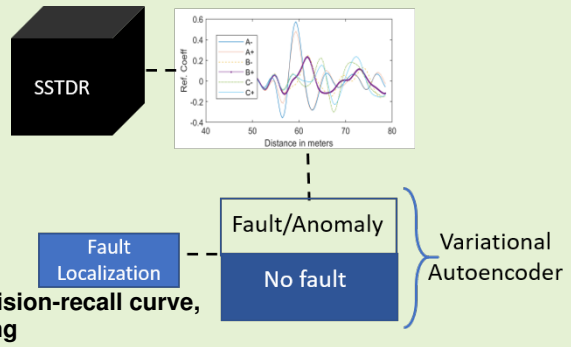
Anomaly detection of disconnects using SSTDR and Variational Autoencoders

Ayobami S. Edun, *Student Member, IEEE*, Cody LaFlamme, Samuel Kingston, Cynthia M. Furse, *Fellow, IEEE*, Michael A. Scarpulla, *Senior Member, IEEE*, and Joel B. Harley, *Member, IEEE*

***Abstract*— This article utilizes variational autoencoder (VAE) and spread spectrum time domain reflectometry (SSTDR) to detect, isolate, and characterize anomalous data (or faults) in a photovoltaic (PV) array. The goal is to learn the distribution of non-faulty input signals, inspect the reconstruction error of test signals, flag anomalies, and then locate or characterize the anomalous data using a predicted baseline rather than a fixed baseline that might be too rigid. The use of VAE handles imbalanced data better than other methods used for classification of PV faults because of its unsupervised nature. We consider only disconnects in this work, and our results show an overall accuracy of 96% for detecting true negatives (non-faulty data), a 99% true positive rate of detecting anomalies, 0.997 area under the ROC curve, 0.99 area under the precision-recall curve, and a maximum percentage absolute relative error of 0.40% in locating the faults on a 5-panel setup with a 59.13 m leader cable.**



***Index Terms*— Variational Autoencoders, reflectometry, SSTDR, disconnects, faults**

I. Introduction

ANOMALY detection is an important task in several fields, requiring the ability to detect and isolate unwanted data, input, or outcomes (”anomalies”, also sometimes called ”outliers”). If these anomalies are not detected, isolated, or analysed, they can corrupt the data, distort results, and can sometimes lead to wrong conclusions. To detect such anomalies, algorithms such as nearest neighbors [1], clustering [2], and variational autoencoders [3] have been used. In this work, we combine the abilities of a variational autoencoder to capture a low dimensional representation of input data with spread spectrum time domain reflectometry (SSTDR) to detect anomalies in a photovoltaic (PV) array and subsequently characterize such anomalies for fault detection.

To probe an electrical signal, SSTDR [4] uses a modulated pseudo-noise (PN) code as the incident signal. This signal reflects at points of impedance mismatch to the SSTDR device [5] where it is measured. The time delay between the incident and reflected signals gives the distance to the fault while the strength of the impedance mismatch can be estimated from the amplitude and phase of the reflection [6]. An advantage of SSTDR is that it does not depend on an IV curve like in [7]–[9], it requires just one device unlike the work in [10] that requires a temperature sensor, voltage, current and other measurement sensors, and SSTDR can be used on a live/energized system. However, multiple reflections from the system can make the reflected signal more complex to analyze which is why we need better algorithms, like in this work, to analyze the reflected signals.

Prior works have shown the viability of SSTDR for detecting disconnection faults [11], locating faults [12], characterizing lumped elements [6] in PV arrays, detecting degradation in MOSFETS [13], and detecting degraded/aged cells in a Li-ion battery pack [14]. Authors in [12] gave an overview of the abilities of SSTDR to detect faults, such as ground faults, arc faults, broken cells or modules, open circuit faults, and others. They also provided quantitative comparisons of the SSTDR reflections in each scenario. In most of these works, baseline subtraction is used to classify and locate faults. To classify faults, each test data can be correlated with the baseline and if the correlation coefficient is below some threshold, the test data can be classified as a fault. To locate the fault, a baseline taken from the system when there is no fault is compared with (i.e., subtracted from) the response from a system with a fault to locate the fault. A challenge with baseline subtraction for PV systems is that the reflectometry responses are highly affected by environmental conditions, such as illuminance,

This work was supported by the U.S. Department of Energy’s Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) under Agreement DE-EE0008169.

Ayobami S. Edun, Cody LaFlamme, and Joel B. Harley are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: aedun@ufl.edu; flyingcodfish@ufl.edu; joel.harley@ufl.edu).

Samuel R. Kingston, Evan J. Benoit, and Michael Scarpulla are with the Department of Electrical and Computer Engineering, The University of Utah, Salt Lake City, UT 84112 USA (e-mail: samuel.kingston@utah.edu; evan.benoit@utah.edu; mike.scarpulla@utah.edu).

Cynthia M. Furse is with the Department of Electrical and Computer Engineering, The University of Utah, Salt Lake City, UT 84112 USA, and also with LiveWire Test Labs, Inc., Salt Lake City, UT 84117 USA (e-mail: cfurse@ece.utah.edu).

temperature, and humidity [15].

To mitigate environmental effects, and as an alternative to using correlation coefficient, machine learning techniques have been used with SSTDR to detect, classify, and locate faults in a PV array. For example, authors in [16] used K-SVD (an unsupervised dictionary learning algorithm) and DK-SVD (a supervised dictionary learning algorithm) with SSTDR to detect and classify disconnection faults in a solar array while being robust to the effect of environment on SSTDR signals. While the work in [16] achieved great results, there are two main drawbacks of the work. First, the assumption that faults are a linear combination of some basis signals is not always true. For example, in the presence of moisture, the relative permittivity of the cable changes and subsequently causes a change in the velocity of propagation (VOP) of the underlying SSTDR signal. The effect of this VOP change varies differently in the region of the panels from the region with only cables leading to a non-linear effect. Secondly, if the dictionaries are updated online, they can be filled with one class of data over time and hence become non-representative. This challenge occurs because the amount of anomalous (or faulty) data is much less than that of the non-faulty data, which is often referred to as a data imbalance problem. Hence, unstable results might be obtained if the SSTDR device is deployed to the field with this algorithm.

To address these issues, we introduce a three-step approach. First, we design a variational autoencoder [17] (VAE) that learns the probability distribution of baselines (non-faulty data). Secondly, the VAE network is used for anomaly detection since anomalies are expected to have a distribution different from that of the baselines. Thirdly, the data flagged as anomalies are then inspected to detect, locate faults, or characterize such anomalies. To locate the faults, the VAE predicts the best baseline from its latent space to be used for baseline subtraction rather than using a fixed baseline which may be too rigid. This is another advantage of our method. Variational autoencoders have been used in various fields, such as speech processing [18], natural language processing [19], and image processing [20]. Specifically for photovoltaics, VAEs have been used to detect microcracks in photovoltaic silicon wafers [21] and for short-term forecasting of photovoltaic power production [22]. Yet, there is no work that combines VAE with SSTDR for detecting anomalies and locating faults.

VAE is used for discriminative and generative tasks because it has the ability to represent complex data in a low-dimensional latent space for numerous purposes and specifically, anomaly detection. Our approach uses a VAE to train a generative model using non-faulty baseline data. Our goal is for the VAE to learn to generate an optimal baseline signal without faults that the current measurement can be compared with to aid in the detection and localization of anomalous fault signatures. The algorithm relies on an encoder to project the input data to a low-dimensional space called latent variables. These latent variables represent the mean and standard deviation of each dimension in this low-dimension space. The decoder then samples from the latent variables to reconstruct each input data (baseline). The objective of the VAE is to reduce the reconstruction loss as much as possible so that anomalies can be flagged by examining the reconstruction error of the test data.

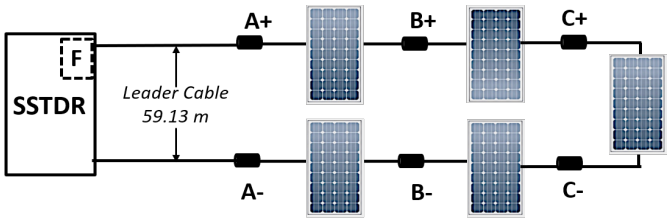


Fig. 1: Experimental Setup

In this work, we combine the ability of SSTDR to probe an energized PV array intermittently with the representation ability of variational autoencoders to detect anomalous SSTDR signals. These anomalous signals are then further inspected to characterize them. The goal is to learn the distribution of non-faulty input signals, inspect the reconstruction error of test signals, flag anomalies, and then locate or characterize the anomalous data. In prior work that uses DK-SVD [16], sparse coefficients of a learned dictionary were used to classify faults, but the learned dictionary can become unrepresentative over time if updated online and populated with data of the same class. Our approach removes the need to depend on the sparse coefficient of a dictionary to determine if a signal is faulty or not and avoids the possibility of having an unrepresentative dictionary. Also, by learning a distribution of the non-faulty (non-anomalous) data, it removes the assumption made by DK-SVD that a signal is a linear combination of some basis signals. We consider only disconnection faults in this work, and our results show an accuracy of 96% for detecting true negatives (non-faulty data), a 99% true positive rate of detecting anomalies, and an overall accuracy of 97% for detecting true negatives (non-faulty data) and true positives (anomalies) on a 5-panel setup with a 59.13m leader cable.

The rest of this article is organized as follows. In Section III, we give an overview of variational inference and variational autoencoders and how faults can be detected and located through them. Our experimental setup and data generation process are detailed in section II, while the results are presented in section IV. Finally, in section V, we give our conclusions and future work.

II. Experimental Setup

Fig. 1 shows the experimental setup used to collect data. The VAE algorithm was then deployed on the data and the VAE results are later compared to two other algorithms: DK-SVD and correlation coefficient. A Wilma SSTDR [5] was connected to five (5) 36-cell PV panels with a leader cable of 59.13m. The characteristics of the panels are same as in [16]. SSTDR has two test modes: static and intermittent. In static test mode, the PV array setup is probed once to determine if there is a fault or not. In the intermittent test modes, the SSTDR input signal is sent about 256 times in a second to constantly monitor or gather data from the array. The intermittent test mode was used for our experiments.

The intermittent experiment was conducted for six days in October 2019. While the SSTDR was connected to the PV

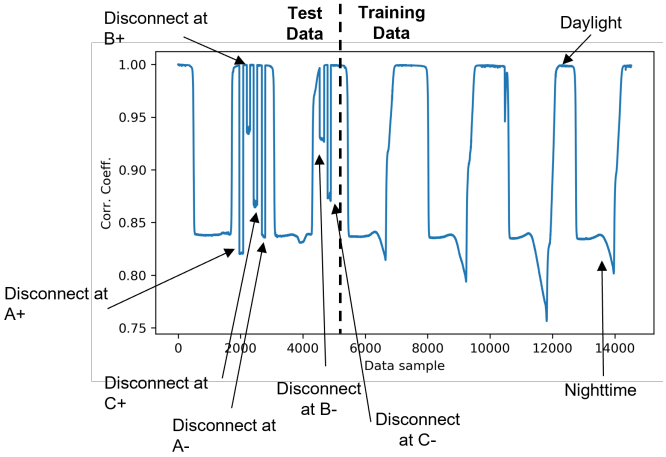


Fig. 2: Correlation coefficient of data samples with the first data sample

arrays, we induced disconnections at each location (A+, A-, B+, B-, C+, C-) as annotated in Fig. 1. The MC4 connector at each location was disconnected for about 1 hour to gather enough test data. Our experiments resulted in a data set of about $7,642,308$ reflection signatures (or measurements) for the six days. To reduce run-time and memory consumption of the algorithm, we considered only every 500^{th} signal in the full data set, resulting in $14,528$ measurements. This downsampled dataset is then split into two sets to obtain the training data and test data. The test data contains the first 35% of the downsampled experimental data, resulting in $5,085$ SSTDR measurements with 14% of faulty data. The remaining 65% of the data, resulting in $9,443$ SSTDR measurements, were used for training and they contain only non-faulty SSTDR measurements. To validate our algorithm performance, we split the training data into 70% training set and 30% validation set.

Fig. 2 shows the correlation, or cosine similarity, between each data sample with the first data sample in the data set (i.e., a baseline). Correlation values close to 1 represent high similarity to the baseline and correspond to data taken in daylight. The dips correspond to data collected at nighttime as well as other induced faults as marked in the figure. The dotted line shows the demarcation between test data and training data. Note that the test data and training data correspond to data collected on different days. A sample of daylight, nighttime, and disconnect data are shown in Fig. 3.

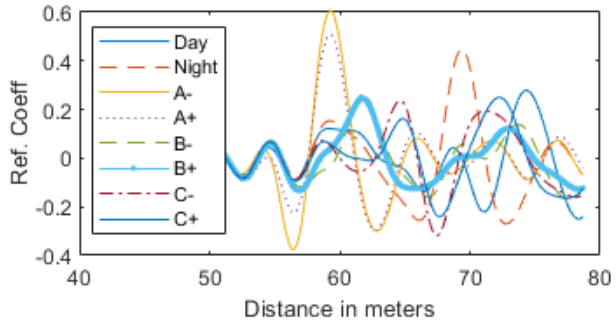


Fig. 3: Time domain reflection signature of daylight, nighttime, disconnects at locations A, B and C..

III. Fault Detection and Localization with Variational Autoencoders

A typical autoencoder consists of an encoder network and a decoder network. In this method, the encoder network, which is a neural network with posterior distribution $q_\theta(\mathbf{z}|\mathbf{x})$, transforms the input data $\mathbf{x}$ into a lower-dimensional representation, called the latent variables $\mathbf{z}$. The corresponding low-dimensional latent variable space is learned by the encoder such that it has enough information to accurately reconstruct the training data. The decoder network, also a neural network but with a likelihood $p_\phi(\mathbf{x}|\mathbf{z})$, uses these latent variables to reconstruct the input data while minimizing the reconstruction loss. To ensure the latent variables are representative of the input data and the reconstruction loss is minimized, the encoder and decoder are trained together and optimized via backpropagation with a chosen learning rate.

The latent variables of a typical autoencoder are numbers on the real line (i.e., unconstrained vectors). In a variational autoencoder, instead of encoding input data as a single point (as in a typical autoencoder), we encode the input data as a distribution over the latent space. Hence, the latent variables are regularized using the Kullback–Leibler (KL) divergence [23] during the training stage to ensure they can be used to generate new sets of data. In other words, the latent variables are encoded as a distribution and trained to learn the mean and covariance of the distribution. The distribution returned by the encoder is enforced to be close to a standard normal distribution. Figure 4 shows the VAE network used in this paper. The encoder network has three hidden layers with 32, 16, and 8 neurons, respectively, and a latent space of dimension 2. We chose a dimension of 2 with the intuition that one latent variable should control reflections from the cable while the second latent variable controls the reflections from the panels. When using a 3D latent space, there was no significant difference in the results. Each input and output layer has a dimension of 82, which corresponds to the length of the SSTDR data. The network was trained with a learning rate of 1×10^{-4} (chosen after tuning with the training data) and the ADAM optimizer [24].

The VAE loss function is given as

$$L(\theta, \phi) = \lambda \mathbb{E}_{z \sim q_\theta(\mathbf{z}|\mathbf{x})} \log(p_\phi(\mathbf{x}|\mathbf{z})) - \mathrm{KL}(q_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \quad (1)$$

where θ and ϕ are the parameters of each distribution. The first term in the loss function represents the reconstruction loss

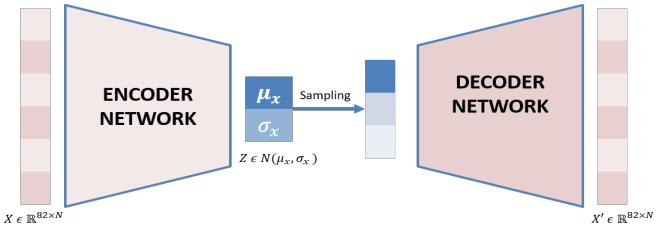


Fig. 4: Variational Autoencoder Network

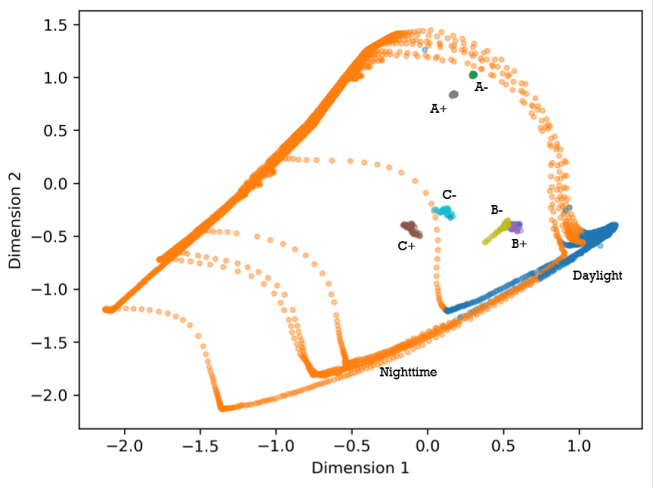


Fig. 5: Latent space of the mean of input samples learned using our VAE network where blue dots corresponds to data obtained during the day, orange dots corresponds to data obtained at night, and data with faults are as labelled.

(chosen to be the mean absolute error) and the second term ensures that the learned distribution q of the encoder is similar to the true prior distribution p. The operator $\mathbb{E}_z$ represents the expected value with respect to the distribution of z and $\mathrm{KL}(\cdot)$ represents the KL-divergence [23]. Note that the reconstruction loss was given a weight, λ of 1000 to place more emphasis on the reconstruction error. Several hyperparameters were tuned manually including the number of layers, the dimension of the latent space, the learning rate, and the weight on the regularization. A search was done over a range of parameter values. This includes learning rates of 0.1,0.01,0.001, and 0.0001; regularization weight of 1, 10, 100, and 1000; and a latent space dimension of 2 and 3. The chosen weight of 1000 gave the best separability of training data in the latent space. Also, the weight allowed us to avoid the Kullback-Leibler (KL) divergence collapse problem where the latent space becomes unrepresentative and collapses into a single point [19].

A. Obtaining Baselines from the VAE

In various works with SSTDR, baseline subtraction has been used to detect and locate faults [25]. In these schemes, the test signal is subtracted from a stored baseline to isolate changes caused by the faults. Such baselines can be obtained when the PV array is set up before operations or intermittently when there is no fault in the setup. This method is useful because it removes reflections from cables and connectors. However, the baseline is highly affected by environmental effects, making them unstable for localization [16], [26], [27].

In our work, data were obtained during the daylight and at nighttime. The data that forms our baselines were collected when there was no fault in the PV array setup. Sample daylight and nighttime time domain SSTDR data are shown in Fig. 3. Fig. 5 shows the corresponding latent space of the data obtained during the daylight (in blue) and at nighttime (in orange). Observe the wide variation within the mean of the baselines in a two-dimensional latent space. We observed that as dimension 1 of the latent space decreases from positive to negative, the magnitude of the input SSTDR around the 60^{th} sample increases. This region corresponds to reflections from the solar panels. Hence, each baseline when used for a baseline subtraction will produce different results thereby leading to instability of fault location.

To find the best baseline for locating faults in a test signal, the variational autoencoder (VAE) can be used. Recall that the VAE is trained on only baseline (non-faulty) data without any example of faulty data. Therefore, by the end of the training, the architecture learns the distribution of the non-faulty data and can act as a generative network. Specifically, when a test data representing a faulty condition is passed through the architecture, the network tries to reconstruct the test data with minimal error. However, because the network only knows the distribution of the non-faulty data, it effectively searches for and generates a baseline signal that minimizes the reconstruction error. Hence, the reconstructed signal serves as the best baseline for estimating the location of the faults.

B. Fault Detection and Location

Our VAE network is trained with non-faulty training data (baselines) to learn the distribution of the latent space while minimizing the reconstruction error. For each training data that is passed through the VAE network, we record the reconstruction loss, that is, the mean absolute error of each input training data and its reconstruction.

To detect faults, we set a threshold using all of the training reconstruction losses. The threshold is set to be the mean of the training reconstruction loss plus 3 times the standard deviation of the training reconstruction loss. To detect faults, the test data is passed through the VAE network which creates a latent space representation from the encoder and then reconstructs the data with the decoder. The reconstruction loss for the test data is then compared to the threshold learned from the training data. Any test data whose loss is greater than the threshold is deemed a fault (i.e., an anomaly). A threshold of mean plus 3 standard deviations was chosen because based on statistics, we know that 99% of the data should lie within 3 standard deviations. We confirmed using the precision-recall curve that this threshold gave a high precision and recall.

As explained in Section III-A, the reconstructed data serves as the best baseline for locating faults in the test data. Hence, to locate faults, the test data is subtracted from the reconstructed data (i.e., the predicted VAE baseline). The peak of the baseline subtracted data is then inspected to locate the fault distance because the strength of the fault can be estimated from the peak of the reflected signal.

In summary, in the training stage, we optimize the network to minimize the reconstruction loss of the training data, we set a threshold using the reconstruction loss. The test data is flagged as a fault if the reconstruction loss exceeds the set threshold. This fault is then inspected for the location of the fault using baseline subtraction where the baseline is the reconstructed output of the decoder. The flowchart of the process is shown in Fig. 6.

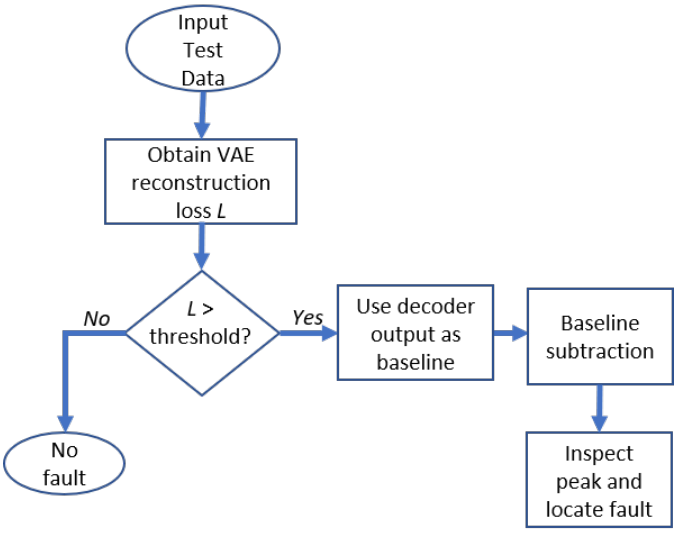


Fig. 6: Flowchart showing the procedure to detect faults (or anomalies) and to locate the faults.

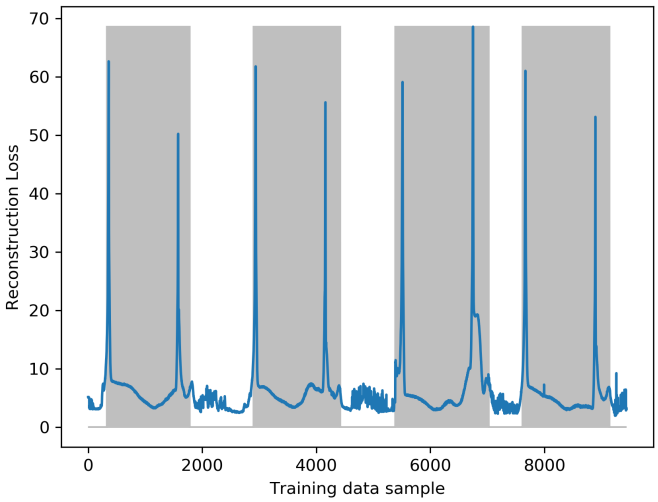


Fig. 8: Reconstruction Loss for training data with the training label where the grey and white background depicts baseline measurements taken at nighttime and at daylight respectively.

IV. Results

The VAE was trained for 100 epochs with parameters described in section III. Fig. 7 shows the training loss. Observe the decrease in the loss as the number of epochs increases. No significant reduction was obtained for epochs above 100.

A. Anomaly Detection

Fig. 8 shows the reconstruction loss for the training data where the grey background depicts regions of baseline measurements taken at nighttime, while the other baseline measurements were taken in daylight. We note that the spikes in the reconstruction loss corresponds to transitions from day to night and night to day. To detect anomalies in the test data, we obtain the threshold from the training reconstruction loss. For our experiment, the threshold is set to be the mean of the training reconstruction loss plus 3 times the standard deviation of the reconstruction loss and is estimated to be 23.28.

From the test data set, we created a vector with a length equal to the number of samples in the test data. All daylight and nighttime data were given a value of 0, while data that corresponds to a fault were given a value of 1. This vector represents a ground truth that shows if a test data is faulty or not. We pass the test data through the VAE network and record the respective reconstruction loss. Any test data whose loss is greater than the threshold is deemed an anomaly and sets the anomaly value to 1. The predicted anomaly value (0 or 1) is compared to the ground truth to obtain the anomaly accuracy.

Fig. 9 shows the reconstruction loss of the test data and the threshold. Comparing Fig. 9 with the test data portion in Fig. 2, we observe that all faults are above the threshold. The figure also reveals that all daylight data are below the threshold but data during the transition from daylight to nighttime and vice-versa are above threshold. This is a consequence of the high reconstruction loss during transition phases, which we saw in Fig. 8. This also reveals the high variability of data we obtain at nighttime. The algorithm shows a 96.61% and a

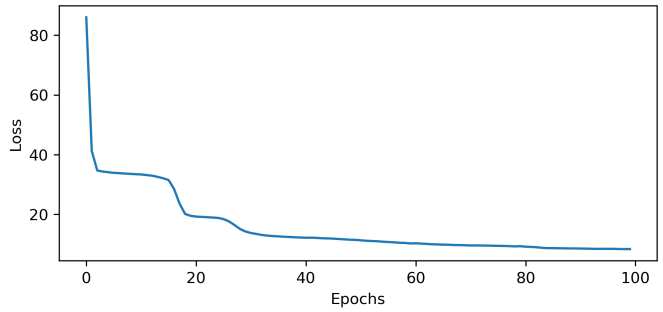


Fig. 7: Training loss for the VAE network described in section III

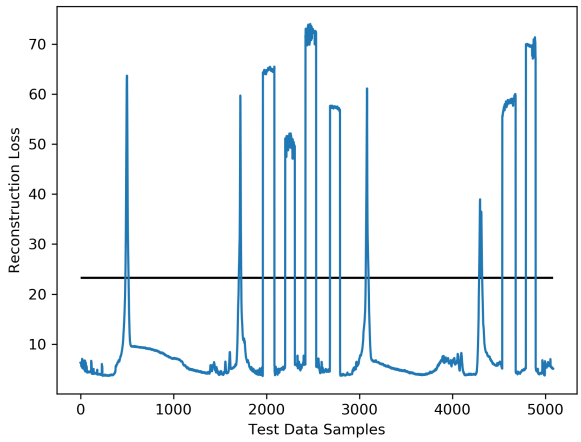


Fig. 9: Reconstruction loss for test data with the threshold line for anomaly detection

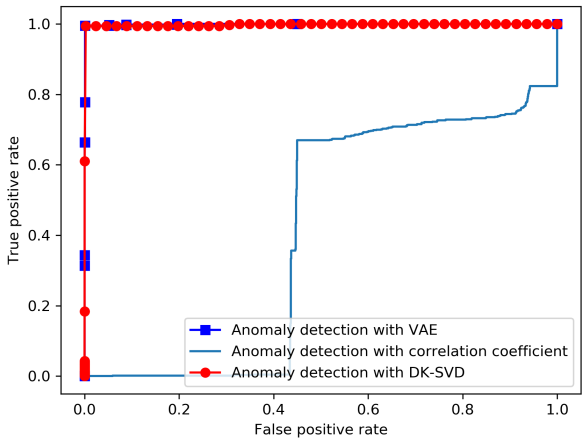


Fig. 10: ROC curve showing the true positive rate and false positive rate for different thresholds using the reconstruction loss from our VAE in comparison with correlation coefficients from Fig. 2 and DK-SVD from [16].

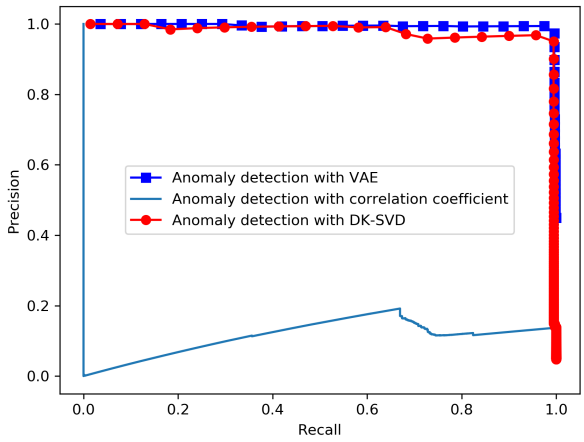


Fig. 11: Precision-recall curve corresponding to different thresholds using the reconstruction loss from our VAE in comparison with correlation coefficients from Fig. 2 and DK-SVD from [16].

97.42% anomaly detection accuracy with a threshold of mean plus 2 standard deviation (17.54) and mean plus 3 standard deviation (23.28), respectively.

We compare our results with the two algorithms mentioned in the introduction - DK-SVD and correlation coefficient. DK-SVD relies on building a dictionary from the training data and representing each test data as a sparse linear combination of the dictionary elements. We train the DK-SVD algorithm using the same data and hyperparameters used in [16] - a dictionary of 110 columns (or atoms) with a sparsity of 4 was trained using a regularization parameter $\alpha = 4$ for 100 iterations. This means an SSTDR data is represented as a linear combination of 4 columns within the dictionary. We then test the algorithm on our test data. Similarly, for the correlation coefficient, we used the test data described in section II and as shown in Fig. 2.

We ran the DK-SVD algorithm on our test data and then compared the confusion matrix with the confusion matrix obtained from VAE. The result shows an anomaly detection accuracy of 99% for both VAE and DK-SVD. Fig. 10 shows the ROC curve for the anomaly detection when reconstruction loss from our VAE was used to detect anomalies compared to when correlation coefficients from Fig. 2 and DK-SVD from [16] are used to detect anomalies. The area under the curve (AUC) is 0.997 for both VAE and DK-SVD. Observe that the true positive rate is very high even when the false positive rate is very low for the VAE and DK-SVD approach. When correlation coefficient is used to detect anomalies, the ROC shows very low true positive rates and does worse than a naive classifier. The area under the curve for this case is 0.435. An advantage of our work is that it is unsupervised while DK-SVD is supervised and requires prior training data of faults which may be difficult to obtain in many situations. Also, there is a possibility of unrepresentative dictionary elements when there is more non-faulty data but VAE relies on only non-faulty data and is not affected by data imbalance problem.

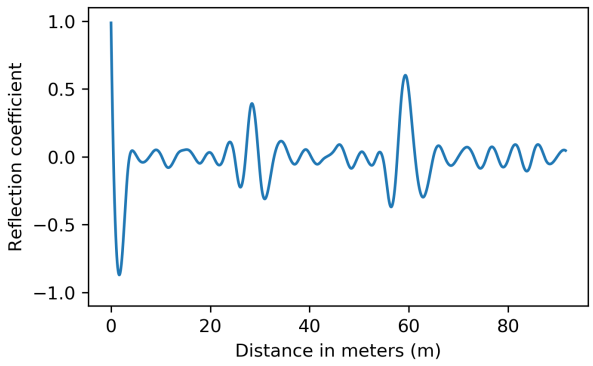
Fig. 11 shows the precision-recall curve for the anomaly detection when reconstruction loss from our VAE is used to detect anomalies compared to when correlation coefficient from Fig. 2 and DK-SVD was used to detect anomalies. The AUC for VAE, DK-SVD, and correlation coefficient are 0.990, 0.978, and 0.040 respectively. While VAE and DK-SVD had the same AUC for the ROC curve, we see that the AUC under the precision recall curve is better with VAE. This shows that the proportion of positive identifications that was actually correct relative to the proportion of actual positives that was identified correctly is very high.

B. Fault Location

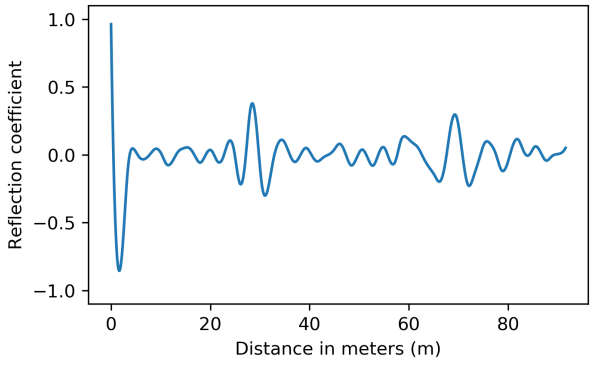
Baseline subtraction is a viable technique for locating the distance to the fault, because it removes the normal reflections in the system, not related to the fault. This includes reflections from connectors and panels. However, the proper baseline must be chosen for accurate localization of faults. Prior work [28] stores a fixed baseline for comparison with new reflection data. A challenge with a fixed baseline is that it does not reflect the current state of the system, as it might have been taken days or weeks before the fault occurred.

To circumvent the issues with a fixed baseline, our autoencoder was used to generate a baseline. In this work, the reconstructed signal from the decoder is used as the baseline for each anomaly. Fig. 12a, 13a, and 14a shows an example anomaly that corresponds to a disconnect at A-, B-, and C-, their respective reconstructed signal (used as baseline) in Fig. 12b, 13b, and 14b, and the resulting baseline subtraction using the reconstructed signal is shown in Fig. 12c, 13c, and 14c.

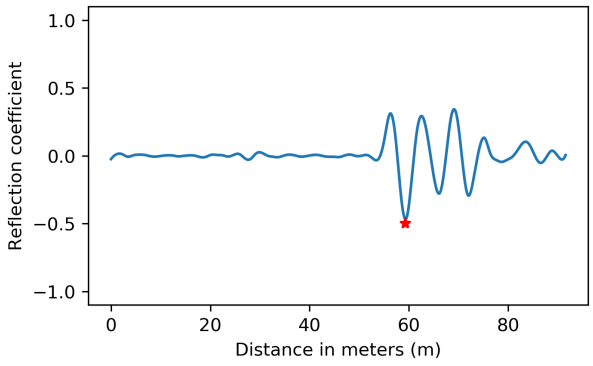
To detect the location of the fault, we inspect the peak of the baseline subtracted signal. The red star in Fig. 12c, 13c, and 14c shows the peak of the baseline subtraction which corresponds to the location of each fault. Table I shows the



(a)

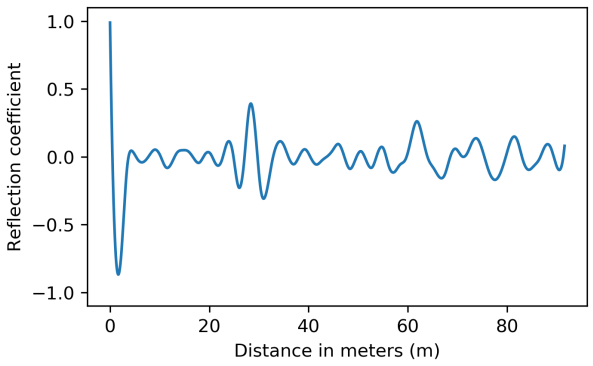


(b)

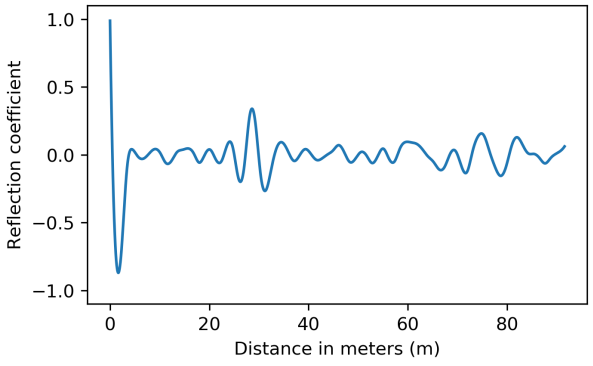


(c)

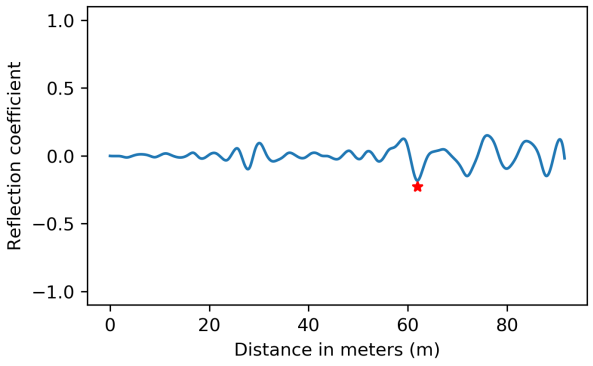
Fig. 12: Comparison of the (a) input data, (b) reconstructed data used as baseline, and (c) the baseline subtraction for disconnect at A+



(a)



(b)



(c)

Fig. 13: Comparison of the (a) input data, (b) reconstructed data used as baseline, and (c) the baseline subtraction for disconnect at B+

result of locating the faults. We compare our results with those obtained in [16] and [28]. The work in [28] located a fault by inspecting the point at which signal deviates from the zero line with a value that exceeds a set threshold while the work in [16] uses a dictionary learning method and they locate the fault by inspecting the peak of the second most prominent dictionary element.

We compare our results to the case where a fixed baseline (column 5) is used for baseline subtraction, and the closest baseline (column 6) is used for baseline subtraction (i.e, choosing a baseline by searching the database of baseline for the closest baseline in correlation). Our results (column 7), as seen in Table I, show that we achieve better results than both a fixed baseline, as in [28], and the closest baseline choice. Also, in comparison to the state-of-the-art work in [16] (column 4), we achieve further improvement. Note that the second column in Table I gives the location of the faults without including the effective electrical length of the panels while the third column accounts for the effective panel length. All results were compared to the values in the third column of Table I.

V. Conclusion and Future Work

In this work, we have shown the viability of combining the ability of spread spectrum time domain reflectometry to probe energized electrical systems with the representation ability of variational autoencoders to detect anomalies (or faults) in a

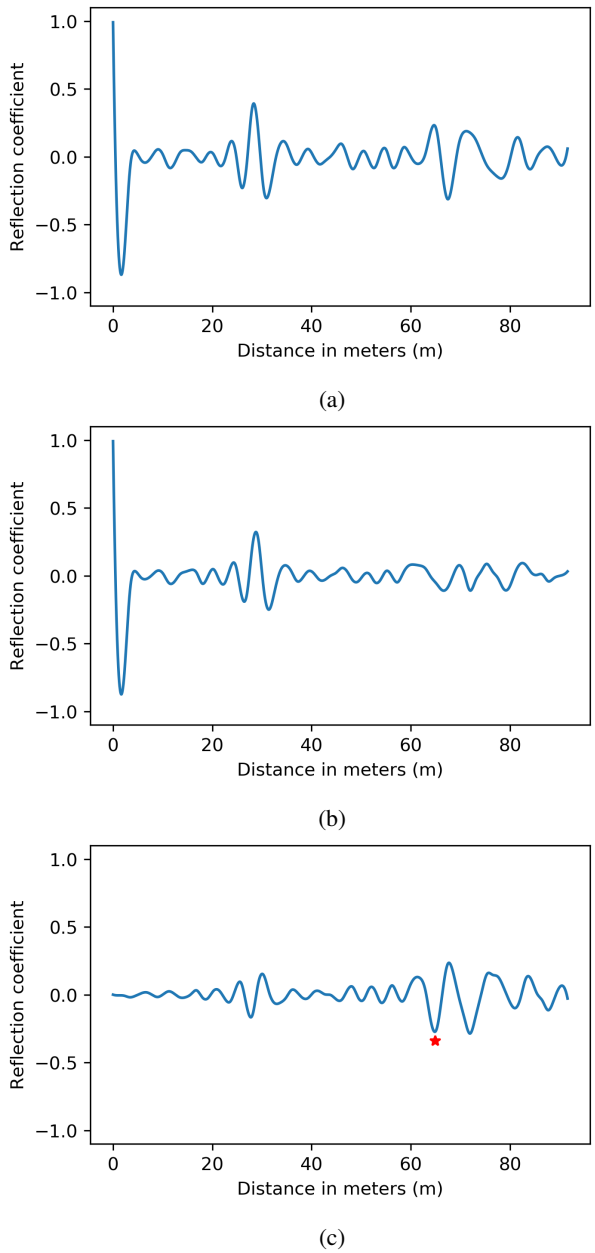


Fig. 14: Comparison of the (a) input data, (b) reconstructed data used as baseline, and (c) the baseline subtraction for disconnect at C+

photovoltaic setup. The ROC and precision-recall curve show a very high true positive rate with an AUC of 0.99. Also, rather than using a fixed baseline, the variational autoencoder can generate a baseline in an unsupervised manner for locating the faults with baseline subtraction. Our results show a maximum percentage absolute relative error of 0.40% on a 5-panel setup with a 59.13m leader cable.An advantage of this method is that it can be used for continual testing of energized systems. As we get more and more baseline measurements, the data fills up the low-dimensional space, and the VAE can better predict the best baseline. Hence, we do not suffer from the class imbalance problem. In a future work, different kinds of faults will be considered to test the robustness of this approach.

TABLE I: Results of the localization of disconnects

Disconnection Localization						
Label	Disconnect Location (m) (cables only)	Corrected Discon-nect Location (m)	Absolute relative error (%) [16]	Absolute relative error (%) [28] (fixed base-line)	Absolute relative error (%) (closest base-line)	Absolute relative error (%) (Our method)
A-	59.13	59.13	0.34	0.27	17.10	0.27
A+	59.13	59.13	0.34	4.13	17.10	0.27
B-	60.05	61.87	0.24	43.33	15.03	0.11
B+	60.05	61.87	0.24	4.51	15.03	0.11
C-	60.96	64.62	0.68	30.15	10.18	0.40
C+	60.96	64.62	0.68	42.09	10.18	0.40

VI. Disclosure

Dr. C.M. Furse is a co-founder of Livewire Innovation, Inc. which is commercializing SSTDR technology, and therefore has a financial conflict of interest with this company

References

[1] Y. T. Quek, W. L. Woo, and T. Logenthiran, "Anomaly warning and fault detection in dc pico-grid with enhanced k-nearest neighbours technique," in *2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, 2018, pp. 728–733.

[2] P. Li and O. Niggemann, "Improving clustering based anomaly detection with concave hull: An application in fault diagnosis of wind turbines," in *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)*, 2016, pp. 463–466.

[3] X. Wang, P. Cui, Y. Du, and Y. Yang, "Variational autoencoder based fault detection and location method for power distribution network," in *2020 8th International Conference on Condition Monitoring and Diagnosis (CMD)*, 2020, pp. 282–285.

[4] Z. Wenqi, W. Li, and C. Wei, "Theoretical and experimental study of spread spectral domain reflectometry," in *2012 Electrical Systems for Aircraft, Railway and Ship Propulsion*, 2012, pp. 1–5.

[5] "Live cable fault detection by livewire innovation," https://www.livewireinnovation.com/, accessed: 2021-07-15.

[6] A. S. Edun, N. K. Tumkur Jayakumar, S. R. Kingston, C. M. Furse, M. A. Scarpulla, and J. B. Harley, "Spread spectrum time domain reflectometry with lumped elements on asymmetric transmission lines," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 921–929, 2021.

[7] S. Zhou, M. Mao, L. Zhou, Y. Wan, and X. Xi, "A shadow fault diagnosis method based on the quantitative analysis of photovoltaic output prediction error," *IEEE Journal of Photovoltaics*, vol. 10, no. 4, pp. 1158–1165, 2020.

[8] Y. Hu, J. Zhang, W. Cao, J. Wu, G. Y. Tian, S. J. Finney, and J. L. Kirtley, "Online two-section pv array fault diagnosis with optimized voltage sensor locations," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 11, pp. 7237–7246, 2015.

[9] H. Momeni, N. Sadoogi, M. Farrokhifar, and H. F. Gharibeh, "Fault diagnosis in photovoltaic arrays using gbssl method and proposing a fault correction system," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5300–5308, 2020.

[10] P. Xi, P. Lin, Y. Lin, H. Zhou, S. Cheng, Z. Chen, and L. Wu, "Online fault diagnosis for photovoltaic arrays based on fisher discrimination dictionary learning for sparse representation," *IEEE Access*, vol. 9, pp. 30 180–30 192, 2021.

[11] A. S. Edun, S. Kingston, C. LaFlamme, E. Benoit, M. A. Scarpulla, C. M. Furse, and J. B. Harley, "Detection and localization of disconnections in a large-scale string of photovoltaics using sstdr," *IEEE Journal of Photovoltaics*, pp. 1–8, 2021.

[12] M. U. Saleh, C. Deline, E. Benoit, S. Kingston, A. S. Edun, N. K. T. Jayakumar, J. B. Harley, C. Furse, and M. Scarpulla, "An overview of spread spectrum time domain reflectometry responses to photovoltaic faults," *IEEE Journal of Photovoltaics*, vol. 10, no. 3, pp. 844–851, 2020.

[13] A. Hanif and F. Khan, "Degradation detection of thermally aged sic and si power mosfets using spread spectrum time domain reflectometry (sstdr)," in *2018 IEEE 6th Workshop on Wide Bandgap Power Devices and Applications (WiPDA)*, 2018, pp. 18–23.
[14] S. Roy and F. Khan, "Detection of degraded/aged cell in a li-ion battery pack using spread spectrum time domain reflectometry (sstdr)," in *2020 IEEE Applied Power Electronics Conference and Exposition (APEC)*, 2020, pp. 1483–1488.
[15] C. LaFlamme, E. Benoit, A. Edun, C. M. Furse, P. K. Kuhn, M. A. Scarpulla, and J. B. Harley, "Quantifying the environmental sensitivity of sstdr signals for monitoring pv strings," *IEEE Journal of Photovoltaics*, pp. 1–7, 2021.
[16] A. S. Edun, C. LaFlamme, S. R. Kingston, H. V. Tetali, E. J. Benoit, M. Scarpulla, C. M. Furse, and J. B. Harley, "Finding faults in pv systems: Supervised and unsupervised dictionary learning with sstdr," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 4855–4865, 2021.
[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.
[18] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 371–375.
[19] T. Song, J. Sun, B. Chen, W. Peng, and J. Song, "Latent space expanded variational autoencoder for sentence generation," *IEEE Access*, vol. 7, pp. 144 618–144 627, 2019.
[20] T. Inoue, S. Choudhury, G. De Magistris, and S. Dasgupta, "Transfer learning from synthetic to real images using variational autoencoders for precise position detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2725–2729.
[21] Z. Liu, F. Oviedo, E. M. Sachs, and T. Buonassisi, "Detecting microcracks in photovoltaics silicon wafers using varitional autoencoder," in *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, 2020, pp. 0139–0142.
[22] A. Dairi, F. Harrou, Y. Sun, and S. Khadraoui, "Short-term forecasting of photovoltaic solar power production using variational auto-encoder driven deep learning approach," *Applied Sciences*, vol. 10, no. 23, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/23/8400
[23] S. Kullback and R. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.
[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.
[25] M. Kafal, F. Mustapha, W. Ben Hassen, and J. Benoit, "A non destructive reflectometry based method for the location and characterization of incipient faults in complex unknown wire networks," in *2018 IEEE AUTOTESTCON*, 2018, pp. 1–8.
[26] Q. Shi and O. Kanoun, "A new algorithm for wire fault location using time-domain reflectometry," *IEEE Sensors Journal*, vol. 14, no. 4, pp. 1171–1178, 2014.
[27] L. Griffiths, R. Parakh, C. Furse, and B. Baker, "The invisible fray: a critical analysis of the use of reflectometry for fray location," *IEEE Sensors Journal*, vol. 6, no. 3, pp. 697–706, 2006.
[28] M. U. Saleh, C. Deline, S. Kingston, N. K. T. Jayakumar, E. Benoit, J. B. Harley, C. Furse, and M. Scarpulla, "Detection and localization of disconnections in PV strings using spread-spectrum time-domain reflectometry," *IEEE Journal of Photovoltaics*, vol. 10, no. 1, pp. 236–242, 2020.

Ayobami S. Edun received his B.Eng. degree in Electrical and Electronics engineering from Federal University of Technology, Akure, Nigeria in 2014. He received the M.S. degree in Electrical and Computer engineering from University of Florida, Gainesville, FL, USA in 2019. He is currently working towards his PhD degree in Electrical and Computer Engineering at the University of Florida, Gainesville, FL, USA. He currently works as a Research Assistant at the SmartDATA Lab, University of Florida, Gainesville, FL where he focuses on developing algorithms to detect, localize, and characterize faults within solar arrays. His research interests include data science, machine learning, renewable energy, and smart grids. Ayobami's awards include 2020 College of Engineering outstanding achievement award and 2020 Scarborough-Maud Fraser award for commitment and dedication to academics and service within the University of Florida community.

Samuel R. Kingston was born in Salt Lake City, Utah, UT, USA in 1991. He received the A.S. degree in business from Salt Lake Community College, Salt Lake City, in 2011. He received a B.S. degree in electrical computer engineering from the University of Utah, Salt Lake City, UT, in 2016 and is currently working on a Ph. D from the University of Utah, Salt Lake City, UT. From 2017 to current, he has been a Research Assistant with the University of Utah lab working in the algorithms group. He has been working with spread spectrum time domain reflectometry (SSTDR) in being able to detect, localize, and characterize faults within solar panel system. To date, he has written a conference paper for nondestructive health monitoring. He is currently working on several journal papers, where each one will be a building block in achieving the overall research team goal. His research interests are in signal processing used for health monitoring, renewable energy alternatives, and creating successful startups from conceptual ideas. In 2016, his senior project team received an award for best clinic project where he worked with L3 communications to develop a way to detect low probability of intercept (LPI) RADAR signals.

Cody LaFlamme received his B.S. degree in Computer Engineering from the University of Florida in 2019. Currently, he is a Ph.D. student of Electrical and Computer Engineering and researcher at the University of Florida, focusing on signal processing and machine learning for signal segmentation and analysis.

Michael A. Scarpulla (M'05-SM'14) earned the B.Sc. degree from Brown University in 2000 and the PhD from UC Berkeley in 2006, both in Materials Science and Engineering. His PhD work focused on laser processing of ion implanted compound semiconductors, carrier mediated ferromagnetism, and multiband semiconductors. From 2006-2008 he was a postdoctoral scholar at UC Santa Barbara working on epitaxial integration of rare-earth pnictides with III-V semiconductors using MBE. Since joining the ECE and MSE faculties at University of Utah in 2008, he has worked in light trapping for photovoltaics, materials processing and characterization of chalcogenide thin film photovoltaics, reflectometry in photovoltaic systems, and defects in wide-bandgap semiconductors. His hobbies include skiing, climbing, and other mountain adventures.

Cynthia M. Furse (M'85–SM'99–F'08) is Professor of Electrical and Computer Engineering at the University of Utah. Dr. Furse received her B.S. in electrical engineering with a mathematics minor in 1985, M.S. degree in electrical engineering in 1988, and her Ph.D. in electrical engineering from the University of Utah in 1994. She has applied her expertise in electromagnetics to sensing and communication in complex lossy scattering media such as the human body, geophysical prospecting, ionospheric plasma, and aircraft wiring networks. She has taught electromagnetics, wireless communication, computational electromagnetics, microwave engineering, antenna design, and introductory electrical engineering and has been a leader in the development of the flipped classroom. Dr. Furse is a Fellow of the IEEE and the National Academy of Inventors. She is a past AdCom member for the IEEE AP society and past chair of the IEEE AP Education Committee. She has received numerous teaching and research awards including the 2009 IEEE Harriett B. Rigas Medal for Excellence in Teaching. She is a founder of LiveWire Innovation, Inc., a spin-off company commercializing devices to locate intermittent faults on live wires.

Joel B. Harley (S'05-M'14) received his B.S. degree in Electrical Engineering from Tufts University in Medford, MA, USA. He received his M.S. and Ph.D. degrees in Electrical and Computer Engineering from Carnegie Mellon University in Pittsburgh, PA, USA in 2011 and 2014, respectively. In 2018, he joined the University of Florida, where he is currently an associate professor in the Department of Electrical and Computer Engineering. Previously, he was an assistant professor in the Department of Electrical and Computer Engineering at the University of Utah. His research interests include integrating novel signal processing, machine learning, and data science methods for the analysis of waves and time-series data. Dr. Harley's awards and honors include the 2021 Achenbach Medal from the International Workshop on Structural Health Monitoring, a 2020 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Society Star Ambassador Award, a 2020 and 2018 Air Force Summer Faculty Fellowship, a 2017 Air Force Young Investigator Award, a 2014 Carnegie Mellon A. G. Jordan Award (for academic excellence and exceptional service to the community). He has published more than 90 technical journal and conference papers, including four best student papers. He is a student representative advisor for the IEEE Ultrasonics, Ferroelectrics, and Frequency Control Society, a member of the IEEE Signal Processing Society, and a member of the Acoustical Society of America.