v_A(t) + v_B(t),$$

where the indices A and B refer to the desired sensors and $v_T(t)$ is the data used in the next steps. Therefore, the number of features selected up to this step is 4000.

Normalization: Data normalization is one of the most important stages of data prepossessing. At this step, all the data were divided by the maximum value so that all the data were in the range of 0 and 1.

Train-Test splitting data: In order to evaluate the trained model, it is necessary to divide the data set into two parts, train and test data. We chose the segmentation ratio of 20 and 80 percent for test and training data, respectively.

Dimensionality reduction: PCA is a technique for reducing the dimensionality of data [13]. By reducing the number of features, PCA can help to:

- reduce the risk of overfitting a model to noisy features.

- Speed-up the training of a machine learning algorithm.

- Make simpler data visualizations. By adjusting n_components = 80 in the scikit-learn library, we reduced the number of input features from 4000 to 80.

2.3 Machine learning implementation

We have implemented three ML-based methods, namely, ANN, Random Forest, and XGBoost. Due to page limitations, we have provided the results specifically for the XGBoost method. XGBoost is a scalable end-to-end tree boosting system [14]. It is an implementation of gradient-boosting decision. It is a robust machine-learning algorithm that can help to understanding data and make better decisions. The XGBoost algorithm is called gradient boosting since the objective function is optimized using the gradient descent algorithm before each new model is added [1]. The objective function consists of two terms: The loss factor, which is the measure of the predictive power, and the regularization factor, which controls the complexity of the model and helps to avoid overfitting.

The XGBoost python package is used to build the classifier. The XGBoost tuned hyperparameters are: n_estimators = 400, max_depth = 60, eta = 0.045, subsample = 0.6, and colsample_bytree = 0.7.
The next step is to select a pair of sensors whose information leads to the generation of a high-accuracy model.

Table 1 shows the accuracy of the trained model using the data from six different pairs of sensors. According to Table 1, the best performance of the model is obtained using sensors 2 and 3. It is also shown that the calculation time is greatly reduced when using the PCA method, from more

than half an hour without using PCA method to less than 90 seconds with PCA.

Table 1. Accuracy of each trained model on the test data of selected pair of voltage sensors

Selected sensors no.	*1&2*	*1&3*	*1&4*	*2&3*	*2&4*	*3&4*
Accuracy (R2 Score)	0.93	0.92	0.99	0.99	0.92	0.93
PCA and fit time (sec)	88	91	85	86	89	85
Fit time Without PCA (sec)	1930	1870	1857	1856	1863	1898

3. Evaluation of the machine learning model

In this study, the evaluation of various machine learning algorithms, including ANN, Random Forest, and XGBoost, was conducted to ensure the selection of a robust and suitable model for lightning localization. Ultimately, the XGBoost algorithm was chosen as the primary machine learning algorithm due to its commendable accuracy and high efficiency. Subsequent activities in the study were carried out exclusively using XGBoost.

After selecting the model as well as the voltage sensors, the model was trained again. Before the model training step, 20% of the data were randomly separated by the scikit-learn train_test_split function as test data. the 10,000 simulated data, 8,000 were used as training data and 2,000 as test data. Also, to achieve the smallest feature dimensions, a process of finding the best accuracy in the smallest dimensions was implemented. Figure 2 shows the accuracy graph for configuration related to Figure 1. This layout is one of several layouts studied by the authors, which has been discussed more because of its specific results. This graph is related to the results of sensors 2 and 3. In this paper, two criteria R2 Score and Root mean squared error (RMSE) are used to evaluate the accuracy of the model. R2 score indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. It is a statistical measure that shows how close the regression line is to the actual data. RMSE measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value. According to the type of deployment of lines and selected sensors, the accuracy of the generated model is 99%. Its RMSE value is 1391 meters. As can be seen from Fig. 2, the accuracy along the x and y axes are different. The value of RMSE along the x-axis is 264 meters and along the y-axis is 1853 meters. Also, higher location errors along the x-axis can be observed for lightning strikes in the immediate vicinity of the sensors.

3.1 Impact of the geometry configuration

As mentioned earlier, the model training operation was done using a pair of sensors. According to Table 1, six different pairs of sensors in different locations were considered for training the models. It should be noted that the use of only one sensor or two sensors at a close distance from each other leads to a large error. The reason for this large error is that the use of only one sensor or two closely located sensors result in an ambiguity and cannot separate the occurrence of lightning on both sides of the transmission line.

In this paper, to investigate different layouts of transmission lines and their effect on the performance of the model, several layouts were investigated and it was observed that if the transmission lines are perpendicular to each other and in line with the sides of the geometry of the problem so that the area is covered by them, the model achieves better performance. Furthermore, if the lines are placed on the sides of the considered area, the accuracy increases. The results of four layouts, two lines parallel and perpendicular to each other as well as lines perpendicular to each other and on the side of the geometry of problem, are shown in Table 2. In the final layout of the transmission lines and sensors, according to Table 2, the accuracy of the model is 99.9 % and the RMSE error reached 400 meters.

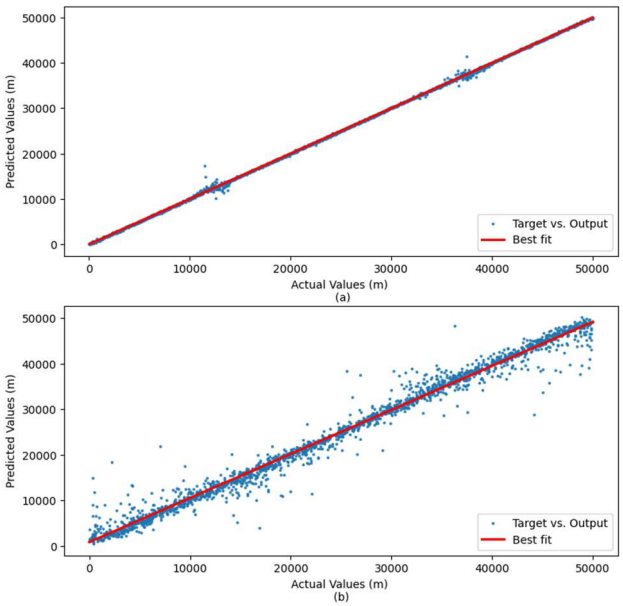

Figure 2. Model evaluation results for (a) the x-coordinate and (b) the y-coordinate of geometry configuration of Figure 1 (sensors 2&3)

3.2 Sensitivity to the Noise Level

The model training operation was performed using noisy data with an SNR of 30 dB. We repeated the training considering SNRs of 20 and 10 dB. In these cases, the accuracy of the model was still remarkable. For an SNR of 10 dB, the model achieved 98% accuracy.

4. Conclusion

In this paper, we showed the efficiency of the XGBoost algorithm to find the lightning location using noisy induced voltage data on power transmission lines. Notably, its robust performance is demonstrated even in the presence of noisy data. Using only two sensors, reducing the dimensions of features from 4000 to 80 by the PCA algorithm, and data with an SNR of 30 dB, the proposed model achieved 99% accuracy. Dimensionality reduction significantly increased the speed of calculations. It was also shown that the location of sensors and the configuration of the transmission lines can impact the performance of the model.

Table 2. Results of some lines and sensors layouts in problem geometry

Parallel transmission lines, TL1: X=12.5km , TL2: X=37.5km Sen1:(12.5km,37.5km), Sen2:(12.5km,12.5km), Sen3:(37.5km,37.5km), Sen4:(37.5km,12.5km)						
	sen1&2	sen1&3	sen1&4	sen2&3	sen2&4	sen3&4
R2	0.926	0.921	0.989	0.990	0.923	0.927
RMSE	3791.38	4098.99	1323.79	1391.51	3960.41	3951.09
perpendicular transmission lines: TL1: Y=5km , TL2: X=5km Sen1:(10km,5km), Sen2:(45km,5km), Sen3:(5km,10km), Sen4:(5km,45km)						
	sen1&2	sen1&3	sen1&4	sen2&3	sen2&4	sen3&4
R2	0.992	0.992	0.993	0.993	0.994	0.992
RMSE(m)	1277.5	1238.06	1129.66	1093.45	1037.04	1277.5
perpendicular transmission lines: TL1: X = 25km , TL2: Y=25km Sen1:(25km,37.5km), Sen2:(25km,25km), Sen3:(12.5km,25km), Sen4:(37.5km,25km)						
	sen1&2	sen1&3	sen1&4	sen2&3	sen2&4	sen3&4
R2	0.452	0.960	0.965	0.975	0.970	0.467
RMSE	10736.3	2593.03	2548.96	1946.2	2441.98	10646.4
Perpendicular transmission lines: TL1: Y=0 , TL2: X = 0 Sen1:(0km,0km), Sen2:(50km,0km), Sen3:(0km,0km), Sen4:(0km,50km)						
	1&2	1&3	1&4	2&3	2&4	3&4
R2	0.998	0.998	0.999	0.999	0.998	0.997
RMSE	658.39	513.19	447.0	400.02	374.6	681.36

References

[1] H. Karami, A. Mostajabi, M. Azadifar, M. Rubinstein, C. Zhuang, and F. Rachidi, "Machine Learning-Based Lightning Localization Algorithm Using Lightning-Induced Voltages on Transmission Lines," IEEE Trans. Electromagn. Compat., pp. 1–8, Mar. 2020.

[2] W. Schulz, V. A. Rakov, and M. Bernardi, "Review of CIGRE Report Cloud-to-Ground Lightning Parameters Derived from Lightning Location Systems – The Effects of System Performance," 2009.

[3] A. Alammari et al., "Lightning mapping: Techniques, challenges, and opportunities," IEEE Access, vol. 8, pp. 190064–190082, 2020.

[4] T. Tantisattayakul, K. Masugata, I. Kitamura, and K. Kontani, "Broadband VHF sources locating system using arrival-time differences for mapping of lightning discharge process," J. Atmos. Solar-Terrestrial Phys., vol. 67, no. 1031–1039, 2005.

[5] E. P. Krider, R. C. Noggle, and M. A. Uman, "A Gated, Wideband Magnetic Direction Finder for Lightning Return Strokes," J. Appl. Meteorol. Climatol., vol. 15, no. 3, pp. 301–306, Mar. 1976.

[6] A. Mostajabi, H. Karami, M. Azadifar, A. Ghasemi, M. Rubinstein, and F. Rachidi, "Single-Sensor Source Localization Using Electromagnetic Time Reversal and Deep Transfer Learning: Application to Lightning," Sci. Rep., vol. 9, no. 1, pp. 1–14, Dec. 2019.

[7] G. Lugrin, N. M. Parra, F. Rachidi, M. Rubinstein, and G. Diendorfer, "On the location of lightning discharges using time reversal of electromagnetic fields," IEEE Trans. Electromagn. Compat., vol. 56, no. 1, pp. 149–158, 2014.

[8] T. Wang, S. Qiu, L.-H. Shi, and Y. Li, "Broadband VHF Localization of Lightning Radiation Sources by EMTR," IEEE Trans. Electromagn. Compat., vol. 59, no. 6, pp. 1949–1957, Dec. 2017.

[9] K. Mehranzamir, Z. Abdul-Malek, H. Nabipour Afrouzi, S. Vahabi Mashak, C. leong Wooi, and R. Zarei, "Artificial neural network application in an implemented lightning locating system," J. Atmos. Solar-Terrestrial Phys., vol. 210, p. 105437, Nov. 2020.

[10] X. Wang, K. Hu, Y. Wu, and W. Zhou, "A Survey of Deep Learning-Based Lightning Prediction," Atmos. 2023, Vol. 14, Page 1698, vol. 14, no. 11, p. 1698, Nov. 2023.

[11] M. Lu et al., "Lightning Strike Location Identification Based on 3D Weather Radar Data," Front. Environ. Sci., vol. 9, p. 714067, Aug. 2021.

[12] S. Rusck, Induced Lightning Over-Voltages on Power-Transmission Lines With Special Reference to the Over-Voltage Protection of Low-Voltage Networks. Stockholm, Sweden: KTH, 1958.

[13] I. T. Jollife and J. Cadima, "Principal component analysis: a review and recent developments," Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., vol. 374, no. 2065, Apr. 2016.

[14] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-August-2016, pp. 785–794, Mar. 2016.