# Feature Analysis for PV Fault Detection Neural Network Using Linear PCA and Random Forest


Skyler Verch
*School of ECEE*
*Arizona State University*
sverch@asu.edu

Gowtham Muniraju
*School of ECEE*
*Arizona State University*
gmuniraj@asu.edu

Andreas Spanias
*School of ECEE*
*Arizona State University*
spanias@asu.edu

Yiannis Tofis
*KIOS center*
*University of Cyprus*
tofis.n.yiannis@ucy.ac.cy



***Abstract*—Photo Voltaic (PV) smart monitoring devices (SMD's) provide 10 features of data such as current, voltage, irradiance, etc. These features have been used by Neural networks to detect and classify faults in a solar array with 90% accuracy. The purpose of this work will be to identify which of the 10 features contributes most significantly to the accuracy of the fault detection and classification neural network. Using linear principal component analysis as a dimensionality reduction technique, and a random forest model to determine feature importance, we show that the number of features can be reduced while retaining high classification accuracy.**



***Index Terms*—linear principal component analysis, feature analysis, neural networks, machine Learning, PV modules**


## I. Introduction

Utility-scale solar farms greatly benefit from sensor monitoring systems with the capacity to automatically and remotely detect array faults and anomalies. A system with such capabilities not only reduces monitoring and maintenance cost but also elevates the efficiency and robustness of a PV power plant, as the sensor data provided by such a system can also be used to optimize power output via topology reconfiguration. [1]

## II. Machine learning and neural network

Previous work in SenSIP lab addressed several problems in solar array monitoring, control and optimization [1-11]. Initial work was reported in [2] where traditional statistical methods were proposed. Later machine learning methods [3] were considered including PU Learning [4]. Fault detection using neural nets was reported in [5,6] and optimization methods were reported in [7]. PU learning for fault detection was reported in [8] and a recent study including neural net fault detection experiments and simulations on a quantum computer simulator was published in [9].

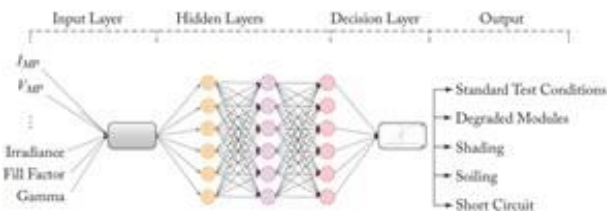


Fig. 1. Fault classification using a neural network [1]

A visualization of the unsupervised neural network architecture used in this project is shown in figure 1. A set of data (one measurement corresponding to each feature) is passed through the neural network to be classified as clean "no fault", or one of 4 faults. The shown architecture has produced 90% classification accuracy [1]. In order to optimize such a network, however, it may be useful to identify and reduce redundancies in the data set while maintaining high degrees of classification accuracy. This simplification could reduce computational resources, prevent over fitting and improve classification accuracy.

## III. Experimental methods

Several techniques are commonly used to identify the data features which contribute most significantly to a network's classification ability [12]. This work will focus specifically on linear principal component analysis and random forest as dimensionality reduction techniques. Principal Component Analysis (PCA) involves finding the eigenvectors of the data set's covariance or correlation matrix. This technique effectively reduces the number of dimensions in a data set, by creating new variables (principal components), as linear functions of original variables [13]. Figure 2 shows the directions, or rather, the principal components (PC1 and PC2) that capture the largest variance in the data. In this case, projecting the data points to the singular dimension of PC1 retains a wide range of data with minimal overlap. Nonlinear principal component analysis may be used in conjunction to capture trends in data that may not be accurately described via a linear function.

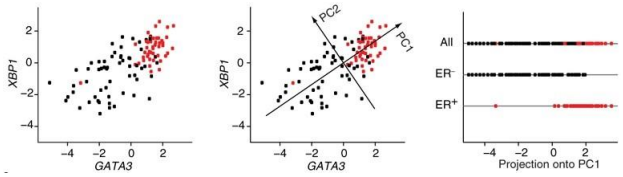


Fig. 2. PCA of a gene expression data set [14]

Since the principal component method reduces the feature space by combining the original features into more significant ones, the original feature space is lost. Thus, the importance of the original features cannot be derived with this method. We will instead use random forest analysis to determine these importances. Random forest is an important machine learn- ing classification algorithm, which uses bagging and feature randomness to generate many uncorrelated decision trees that independently and then aggregately predict a classification. By creating decision trees that split the data with a random selection of features, and by training these trees on different sets of data via bagging, the final aggregate classification is

more comprehensive than the prediction of any individual tree. The structure of a random forest algorithm is visualized in figure 3. Using this algorithm, it's easy to determine how well any feature splits the data by its Gini impurity. Gini impurity is simply the probability of incorrectly classifying a data point [15]. The primary role of the random forest algorithm in this paper is to provide these feature importances. We then use a neural network model to affirm that the random forest algorithm is correctly determining which features are most important for classification accuracy.

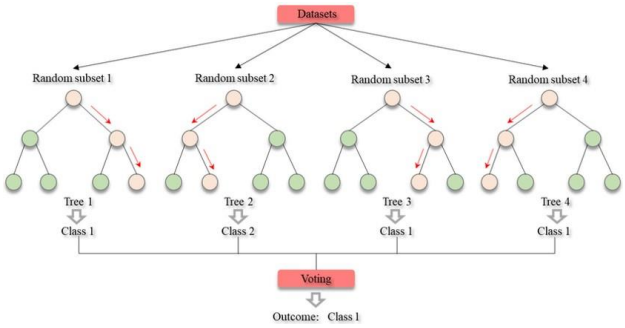


Fig. 3. Random forest structure [16]

## IV. Dataset

Both Machine learning techniques were implemented in python with the use of machine learning and data science toolkits such as scikit learn, keras, pandas, and numpy. We used the PVWatts dataset from the National Renewable Energy Laboratory (NREL) [17]. It contains 21,486 unique data samples for each feature, of which 4297 samples each belong to one of 5 classifications. There are 10 input features of data provided including $V_{OC}$, $V_{MP}$ , $I_{SC}$, $I_{MP}$ , $P_{MP}$ , temperature, irradiance, Fill Factor, gamma and DC power. Gamma and fill factor are dependent features in that they're mathematical combinations of other features. Gamma is the ratio of power over irradiance and fill factor is the ratio of the maximum power from the solar cell over the product of $V_{OC}$ and $I_{SC}$. The 5 classifications as shown in figure 1 include standard test conditions, degraded modules, shading, soiling, and short circuit. Of this data, 70% was used to train the neural network,and the remaining for validation.

## V. Results

We tested the classification accuracy of the neural network with an increasing number of principal components as features. The results of which are shown in figure 4. It's clear that to reach the target 90% classification accuracy and to explain for near 100% of the data variance, the network needs 8 principle component inputs. Additional principal components did not improve the classification accuracy. It's also evident that there are diminishing returns in accuracy beyond 4 principal components, and that it may be beneficial to use 4 components at approximately a 5% loss in peak accuracy.

We then generated feature importances from the random forest algorithm. Figure 5 shows the features in order of

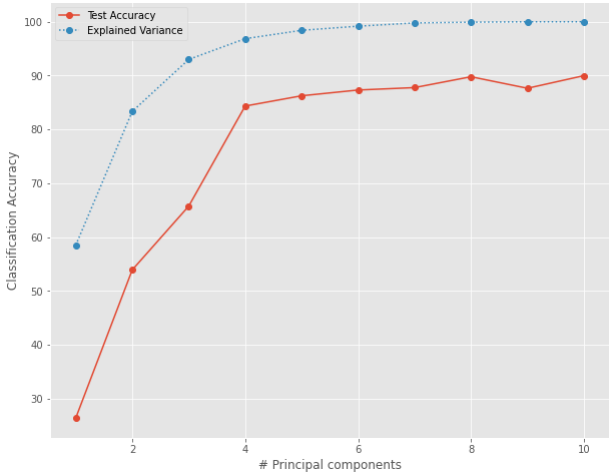


Fig. 4. Linear PCA classification accuracy.

importance. The top 3 features are approximately twice as important as the 4th and 5th most important features. Using these rankings, we tested the neural network's classification ability using an increasing number of features ordered from most to least important. Figure 6 shows that the model reaches target accuracy with only the 4 most important features. Similarly, to PCA, the top 4 features are responsible for most of the accuracy improvement.

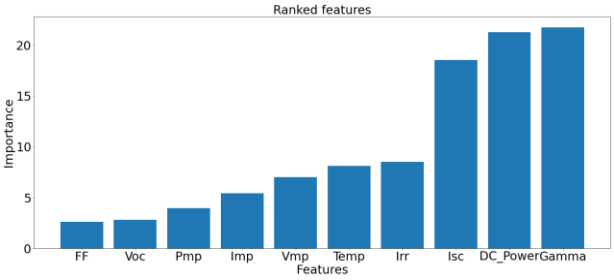


Fig. 5. Random forest feature importances

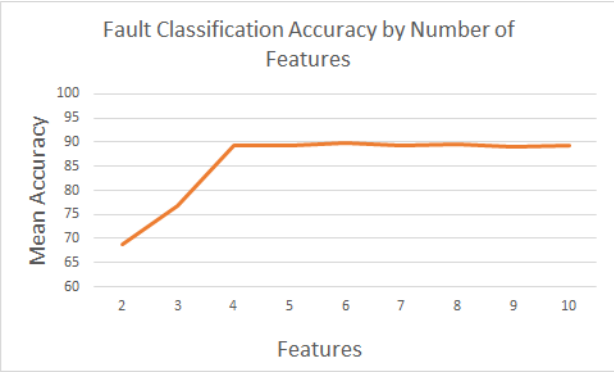


Fig. 6. Classification accuracy from best to worst features

In our analysis of feature importance, we found gamma to be the most important feature. Gamma is the ratio of two other highly important features: power and irradiance. It might make intuitive sense that gamma is the most important because it carries the significance of two features. An effort was made to mathematically combine other features in a similar fashion, to determine whether this intuition was generally true. While other dependent features did often carry more importance than their constituent features, the presence of outliers, inconsistencies, and a general lack of ability to systematically consider all possible combinations of features prevents the assertion that dependent features indeed carry the importance of multiple features. One such inconsistency is fill factor, as it is the least important feature, despite being a ratio of three other features. Another finding, however, shows that dependent features do play a unique role in the neural network's classification ability. In figures 7 and 8, we show the network's performance without dependent features. It's clear that the network loses 5% accuracy. This finding is concurred in figures 9 and 10, as we reintroduce gamma and regain the lost accuracy. It can be concluded then, that the information contained in certain dependent features is essential to the neural network.

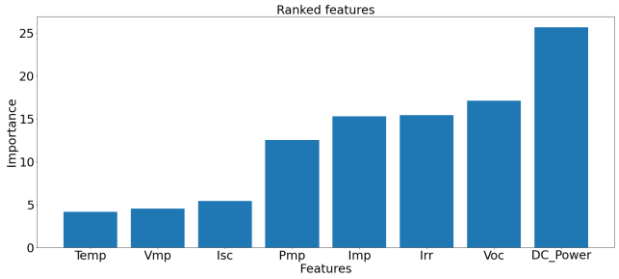


Fig. 7. Feature importance without gamma and fill factor

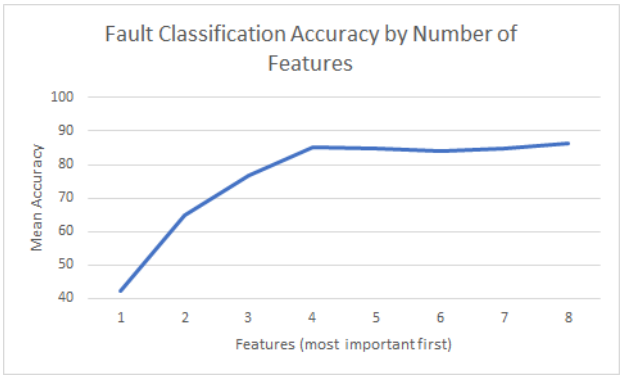


Fig. 8. Classification accuracy without gamma and fill factor

## VI. Conclusion

The feature space can be effectively reduced and optimized with both Linear Principal Component Analysis and Random Forest importance techniques. In either case, the top 4 features allowed the network to reach at least 85% classification accuracy. It cannot be concluded that dependent features are consistently more important, but they do hold a vital role in training the neural network to accurately classify faults.

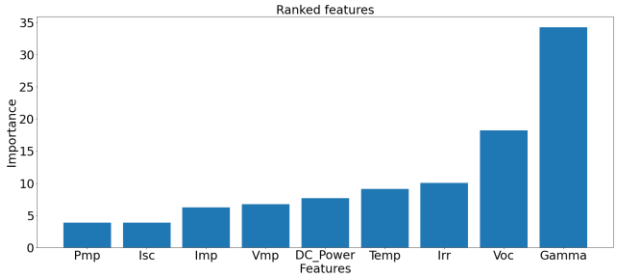


Fig. 9. Feature importance with gamma

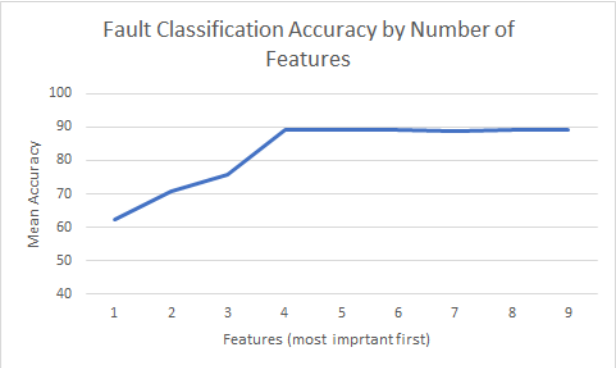


Fig. 10. Classification accuracy with gamma

## Acknowledgment

This research is funded by NSF grant 1854273.

## References

[1] S.Rao, S. Katoch, V. Narayanaswamy, G. Muniraju, C. Tepedelenlioglu, A. Spanias, P. Turaga, R. Ayyanar, and D. Srinivasan, "Machine learning for solar array monitoring, optimization, and control," *Synthesis Lectures on Power Electronics*, vol. 7, no. 1, pp. 1–91, 2020.

[2] H. Braun, S. Buddha, V. Krishnan, C. Tepedelenlioglu, A. Spanias, S.i Takada, T. Takehara, M. Banavar, and T. Yeider., Signal Processing for Solar Array Monitoring, Fault Detection, and Optimization, Synthesis Lectures on Power Electronics, Morgan & Claypool, Book, 1-111 pages, ISBN 978-1608459483, Sep. 2012.

[3] U. Shanthamallu, A. Spanias, C. Tepedelenlioglu, M. Stanley, "A Brief Survey of Machine Learning Methods and their Sensor and IoT Applications," *Proceedings 8th International Conference on Information, Intelligence, Systems and Applications (IEEE IISA 2017)*, Larnaca, August 2017

[4] K. Jaskie and A. Spanias, "Positive and Unlabeled Learning Algorithms and Applications: A Survey," *Proc. IEEE IISA 2019*, Patras, July 2019

[5] Sunil Rao, Andreas Spanias, Cihan Tepedelenliglu, "Solar Array Fault Detection using Neural Networks", IEEE International Conference on Industrial Cyber-Physical Systems (ICPS), Taipei, May 2019..

[6] S. Rao, G. Muniraju, C. Tepedelenlioglu, D. Srinivasan, G. Tamizhmani and A. Spanias, "Dropout and Pruned Neural Networks for Fault Classification in Photovoltaic Arrays, *IEEE Access*, 2021.

[7] Vivek Narayanaswamy, Raja Ayyanar, Andreas Spanias, Cihan Tepedelenlioglu, "Connection Topology Optimization in PV Arrays using Neural Networks'," IEEE International Conference on Industrial Cyber-Physical Systems (ICPS), Taipei, May 2019.

[8] K. Jaskie, J. Martin, and A. Spanias, "PV Fault Detection using Positive Unlabeled Learning," Applied Sciences, vol. 11, Jun. 2021.

[9] Glen Uehara, Sunil Rao, Mathew Dobson, Cihan Tepedelenlioglu and Andreas Spanias, "Quantum Neural Network Parameter Estimation for Photovoltaic Fault," *Proc. IEEE IISA 2021*, July 2021

[10] H. Braun, S. T. Buddha, V. Krishnan, C. Tepedelenlioglu, A. Spanias, M. Banavar, and D. Srinivansan, "Topology reconfiguration for optimization of photovoltaic array output," *Elsevier Sustainable Energy, Grids and Networks* (*SEGAN*), pp. 58-69, Vol. 6, June 2016.

[11] M20-254P Dropout and Pruned Neural Networks for Fault Classification in Photovoltaic Arrays Gowtham Muniraju, Sunil Rao, Andreas Spanias, Cihan Tepedelenlioglu, **Provisional US 63/039,012** , 06/15/2020

[12] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in 2014 *Science and*

*Information Conference*, 2014, pp. 372–378.
[13] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374 no. 2065,p. 20150202, 2016.
[14] M. Ringnér, "What is principal component analysis?" *Nature biotechnology*, vol. 26, no. 3, pp. 303–304, 2008.
[15] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.
[16] J. Yang, G. Qin, P. Tang, Shen, T.-Y. Liu, and S. Gao, "Delineation of urban growth boundaries using a patch-based cellular automata model under multiple spatial and socio-economic scenarios," *Sustainability*, vol. 178, p. 6159, 2019.
[17] A. Dobos, "Pvwatts version 1 technical reference," 2013.