18th International Conference on Renewable Energies and Power Quality (ICREPQ'20)
Granada (Spain), 1st to 2nd April 2020
Renewable Energy and Power Quality Journal (RE&PQJ)
ISSN 2172-038 X, Volume No.18, June 2020

Analysis of public datasets of power quality distortions

S. Dominguez-Gimeno[1], R. Igual[1] and C. Medrano[1]

[1] Department of Electrical Engineering / Electronics Engineering and Communications
E.U.P.T., Universidad de Zaragoza
Campus of Teruel, 44003 Teruel (Spain)
Phone/Fax number:+0034 978645359, e-mail: rigual@unizar.es, ctmedra@unizar.es

Abstract. Automatic classification of power quality distortions has gained interest in research due to the proliferation of distributed power systems with renewable sources. To train and test a classification system, data with power quality distortions are required. Most studies generate synthetic data from mathematical equations, since real distortions are difficult to record. A possible alternative is to use public datasets of real disturbances. However, there are strong differences among public datasets. In this paper, existing datasets of power quality distortions were compiled and their main features were analysed and compared. To the best of our knowledge, this is the first work reviewing these datasets. To identify the datasets, the most cited papers on this topic were surveyed. In addition, systematic searches were conducted in four popular scientific repositories. As a result, four available datasets were identified. They included a limited number of samples (20-44) and types of distortions. Sampling frequencies and recording conditions were appropriate and the two main fundamental grid frequencies (50 and 60 Hz) were also considered. Although these datasets are appropriate for partially testing automatic classifiers, a remaining research effort is to provide comprehensive datasets with hundreds of samples and several types of distortions.

Key words. Public dataset, analysis, power quality, disturbances, critical comparison.

1. Introduction

Power quality is a topic that has been widely studied [1]. In recent years, a growing interest has been identified due to the popularization of renewable power systems. Renewable sources, such as solar photovoltaics, include non-linear components. These elements are sources of power quality distortions [2]. This problem has been aggravated with the development of distributed technologies, which mainly include renewable sources [3].

Mitigation of power quality distortions is an active topic of research [4]. Studies in this field agree that to properly mitigate power quality distortions it is essential to detect the occurrence of the distortions and to identify their specific types [5]. Many automatic classifiers of power quality distortions have already been developed [1]. To provide a measure of their performance, datasets with disturbance samples are required [6]. As the recording of real power quality distortions is a difficult task, most studies in this field use synthetic signals generated from mathematical models. Deokar & Waghmare [7] presented a mathematical model to generate five disturbances: sags, swells, harmonics, fluctuations and transients (low and high frequency). Eight distortions were modelled by Decanini et al. [8], Naderian & Salemnia [9], Abdoos et al. [10], Borges et al. [11] and Huang et al. [12]. They also included interruptions, harmonics with sag and harmonics with swell. Flicker, notching and spikes were considered in the equations implemented by Kumar et al. [13] and Granados-Lieberman et al. [14] to model a total of nine single disturbances. Other authors have considered many more types of combined distortions (Hooshmand & Enshaee [15], Kanirajan & Kumar [16], Kubendran & Loganathan [17]). Igual et al. [6] merged several existing proposals into an integral mathematical model that considered most of the equations implemented by other authors. It is publicly available for download by any interested researcher.

However, the use of synthetic signals have several limitations. It is not clear that they accurately represent the real electrical signals of grids. Therefore, classification systems that provide high performance when validated with synthetic signals may not behave equally well when used in real grids. In addition, some distortions, especially those that combine more than one simple disturbance, are difficult to model mathematically. Thus, real distortions are required to validate power quality classifiers.

Several authors have recorded real distortions from electrical facilities [18], [19]. However, it is not easy to have a complete dataset of real power quality disturbances. Even if grids are accessible, the number of registered samples may be not sufficient to train and test automatic classification systems, since distortions occur occasionally.

A feasible alternative is to use public datasets of power quality distortions. Sharing real disturbances publicly is not common in this field. Most studies that use real distortions to test the classifiers do not provide them as supplementary material. However, there are some public datasets. They were published in a variety of platforms and formats. Therefore, it is not easy to compare them fairly to

know the features of each dataset and to select those that best suit the requirements of a particular study.

In this paper, we have compiled and analysed the existing public datasets of power quality distortions. For that, scientific repositories and the most relevant studies on power quality classification have been examined and those that use public datasets have been identified. In addition, the most important analysis aspects of the datasets have been defined. As a result, we present a critical comparison of the datasets. To the best of our knowledge, this is the first study that compiles and compares public datasets of power quality disturbances. This paper aims to serve researchers in power quality classification to identify the most appropriate public datasets to be used in the validation of their classifiers.

The rest of this paper is organized as follows: Section 2 describes the materials and methods used for this study, including the selection procedure and the items of analysis, Section 3 presents the results of the study, Section 4 discusses those results and, finally, Section 5 draws some conclusions from this work.

2. Materials and methods

A. Selection procedure

The public datasets to analyse should meet the following conditions:

- They should include real distortions from any electrical facility.
- They should provide the distortions in time domain (voltage versus time or current versus time) since real signals are in the grid in time domain. These signals are the inputs of automatic classification systems.
- They should be publicly available for download and reuse.
- Links to the datasets should be active.

To find the datasets, two types of searches were conducted. First, the 15 % most cited papers in the field of automatic power quality classification were found and examined (124 studies). Specifically, the origin of the distortions used in the validation experiments was examined. From this set of studies, only 9 used public datasets in the validation experiments. The rest used synthetic or simulated datasets or real private datasets that were not publicly available. The datasets cited in the 9 papers that met the requirements outlined above were selected for this paper. Two different datasets were identified. Both belonged to the IEEE, one to the "IEEE 1159.2" working group [20] and the other to the "IEEE 1159.3" working group [21].

Second, we searched popular scientific repositories of public datasets: "IEEE Data Port", "Mendeley data", "IEEE Power and Energy Society Open Datasets" and "Harvard dataverse". The searches included the following generic keywords: "power quality". These keywords were selected so as not to leave any existing dataset. Four hundred and twenty-two datasets were found in the generic field of power quality. Their titles and brief descriptions were analysed. From them, 10 candidates were selected. They were examined in more detail and only 2 presented signals in the time domain that fulfilled the conditions established above.

In relation to "IEEE Data Port", three datasets were preliminarily selected [22]–[24]. However, one of them did not include electrical signals but features extracted from them [22]. The other two contained real time-domain signals of power quality distortions. They were entitled "Real life power sags" [23] and "Real life power quality transients" [24].

In relation to the "Mendeley data" searches, only two candidate datasets were identified. However, they were discarded since one of them presented a mathematical model of power quality distortions instead of a public dataset and the other did not include electrical disturbances.

Regarding "IEEE Power and Energy Society Open datasets", only one dataset on power quality was found [25]. However, it did not include real data from the grid but "Laboratorial Essays of Polypropylene and All-film Power Capacitors". Therefore, it was not interesting for automatic classification of power quality distortions.

With respect to the "Harvard dataverse", fifty records were found that contained power quality in the name, but only four of them were related to the electrical field. The four datasets belonged to the same category: "Power Quality and Modern Energy for All" [26]. However, they included voltage data without mentioning specific power quality distortions. Therefore, they are not appropriate for studies of power quality classification.

Figure 1 shows a graphical description of the search and selection procedure.

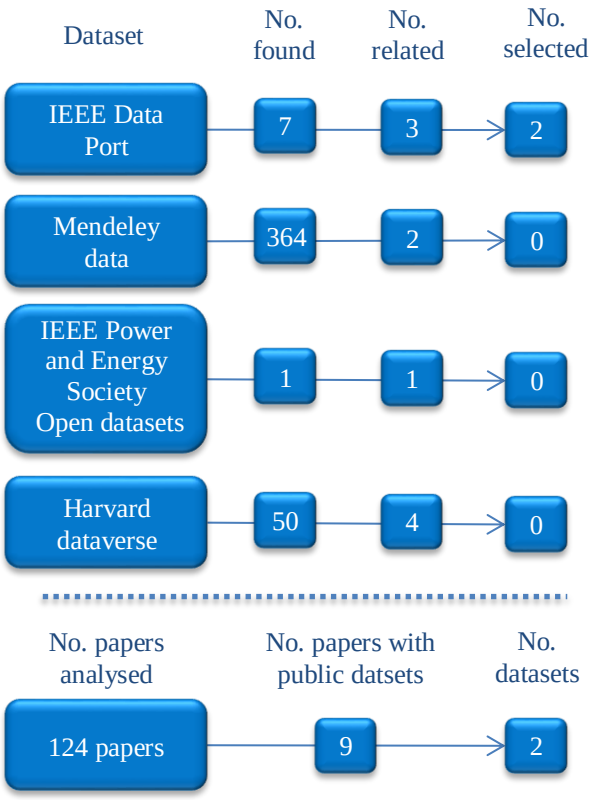

Fig. 1. Results of the search procedure in research studies and in popular scientific repositories of electrical datasets.

B. Items of analysis

Datasets were subjected to a detailed analysis. The following items were obtained for each of them:

- Institution, research group or association that published the public dataset.
- Year of the registration.
- Country in which the real distortions were collected.
- Specific conditions in which the distortions were recorded.
- Number of signals with power quality distortions contained in the dataset.
- Types of distortions registered.
- Sampling frequency of the acquisition setup.
- Fundamental frequency of the grid.
- Number of periods recorded for each distortion.
- Number of points contained in the files for each distortion.
- Data format of the files containing the power quality distortions.
- Acquisition devices and software used to capture the real distortions.
- Whether or not a script was provided to analyse the data.

These items were selected since they cover the main aspects of analysis of datasets for studies on power quality classification [1].

3. Results

The four datasets found were analysed and the items stated in section 2.B were obtained for each of them. The second column of Table I shows the analysis for the "IEEE 1159.2" dataset [20], the third column presents the results for the "IEEE 1159.3" dataset [21], the fourth column analyses "Real life power sags" dataset [23], while the fifth column shows the values for "Real life power quality transients" dataset [24].

As an example, Figures 1 to 4 show four signals with power quality distortions that belong to each dataset.

Table I. – Comparative analysis of the public datasets.

	Dataset			
	IEEE 1159.2 [20]	**IEEE 1159.3** [21]	**Real life power sags** [23]	**Real life power quality transients** [24]
Institution/group of publication	IEEE 1159.2 Working Group	IEEE 1159.3 PDQDIF Taskforce	Dept. of Automation Engineering, Electronics, Architecture and Computer Networks. Polytechnic School of Algeciras, University of Cadiz.	Dept. of Automation Engineering, Electronics, Architecture and Computer Networks. Polytechnic School of Algeciras, University of Cadiz.
Year of registration	1994/1995	1999/2002/2007	2011/2012	2010/2011
Country	-	-	Spain	Spain
Conditions	-	-	According to UNE-EN 61000-4-30	According to UNE-EN 61000-4-30
No. signals	20	37	27	44
Types of distortions	Oscillatory transients, sags, sag with harmonics, swell, swell with harmonics, interruption, among others	Transients, oscillatory transients, sags, among others	Sags	Transients
Sampling frequency	15,370 Hz	7,700 Hz/15,370 Hz	20,000 Hz	20,000 Hz
Fundamental frequency	60 Hz	60 Hz	50 Hz	50 Hz
Periods/distortion	6	4, 5, 6, 7, 9, 10, 11	50, 100	50
No. points/signal	1536	256, 512, 640, 768, 896, 1152, 1536, 2816	20000, 20400, 40400	20000
Data format	.xls	PQDIF	.txt	.txt
Acquisition devices/software	TESTWAV and ORIA-MAC	PQDIFfractor	HAMEG instrument Differential probe HZ 115 + National Instrument Chassis NIcDAQ 9188 + NI 9225 Simultaneous input mode + LabView + general purpose PC to access instruments via Ethernet	HAMEG instrument Differential probe HZ 115 + National Instrument Chassis NIcDAQ 9188 + NI 9225 Simultaneous input mode + LabView + general purpose PC to access instruments via Ethernet
Script provided (Yes/No)	No	Yes (XML)	No	No

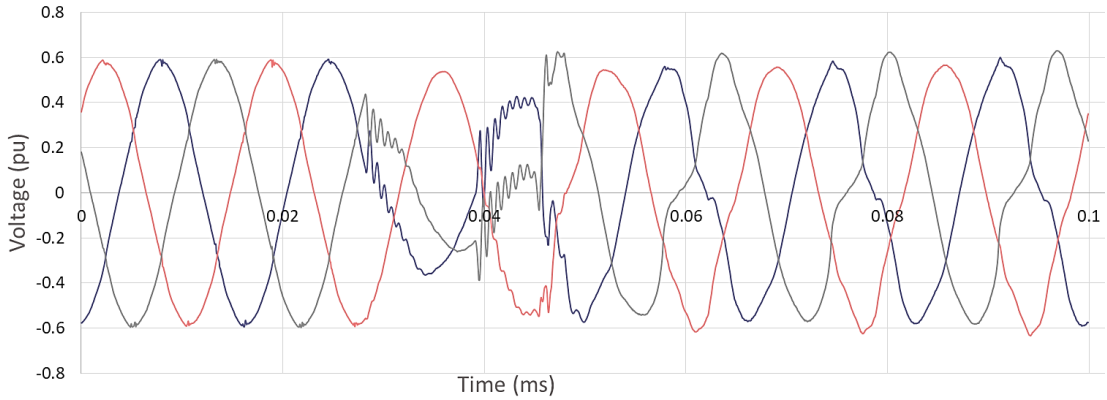

Figure 1. Example signal of the IEEE 1159.2 dataset [20].

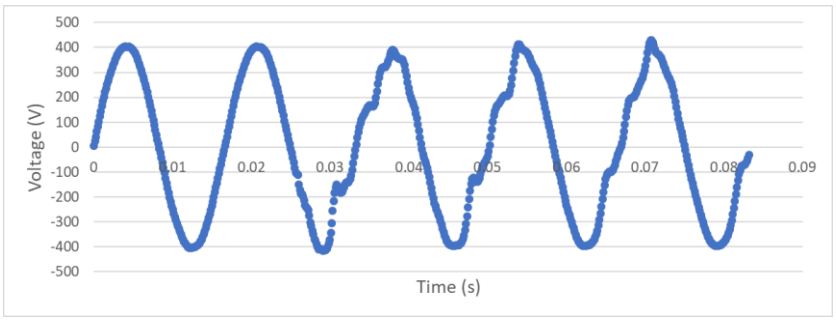

Figure 2. Example signal of the IEEE 1159.3 dataset [21].

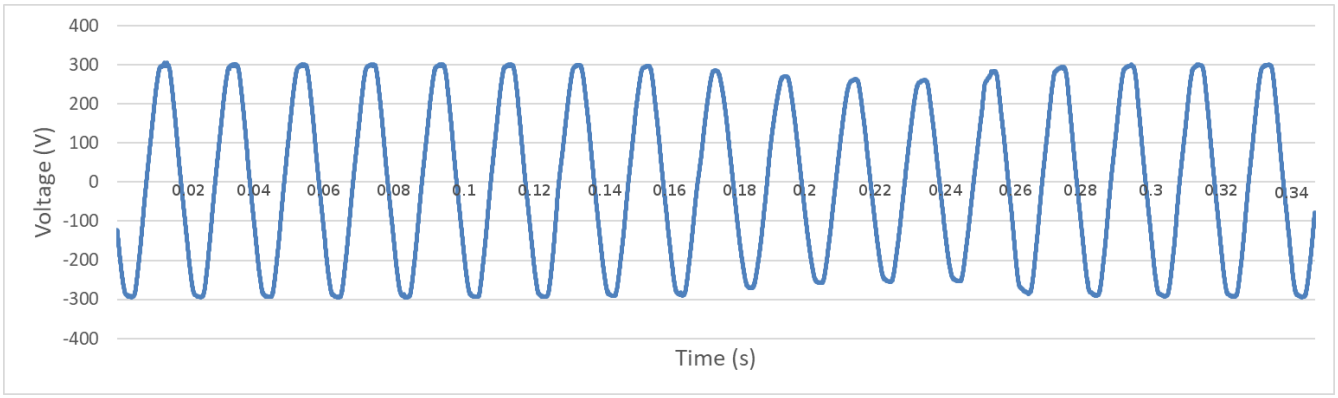

Figure 3. Example signal of the "Real life power sags" dataset [23].

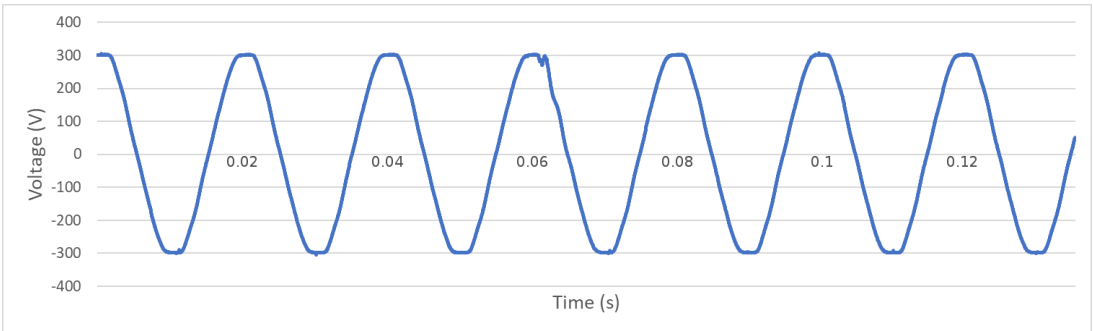

Figure 4. Example signal of the "Real life power quality transients" dataset [24].

4. Discussion

Results in Table I show that existing public datasets of power quality distortions were published several years ago. This is a clear symptom that researchers in this field are not especially likely to provide their data publicly. In fact, the four datasets were published by only two institutions (two each), the IEEE association (1159.2 and 1159.2 working groups) and the University of Cadiz. This contrasts with the number of studies that use real datasets, which is 22.5 % according to our estimates [27]. The lack of public data makes research in this field difficult, since the collection of real data is not available to all researchers.

In view of Table I, it is possible to conclude that the number of distortions included in the datasets is insufficient to train automatic power quality classifiers.

Studies in this field use hundreds or thousands of distortions to train classifiers [28]. Therefore, existing datasets can only be used to test the classification systems, but not to train them. In fact, this approach is adopted by the studies that use them (e.g.: [18], [29]). Most of them show a decrease in performance when using the real datasets [29]–[31]. This can be explained since it is not clear that synthetic data represent real distortions faithfully. Therefore, a classifier trained with synthetic data may malfunction when used in real electrical grids [27].

Table I also shows that public datasets include a limited number of types of distortions. Only sags and transients are included in the datasets. This means that distortions such as swell, interruption, flicker, notching or harmonics are not considered. Thus, if these datasets were used to validate the classifiers, only a limited number of distortions could be assessed. This is a weak point of existing datasets.

Another unanswered question is how these data were labelled. Assigning a particular type of distortion to a given signal is not easy. Human experts following existing standards or recommendations perform this assignment most times [32]. However, some unlabelled distortions may be in the samples. Therefore, classification systems that correctly identify those distortions would have poor performance, although they really worked well.

It is also important to mention that some datasets were not even clearly labelled. This hinders its use in studies of power quality classification, since researchers who want to use them must perform the labelling themselves.

In relation to sampling frequencies, they are generally high (7.7 kHz to 15.37 kHz). These values are suitable for classifying almost all types of power quality distortions, since the most frequency-demanding disturbances would meet the Nyquist criteria. According to the IEEE Recommended Practice for Monitoring Electric Power Quality [32], only very high frequency oscillatory transients, harmonics and interharmonics would remain undetected. Several studies in this field used lower sampling rates, so public datasets have appropriate values. Regarding grid fundamental frequencies, two datasets were recorded in 50 Hz-grids while the other two were collected in 60 Hz-grids. Therefore, the most common values of grid fundamental frequencies are covered by the different datasets.

In relation to the number of signal periods recorded, high variability can be observed: from 4 to 153. Many studies in this field require at least 10 signal periods to extract discriminant features from the recorded distortions (e.g.: [33]). This is a common segmentation step. Therefore, datasets with less than 10 periods could not be suitable for several classification systems. On the other hand, the detection time of distortions with several tens of periods per sample may be excessive for some applications.

Regarding file formats, a great variability was observed in the different datasets. From common txt or Excel formats to pqdiff, which must be processed in custom software. There is no standardization in the way of providing the datasets. Therefore, researchers who want to use them must implement specific processing algorithms to extract the data in a common format. None of the datasets include processing scripts.

With respect to recording conditions or devices, datasets provide detailed information, which is sufficient to contextualize them.

4. Conclusion

Although there is a growing trend towards open access research, it seems that studies on power quality classification are not taking this approach. Only four public datasets of time-domain real distortions could be found. This contrasts with the number of studies that claim to use real-world distortions (22.5 % of all existing studies according to our estimates [27]). Therefore, authors in this field do not publish their datasets, which is a barrier to research. It is possible that more public datasets might appear if other different repositories or keywords were used in the searches.

Existing public datasets have both a limited number of samples and an extremely limited number of types of distortions. Comprehensive public datasets are required. More samples should be included and, at least, the most common types of distortions such as sags, swells, interruptions, harmonics, oscillatory transients, spikes, notching or flicker, should be considered. In addition, combined distortions of two single disturbances are also required. Otherwise, the usefulness of these datasets to validate automatic power quality classifiers is very limited. Therefore, the publication of a complete dataset is a topic of future research, which is still to be addressed.

Acknowledgement

The authors gratefully acknowledge the “Fundación Iberdrola España: Ayudas a la investigación en energía y medio ambiente”, “European Social Fund, Gobierno de Aragón: FEDER, T49_17R”, “Ministerio de Educación, Cultura y Deporte: José Castillejo CAS18/218” and “Universidad de Zaragoza, Fundación Bancaria Ibercaja, Fundación CAI: IT 1/19”.

References

[1] O. P. Mahela, A. G. Shaik, and N. Gupta, “A critical review of detection and classification of power quality events,” *Renew. Sustain. Energy Rev.*, vol. 41, pp. 495–505, 2015.

[2] P. Basak, S. Chowdhury, S. Halder Nee Dey, and S. P. Chowdhury, “A literature review on integration of distributed energy resources in the perspective of control, protection and stability of microgrid,” *Renew. Sustain. Energy Rev.*, vol. 16, no. 8, pp. 5545–5556, 2012.

[3] R. Igual, C. Medrano, and F. Schubert, “Evaluation of automatic power quality classification in microgrids operating in islanded mode,” in *13th IEEE PowerTech*, 2019.

[4] M. P. Kazmierkowski, “Power Quality: Problems and Mitigation Techniques [Book News],” *IEEE Ind. Electron. Mag.*, vol. 9, no. 2, p. 62, 2015.

[5] M. K. Ahsan, T. Pan, and Z. Li, “A Three Decades of Marvellous Significant Review of Power Quality Events Regarding Detection & Classification,” *J. Power Energy Eng.*, vol. 06, no. 08, pp. 1–37, 2018.
[6] R. Igual, C. Medrano, F. J. Arcega, and G. Mantescu, “Integral mathematical model of power quality disturbances,” in *Proceedings of International Conference on Harmonics and Quality of Power, ICHQP*, 2018, vol. 2018-May.
[7] S. A. Deokar and L. M. Waghmare, “Integrated DWT–FFT approach for detection and classification of power quality disturbances,” *Int. J. Electr. Power Energy Syst.*, vol. 61, pp. 594–605, 2014.
[8] J. G. M. S. Decanini, M. S. Tonelli-Neto, F. C. V Malange, and C. R. Minussi, “Detection and classification of voltage disturbances using a Fuzzy-ARTMAP-wavelet network,” *Electr. Power Syst. Res.*, vol. 81, no. 12, pp. 2057–2065, 2011.
[9] S. Naderian and A. Salemnia, “An implementation of type-2 fuzzy kernel based support vector machine algorithm for power quality events classification: An implementation of T2FK-SVM for power quality events classification,” *Int. Trans. Electr. Energy Syst.*, vol. 27, 2016.
[10] A. A. Abdoos, P. Khorshidian Mianaei, and M. Rayatpanah Ghadikolaei, “Combined VMD-SVM based feature selection method for classification of power quality events,” *Appl. Soft Comput. J.*, vol. 38, pp. 637–646, 2016.
[11] F. A. S. Borges, R. A. S. Fernandes, I. N. Silva, and C. B. S. Silva, “Feature Extraction and Power Quality Disturbances Classification Using Smart Meters Signals,” *IEEE Trans. Ind. Informatics*, vol. 12, no. 2, pp. 824–833, 2016.
[12] N. Huang, D. Xu, X. Liu, and L. Lin, “Power quality disturbances classification based on S-transform and probabilistic neural network,” *Neurocomputing*, vol. 98, pp. 12–23, 2012.
[13] R. Kumar, B. Singh, and D. T. Shahani, “Symmetrical Components-Based Modified Technique for Power-Quality Disturbances Detection and Classification,” *IEEE Trans. Ind. Appl.*, vol. 52, no. 4, pp. 3443–3450, 2016.
[14] D. Granados-Lieberman, M. Valtierra-Rodriguez, L. Morales-Hernández, R. Romero-Troncoso, and R. Osornio-Rios, “A Hilbert Transform-Based Smart Sensor for Detection, Classification, and Quantification of Power Quality Disturbances,” *Sensors (Basel).*, vol. 13, pp. 5507–5527, 2013.
[15] R. Hooshmand and A. Enshaee, “Detection and classification of single and combined power quality disturbances using fuzzy systems oriented by particle swarm optimization algorithm,” *Electr. Power Syst. Res.*, vol. 80, no. 12, pp. 1552–1561, 2010.
[16] P. Kanirajan and V. S. Kumar, “Power quality disturbance detection and classification using wavelet and RBFNN,” *Appl. Soft Comput.*, vol. 35, pp. 470–481, 2015.
[17] A. K. Puliyadi Kubendran and A. K. Loganathan, “Detection and classification of complex power quality disturbances using S-transform amplitude matrix–based decision tree for different noise levels,” *Int. Trans. Electr. Energy Syst.*, vol. 27, no. 4, p. e2286, 2017.
[18] Z. Liu, Y. Cui, and W. Li, “A Classification Method for Complex Power Quality Disturbances Using EEMD and Rank Wavelet SVM,” *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1678–1685, 2015.
[19] H. Erişti, Ö. Yıldırım, B. Erişti, and Y. Demir, “Optimal feature selection for classification of the power quality events using wavelet transform and least squares support vector machines,” *Int. J. Electr. Power Energy Syst.*, vol. 49, pp. 95–103, 2013.
[20] [dataset] IEEE Power and Energy, “IEEE 1159.2 Working Group: Test Waveforms.” [Online]. Available: http://grouper.ieee.org/groups/1159/2/testwave.html. [Accessed: 12-Oct-2019].
[21] [dataset] IEEE Power and Energy Society, “IEEE 1159.3 PQDIF Task Force.” [Online]. Available: http://grouper.ieee.org/groups/1159/3/. [Accessed: 12-Oct-2019].
[22] R. Igual, S. Miraftabzadeh, F. Foiadelli, and C. Medrano, “Synthetic data of the paper: ‘Quantification of feature importance in automatic classification of power quality distortions.’” IEEE Dataport, 2019.
[23] [dataset] Florencias-Oliveros, M. J. Espinosa-Gavira, J. J. González de la Rosa, A. Agüera-Pérez, J. C. Palomares-Salas, and J. M. Sierra-Fernández, “Real-life Power Quality Sags,” 2017. [Online]. Available: http://dx.doi.org/10.21227/H2K88D. [Accessed: 12-Oct-2019].
[24] [dataset] Florencias-Oliveros, M. J. Espinosa-Gavira, J. J. Gonzalez de la Rosa, A. Aguera-Perez, J. C. Palomares-Salas, and J. M. Sierra-Fernandez, “Real-life Power Quality Transients,” 2017. [Online]. Available: http://dx.doi.org/10.21227/H2Q30W. [Accessed: 12-Oct-2019].
[25] G. [dataset] Spavieri, R. T. M. Ferreira, R. A. S. Fernandes, G. G. Lage, D. Barbosa, and M. Oleskovicz, “Laboratorial Essays of Polypropylene and All-film Power Capacitors.” IEEE Power and Energy Society, 2017.
[26] V. Jacome, “Power Quality and Modern Energy for All,” 2019. [Online]. Available: https://dataverse.harvard.edu/dataverse/powerqualityandmodernenergyforall. [Accessed: 27-Oct-2019].
[27] R. Igual and C. Medrano, “[Under Review] Research challenges on real-time classification of power quality disturbances: A systematic review,” *Renew. Sustain. Energy Rev.*, vol. Unpublishe, 2019.
[28] S. Jamali, A. R. Farsa, and N. Ghaffarzadeh, “Identification of optimal features for fast and accurate classification of power quality disturbances,” *Measurement*, vol. 116, pp. 565–574, 2018.
[29] H. Dehghani, B. Vahidi, R. A. Naghizadeh, and S. H. Hosseinian, “Power quality disturbance classification using a statistical and wavelet-based Hidden Markov Model with Dempster–Shafer algorithm,” *Int. J. Electr. Power Energy Syst.*, vol. 47, pp. 368–377, 2013.
[30] A. Rodríguez, J. A. Aguado, F. Martín, J. J. López, F. Muñoz, and J. E. Ruiz, “Rule-based classification of power quality disturbances using S-transform,” *Electr. Power Syst. Res.*, vol. 86, pp. 113–121, 2012.
[31] S. Upadhyaya and S. Mohanty, “Localization and Classification of Power Quality Disturbances using Maximal Overlap Discrete Wavelet Transform and Data Mining based Classifiers∗∗Sponsor and financial support acknowledgment goes here. Paper titles should be written in uppercase and lower,” *IFAC-PapersOnLine*, vol. 49, no. 1, pp. 437–442, 2016.
[32] IEEE, *IEEE Recommended Practice for Monitoring Electric Power Quality*, vol. 2009, no. June. 2009.
[33] M. Zhang, K. Li, and Y. Hu, “A real-time classification method of power quality disturbances,” *Electr. Power Syst. Res.*, vol. 81, no. 2, pp. 660–666, 2011.