Partial Discharge Detection with Convolutional Neural Networks

Wei Wang
Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, CA 92521
Email: wwang031@ucr.edu

Nanpeng Yu
Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, CA 92521
Email: nyu@ece.ucr.edu

***Abstract*—Covered conductors are widely adopted in the medium to low voltage systems to prevent faults and ignitions from events such vegetation contacting with distribution lines and conductors slapping together. However, such events could cause partial discharge in deteriorated insulation system of covered conductors and ultimately lead to failure and ignition. To prevent power outages and wildfires, it is crucial to detect partial discharges of overhead power lines and perform predictive maintenance. In this paper, we develop advanced machine learning algorithms to detect partial discharge by using measurements from high frequency voltage sensors. Our innovative approach synergistically combines the merits of spectrogram feature extraction and deep convolutional neural networks. The proposed algorithms are validated using real-world noisy voltage measurements. Compared to the benchmark, our approach achieves notably better performance. Furthermore, the classification results from the neural networks are interpreted with an occlusion map, which identifies suspicious time intervals when partial discharges occur.**

***Index Terms*—Partial discharge, fire prevention, short-time Fourier transform, convolutional neural networks.**

I. Introduction

Power line faults that result in sparks and ignite nearby tree branches contributed to the increase in the number of wildfires in the United States. California had the worst wildfires in the nation in 2018 with damages valued at more than 3.5 billion dollars [1]. It has been confirmed that the deadly Camp Fire in 2018 was caused by the power line owned by Pacific Gas and Electric Company (PG&E) [2]. Although enhanced vegetation management can significantly reduce wildfire risks, it can be very labor-intensive and costly.

An alternative solution to mitigate power line caused ignitions is to replace existing bare conductors with covered conductors. Upgrading bare conductors to covered power lines will significantly reduce fault current from a few amps to milliamps when a foreign object contacts with power lines. In most cases, this reduction in fault current and energy prevents ignition [3]. However, when vegetation comes in contact with covered conductors or conductors slap together, partial discharge could occur in deteriorated insulation systems. A partial discharge (PD) is a small electrical spark that occurs across the surface of insulating material where the electric field strength exceeds the breakdown strength of the insulating material [4], [5], [6], [7]. Thus, it is crucial to detect PD incidents and prevent ignitions and deadly wildfires.

The PD detection problem has been extensively studied in the existing literature from many different aspects. A detailed description for the mechanism and phenomena of PDs in capacitors, transformers, rotating machines, and power cables is provided in [7], [8]. Comprehensive experiments and measurements for PD events are performed with various sensors such as inductors and Rogowski coils [6], [7], [8]. The PD detection problem is formulated as a classification problem and solved with different methods such as statistical learning, signal processing algorithms, support vector machine [9] and artificial neural networks [5]. Deep neural networks were applied to detect PD problems in recent years. The 1D convolutional network is adopted in [10] to detect PD. The long short-term memory model is used to classify different categories of PD signals in power electronic devices [11]. However, most PD detection methods are developed and evaluated based on high quality signals measured in laboratory environment. Recognizing this problem, researchers recently started investigating PD detection with real-world noisy measurements from high frequency voltage sensors. By selecting input features based on domain knowledge and adopting the random forecast algorithm, state-of-art performance for PD detection was achieved and verified in field deployment [3].

In this paper, we propose an innovative data-driven approach to detect PD. The unique contributions of our proposed method are three-fold. First, we overcome the challenges associated with noisy real-world measurements by deploying high-pass filters and discrete wavelet transform (DWT) to remove the 50Hz base waveform and suppress high-frequency noise. Second, we convert the PD detection task from a time series classification problem to an image classification problem with short-term Fourier transform. Compared to the ad hoc feature extraction approach [3], our proposed feature transformation from 1D voltage time series to 2D spectrogram better preserves useful information in the high-dimensional dataset. Third, by combining two state-of-the-art convolutional neural networks to solve the image classification problem, our proposed algorithm significantly improves the PD detection performance. Finally, by plotting the occlusion map of the trained neural networks, we identified time intervals when

978-1-7281-2822-1/20/$31.00

partial discharges may occur.

The remainder of the paper is organized as follows. Section II describes the PD detection problem and provides a detailed description of the real-world noisy voltage sensor data. The technical methods for the PD detection problem are presented in Section III. Section IV conducts a comprehensive validation study for the benchmark and our proposed algorithms. Section IV concludes the paper.

II. Problem Description

This paper aims to detect the presence of PD in high frequency noisy voltage measurements. With sufficient PD labels, the PD detection problem can be naturally framed as a time series classification problem. In this section, we first provides an overview of the real-world high frequency voltage sensor data. Then, we discuss the technical challenges associated with detecting PD events in the real-wold dataset.

A. Data Description

We gather the power line voltage measurements from an online repository [12] to perform PD detection. The voltage measurements are recorded by inductors installed on three-phase power lines of a electric grid whose operating frequency is 50Hz. The voltage measurements are collected from all three phase wires simultaneously. Each single-phase voltage signal sample contains 800,000 measurements over 20ms, which covers a full operating cycle. The voltage magnitudes are scaled and rounded into integers within the range of $[-128, 127]$. Two voltage signal samples with and without PD are plotted in the Fig. 1. The binary labels indicating whether PDs occur during the individual voltage time series are also provided. The PD detection problem is framed as a binary classification task.

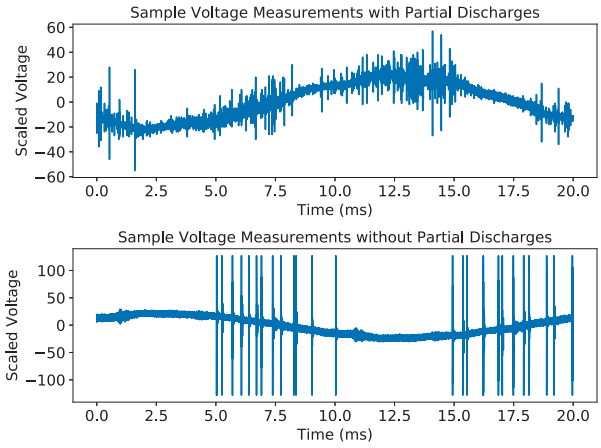

Fig. 1. Sample voltage measurements with and without partial discharges

B. Technical Challenges

There are three major technical challenges in detecting PDs based on noisy and high frequency real-world voltage measurements and label data. First, the real-world voltage measurements contain a significant amount of background noise from several sources including radio emissions, power electronic devices, lightning, switching operations, and corona. In addition, the background noise time series of different voltage sensors do not have a consistent pattern. For example, radio stations with varying distances to the power lines broadcast radio waves during different operating hours.

Second, the dimensionality of the voltage signals is significantly larger than the sample size. There are only 8,712 voltage time series available and each signal contains 800,000 measurement points. Moreover, the information regarding the start and end times of PD events in the time series are not available. In general, it is extremely difficult to identify the sub-sequences associated with PDs in each raw voltage time series. Therefore, in order to avoid over-fitting, we must retain the most distinguishable features for PD events by performing effective data preprocessing and dimensionality reduction.

Third, the sample dataset is highly imbalanced, where only 525 or 6% of the voltage signals have PD events. The highly imbalanced dataset can easily lead to either high false negatives or high false positives.

III. Technical Methods

The overall framework of data-driven PD detection contains three stages, signal preprocessing, data conversion, and sample classification. In the first stage, we apply signal preprocessing techniques on the raw voltage time series data. In the second stage, we convert 1D preprocessed voltage time series into 2D spectrogram with short-time Fourier transformation. In the third stage, we adopt convolutional neural networks to classify the images corresponding to voltage time series into groups with and without PDs.

A. Signal Preprocessing

In the signal preprocessing stage, we first take two de-noising steps to remove the normal frequency component of the voltage data and the high frequency background noise. At last, we apply max-pooling to reduce the dimensionality of the high frequency voltage time series.

The 50Hz component of the voltage signal is always present in a voltage waveform. Thus, it does not provide any useful information for the PD detection problem. To remove the 50Hz component, we first normalize the original voltage signal with integer data points into the range of [-1, 1] and then pass it through a discrete-time high-pass filter:

$$x^h[n] = \alpha x^h[n-1] + \alpha(x[n] - x[n-1]), \qquad (1)$$

where $x^h[n]$ and $x[n]$ denote the n-th output and input data points of the high-pass filter. The coefficient α is set as $1/(2\pi\Delta T f_c)$, where ΔT is the time interval between sampled points and f_c is the cut-off frequency.

To remove the high frequency background noise in the voltage time series without diminishing the voltage spikes of

interests, we apply the discrete wavelet transform (DWT) [13] to decompose the outputs of the high-pass filter as follows:

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k} W_\phi[j_0, k]\phi_{j_0,k}[n] + \frac{1}{\sqrt{N}} \sum_{j=j_0}^{\infty} \sum_{k} W_\psi[j, k]\psi_{j,k}[n], \quad (2)$$

where x, ϕ, and ψ are the discrete signal, the scaling function, and the wavelet function. n denotes the index of data points. W_ϕ and W_ψ are the approximation and detail coefficients. $j \in \mathbb{Z}$ is the dilation factor and j_0 is the base dilation factor. $k \in \mathbb{Z}$ is the translation integer. In practice, the DWT can be implemented by passing the discrete signal through a cascaded filter bank. The high frequency background noise can be suppressed by truncating the detail coefficients [14]. The truncation threshold at dilation level j is calculated based on the mean absolute deviation (MAD) of the absolute value of the detail coefficients $W_\psi[j, \cdot]$ as [15]:

$$threshold_j = \frac{1}{0.6745} MAD(|W_\psi[j, \cdot]|)\sqrt{2 \ln N} \quad (3)$$

where N is the number of data points in the time series.

The denoised voltage signals have extremely high resolution. To reduce the dimensionality of the voltage timer series, we perform max-pooling, which keeps the largest value for every N_m points. This operation maintains the positive spikes, which could indicate the presence of PDs [3].

B. Convert 1D Signal to 2D Image

PDs could occur in any subsequence(s) of the preprocessed 1D voltage time series. Thus, it is more appropriate to extract information from each of the subsequences in the preprocessed voltage time series. This can be achieved by short-term Fourier transform (STFT), which has been widely applied in sound recognition and enhancement [16]. The discrete STFT first divides the whole time series into subsequences with overlap. Then, the discrete Fourier transform (DFT) is performed with the multiplication of the window function and the signal function over each subsequence. Denote the subsequence size or window size as N_w and the overlap between consecutive subsequences as N_o. The hop size is defined as $N_h = N_w - N_o$. For the i-th subsequence, the STFT is performed as:

$$X(i, k) = \sum_{m=0}^{N_w-1} x[iN_h + m]w[m]e^{\frac{-j2\pi km}{N}}, 0 \le k \le N-1, \quad (4)$$

where $X(i, k)$ is the STFT of subsequence i with frequency k, w is the window function with window size N_w.

Now, we can finally convert the 1D voltage time series into a 2D image representing the log-spectrogram, which can be obtained by $spectrogram(i, k) = log(|X(i, k)|^2)$.

C. Convolutional Neural Networks

The PD detection task is converted from a time series classification problem to an image classification problem in Subsection III.B. In this subsection, we use convolutional neural networks (CNNs) to detect if PD is present in the images, which correspond to the sample voltage time series. We adopt two state-of-the-art CNNs, VggNet[17] and ResNet [18] to solve the image classification problem.

The architectures of the two CNNs are shown in Fig. 2. Both networks are constructed with basic building blocks. The basic building block of VggNet consists of stacked convolution layers and a max-pooling layer, where the max-pooling layer is used to gradually reduce the feature dimensionality. The ResNet introduces a by-passing route from the input to the output in the basic building block by hypothesizing that it is easier to optimize the residual mapping $F(x_i, W_i) = o_i - x_i$, where x_i, o_i and W_i are the input, output and weights of the block respectively. For both building blocks, the batch-normalization [19] layer is adopted after each convolution layer to relieve the gradient vanishing problem for deep networks and accelerate network training.

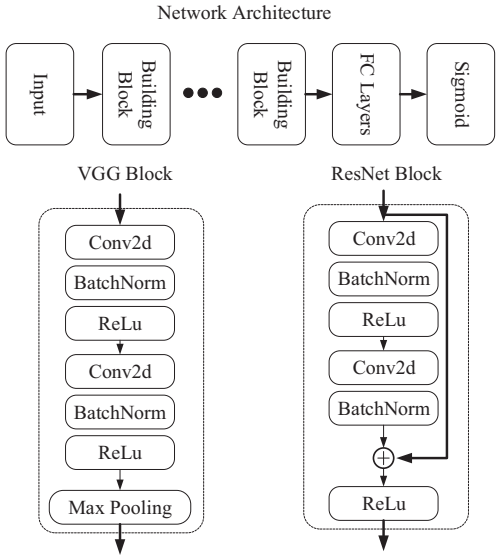

Fig. 2. Network architecture and building blocks of VggNet and ResNet

The fully connected (FC) hidden layers with ReLU activation function are used right after the building blocks. For the binary classification problem, the output layer is a single neuron with the Sigmoid activation function. The binary cross-entropy is chosen as the loss function to be minimized:

$$f_{loss} = -y \ln(\tilde{y}) - (1 - y) \ln(1 - \tilde{y}) \quad (5)$$

where $y \in \{0, 1\}$ is the true binary class label and $\tilde{y} \in [0, 1]$ is the prediction from the neural network.

As the dataset is highly unbalanced, we create balanced training batches, which contain the same amount of samples with and without PD patterns. To further reduce the variance

of the model ouptut, a bag of the ResNets and VggNets are trained and ensembled [20] with average voting.

IV. Numerical Study

A. Benchmark Algorithms

The empirical feature extraction based approaches [3], [21] are selected as two of the benchmarks. In these approaches, every signal is divided into four parts, which are the 90 degree sub-intervals of a sine wave. The expert selected statistics of the peaks in all four parts are used as features, which include the mean value, max value and max ratio between consecutive peaks. The random forest [3] and gradient boosted tree [21] are used as classifiers in the benchmark.

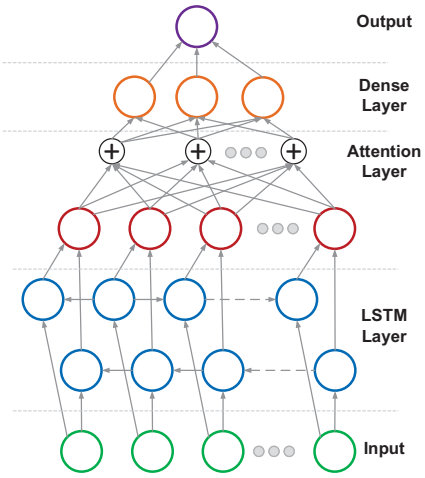

Fig. 3. Network architecture of bi-directional LSTM with attention mechanism

The LSTM model is selected as the third benchmark, which is one of the most popular neural networks to deal with time-series data. The network architecture of recently proposed bi-directional LSTM with attention mechanism [22] is adopted and depicted in Figure 3. The bidirectional LSTM consists of two layers, one layer with the same direction as the data sequence and the other layer with the reverse direction. The bidirectional structure has been shown to further improve the performance in both time-series classification [23] and regression [24] problems. Moreover, the attention mechanism allows modeling of dependencies without regard to their distance in the input or output sequences. The LSTM model typically could handle time series with hundreds of time steps. To apply the LSTM model for the voltage measurements with 800,000 time steps, we divide each signal into 160 periods. The statistics of the peaks are collected over the 5,000 measurement points of each period. The sizes of the LSTM layer, attention layer, and fully connected layer are chosen as 256, 512, and 256 respectively. Note that the ensemble method is also adopted. To be consistent, it is applied for all the benchmark and the proposed algorithms.

B. Data Pre-processing and Model Training

The normalized voltage measurement data is accompanied with the 50Hz base waveform and high frequency noises as shown in the first subplot in Fig. 4. The denoised voltage measurements of the sample signal is shown in the second subplot, where the base waveform and the high frequency noise are removed and suppressed successfully with the high-pass filtering and DWT. The cut-off frequency of the high-pass filter is $f_c = 100Hz$ and Haar wavelet is chosen for the DWT.

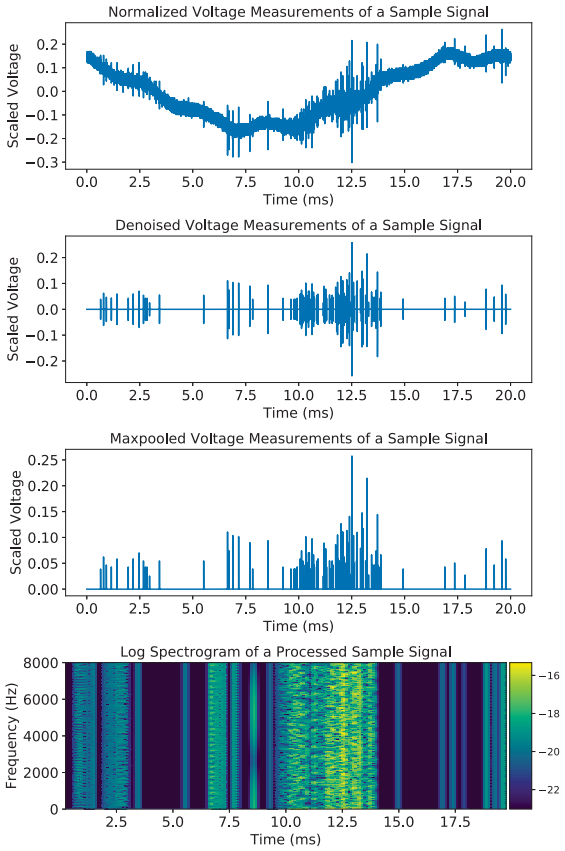

Fig. 4. Preprocessed voltage time series data

As shown in the third subplot, the dimensionality of the original signal is reduced from 800,000 to 16,000 with max-pooling over every 50 data points. At last, the STFT transforms the 1D signal into a 2D spectrogram with a dimensionality of (201, 196) by performing DFT on every 400 data points with hop size 80. As shown in the last subplot, the spectrogram extracts features in both frequency and time domains.

The ResNet with 18 layers (ResNet18) and VggNet with 11 layers (VggNet11) are chosen as the classifiers based on the size of the spectrogram. To train each model, the data is randomly split into a training set (60%), a validation set (20%), and a test set (20%). A CNN is trained over 100 epochs with a batch size of 128. The oversampling technique is adopted during training to handle the unbalanced classes issue, where the same amount of samples with or without PDs are collected

in each batch. The model version with the best Matthews correlation coefficient (MCC) in the validation set is chosen as the final model for testing purpose. Moreover, the best threshold turning the continuous output of neural networks to binary labels is determined by exhaustive search in range (0, 1) with a step size of 0.001 on the validation results.

C. Performance Evaluation

Matthews correlation coefficient (MCC), a common measure of the quality of binary classification algorithms, is used to evaluate the performance of the benchmark and proposed algorithms. The MCC is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (6)$$

where TP, TN, FP, and FN are the true positive rate, true negative rate, false positive rate and false negative rate respectively. MCC is deemed as a balanced measure for the datasets with binary classes of different sizes.

The output $\tilde{y}$ of all algorithms are in the range of $[0, 1]$. The final binary class label is obtained by $\mathbb{1}(\tilde{y} > y_{thr})$. $\mathbb{1}$ is the indicator function and $y_{thr} \in [0, 1]$ is the threshold. For each algorithm, a bag of models are trained by repeating the 5-fold cross-validation multiple times with different random seed. The threshold that achieves the best MCC score on the validation data using the average voting results is selected as the optimal threshold.

D. Effect of Noise Filtering on Convolutional Networks

The CNNs are generally robust to noise. However, the presence of noise may reduce learning efficiency for our small dataset with only 8712 samples. Thus, we test the performance of CNNs with and without noise filtering. For both ResNet18 and VggNet11, a bag of 20 CNNs are trained. The MCC scores of the average voting results over the training, validation, and testing data are as shown in Table I. As ResNet optimizes over residuals, it is more likely to be affected by spikes and non-Gaussian noises from a small dataset. However, the noise filtering step could remove some useful information in the raw data. This is why noise filtering causes the performance of VggNet to deteriorate.

TABLE I
MCC SCORES OF PARTIAL DISCHARGE DETECTION METHODS

Method	Noise Filtering	Train	Validation	Test
ResNet18	Yes	1	0.761	0.738
	No	0.997	0.693	0.656
VggNet11	Yes	0.997	0.761	0.729
	No	0.979	0.769	0.744

As shown in Table I, ResNet, which optimizes the residuals over the by-passing route is more sensitive to noise. The test MCC score drops significantly (11%) when the noise filter is removed. In contrast, VggNet is not very sensitive to noise filtering. The test MCC score actually increases by 2% when the noise filter is removed.

E. Performance Comparison

The MCC scores over the training, validation, and testing data of the benchmark (random forest, gradient boosted tree, and bidirectional LSTM) and the proposed algorithms are provided in Table II. For the benchmark algorithms, the average voting is conducted over a bag of 125 models for the random forest and the gradient boosted tree and a bag of 20 models for the bidirectional LSTM. The final proposed algorithm blends the two CNNs together by average voting.

TABLE II
MCC AND ACCURACY SCORES OF PARTIAL DISCHARGE DETECTION METHODS

Method	MCC / Accuracy		
	Train	Validation	Test
Random Forest	0.926 / 0.991	0.714 / 0.968	0.712 / 0.968
Gradient Boosted Tree	0.887 / 0.986	0.721 / 0.969	0.720 / 0.969
Bidirectional LSTM	0.919 / 0.990	0.661 / 0.959	0.621 / 0.955
Resnet18	1 / 1	0.761 / 0.973	0.738 / 0.970
VggNet11	0.979 / 0.998	0.769 / 0.971	0.744 / 0.968
Resnet18 + VggNet11	1 / 1	0.791 / 0.977	0.757 / 0.973

As shown in Table II, the accuracy scores of all methods are very similar because the majority of the samples is without PD. The MCC is a better performance metric for extremely unbalanced dataset. It can be seen that the two CNNs based methods achieve better performance than the empirical feature extraction based methods. The PD detection performance is further improved by blending the two CNNs and introducing diversity in network structure. The 5.1% improvement in MCC of the combined CNN over the gradient boosted tree is notable considering the extremely unbalanced dataset and the small sample size.

F. Interpretation of CNNs

The methods for interpreting CNNs include filter visualization, saliency map and occlusion experiment [25]. As PD patterns do not have any obvious or specific shapes, we adopt the occlusion map to visualize and understand trained CNNs. The occlusion map can be constructed as follows. For each time step, a small portion of the image is occluded with a mask. Then, the same classifier makes a prediction based on the new image. At last, the occlusion map is obtained by stitching prediction results together from shifting the mask across the whole image. The occluding mask is chosen as a zero matrix with size 10×10. The stride size of mask shifting on both vertical and horizontal directions is set as 2.

Figure 5 shows the results of occlusion experiment of a voltage signal with partial discharge (class value 1). As shown in the third subplot, the predicted class value has a significant dip when the mask is placed around 4.5 and 12.5 ms with frequency range from 2000 Hz to 4000 Hz. The corresponding time intervals in the original time series could be identified as potential voltage signatures for PD events.

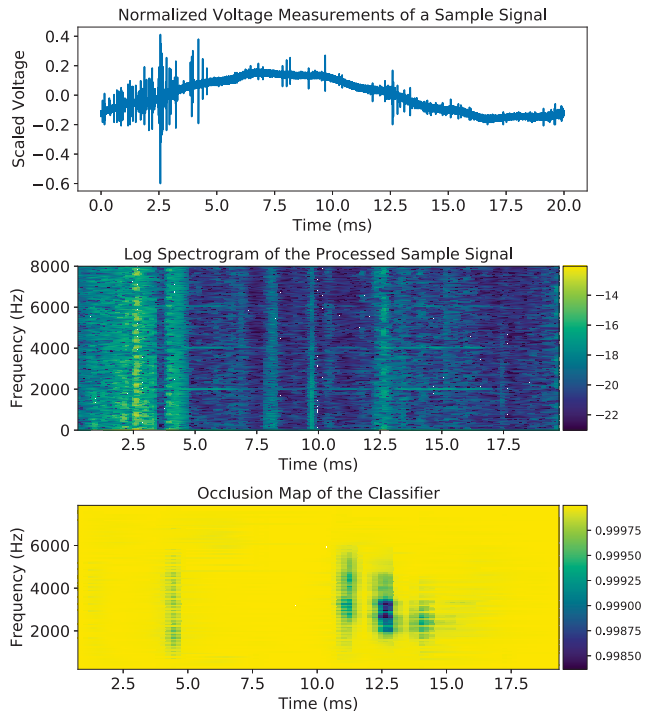

Fig. 5. CNNs interpretation with occlusion map

V. Conclusion

This paper proposes an innovative machine learning algorithm to detect partial discharges of covered conductors. The proposed algorithm not only helps assess the health condition of covered power lines but also prevents potential ignitions. By transforming the 1D voltage signal to 2D spectrogram, we can leverage the powerful CNNs to detect partial discharges based on noisy real-world voltage sensor data. Although the CNNs are generally robust to noises, the noise filtering step can improve learning efficiency and accuracy of ResNet. A comprehensive evaluation is conducted for both benchmark and the proposed machine learning algorithms. The evaluation results show that our proposed algorithms achieve notably improved MCC scores. In addition, we construct an occlusion map of CNNs, which helps identify suspicious time intervals when partial discharges occur.

References

[1] Wiki. (2018) California wildfires. [Online]. Available: https://en.wikipedia.org/wiki/2018_California_wildfires

[2] California Department of Forestry and Fire Protection. (2019, May) Cal fire investigators determine cause of the camp fire. [Online]. Available: https://www.fire.ca.gov/media/5038/campfire_cause.pdf

[3] S. Misk, J. Fulnecek, T. Vantuch, T. Burinek, and T. Jezowicz, "A complex classification approach of partial discharges from covered conductors in real environment," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 24, no. 2, pp. 1097–1104, Apr. 2017.

[4] S. Mik and V. Pokorn, "Testing of a covered conductors fault detectors," *IEEE Transactions on Power Delivery*, vol. 30, no. 3, pp. 1096–1103, Jun. 2015.

[5] N. C. Sahoo, M. M. A. Salama, and R. Bartnikas, "Trends in partial discharge pattern classification: a survey," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 12, no. 2, pp. 248–264, Apr. 2005.

[6] G. M. Hashmi, M. Lehtonen, and M. Nordman, "Modeling and experimental verification of on-line PD detection in MV covered-conductor overhead networks," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 17, no. 1, pp. 167–180, Feb. 2010.

[7] H. A. Illias, M. A. Tunio, A. H. A. Bakar, H. Mokhlis, and G. Chen, "Partial discharge phenomena within an artificial void in cable insulation geometry: experimental validation and simulation," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 23, no. 1, pp. 451–459, Feb. 2016.

[8] R. Bartnikas, "Partial discharges. their mechanism, detection and measurement," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 9, no. 5, pp. 763–808, Oct. 2002.

[9] L. Hao and P. L. Lewin, "Partial discharge source discrimination using a support vector machine," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 17, no. 1, pp. 189–197, Feb. 2010.

[10] M. A. Khan, J. Choo, and Y.-H. Kim, "End-to-end partial discharge detection in power cables via time-domain convolutional neural networks," *Journal of Electrical Engineering & Technology*, vol. 14, no. 3, pp. 1299–1309, 2019.

[11] E. Balouji, T. Hammarstrm, and T. McKelvey, "Partial discharge classification in power electronics applications using machine learning," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov. 2019, pp. 1–5.

[12] ENET Centre at Technical University of Ostrava. VSB power line fault detection data. [Online]. Available: https://www.kaggle.com/c/vsb-power-line-fault-detection/data

[13] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[14] D. L. Donoho and I. M. Johnstone, "Threshold selection for wavelet shrinkage of noisy data," in *Proceedings of 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, Nov. 1994, pp. 24–A25.

[15] T. Vantuch, "Analysis of time series data," Ph.D. dissertation, Technical University of Ostrava, 2018.

[16] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, May 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 07–09 Jul 2015, pp. 448–456.

[20] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15.

[21] Kaggle. (2019) VSB power line fault detection. [Online]. Available: https://www.kaggle.com/c/vsb-power-line-fault-detection/discussion/87038#latest-521846

[22] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, Aug. 2016, pp. 207–212.

[23] S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional long short-term memory networks for relation classification," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, Oct. 2015, pp. 73–78. [Online]. Available: https://www.aclweb.org/anthology/Y15-1009

[24] S. Wang, X. Wang, S. Wang, and D. Wang, "Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting," *International Journal of Electrical Power & Energy Systems*, vol. 109, pp. 470 – 479, 2019.

[25] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, Sept. 2014, pp. 818–833.