Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

An image-based deep transfer learning approach to classify power quality disturbances

Grazia Todeschini [a,*], Karan Kheta [b], Cinzia Giannetti [b]

[a] Department of Engineering, King's College London, The Strand, London, WC2R 2LS, England, United Kingdom
[b] Department of Mechanical Engineering, Swansea University, Bay Campus, Fabian Way, Swansea, SA1 8EN, Wales, United Kingdom

ARTICLE INFO

Keywords:
Convolutional neural networks
Power quality disturbance
Power quality
Monitoring
Transfer learning
Voltage sag
Voltage swell

ABSTRACT

Power quality disturbances (PQDs) consist in deviation of voltage and current waveforms from the ideal sinusoid at fundamental frequency, and need to be monitored to ensure a reliabile electrical supply. While, traditionally, power quality monitoring has been performed using signal processing techniques, coupled with shallow Machine Learning classifiers or wave change detection methods, more recently, new approaches, based on Deep Learning, have been proposed. These methods have the potential to achieve high classification accuracy and to remove the need of extensive data pre-processing, hence being more suitable for real-time deployments. However, high classification performance has been only demonstrated using synthetically generated data. In order to address limitations related to processing time and accuracy, this paper proposes a novel end-to-end framework for automated detection of PQDs based on Deep Transfer Learning. The proposed approach uses a small set of images of voltage waveforms to train the model and classify different types of PQDs. This method leverages on the high performance of existing pre-trained models for image classification and shows consistent high accuracy for data with varying resolution. The proposed methodology provides a pathway towards effective deployment of Deep Learning in power quality monitoring systems and real-time applications.

1. Introduction

In an ideal power system, voltage and current waveforms are sinusoid at constant frequency (50 Hz or 60 Hz). Deviations from the ideal waveform are referred to as 'power quality disturbances' (PQDs). Numerous types of PQDs exist, and involve both waveform amplitude and frequency. Several standards have been developed to define measurement procedures and to harmonize PQD classification [1, 2].

PQDs are a concern for system operators and for customers because they may cause a number of operational problems. Severe PQDs may result in unacceptable operating conditions, thus causing in unplanned outages and interruption of supply to customers. Moderate PQDs bring their concerns too: if not mitigated, they may cause damage to the equipment on the long term, or reduction of its lifetime. For example, small levels of harmonic distortion may lead to pulsating torques in motors or hot-spot temperatures in transformers, that may not be detected and therefore degrade the equipment [3]. Various solutions have been developed to mitigate power quality disturbances, such as passive filters or STATCOMs. In order to choose the location and the rating of such equipment, it is necessary to undertake monitoring of the power system performance.

With the increase of power-electronics based devices installed on the power grid, power system operators worldwide have been observing an increase in the occurrence of PQDs, thus causing growing power quality concerns [4]. This situation, combined with the decreasing cost of measuring equipment, has been resulting in a proliferation of power quality monitoring. The most traditional approaches to power quality monitoring employ triggering features to detect PQDs and store them as individual events. More recently, adoption of triggerless continuous measurements of instantaneous voltage and current waveforms has been observed [5, 6]. As a result of the above trends, a large amount of power quality recordings is becoming available, thus leading to challenges in terms of data processing, management and storage.

Numerous methods have been proposed to perform automatic PQD classification. These methods generally require a feature extraction step, followed by a classification algorithm [7]. The most commonly employed methods for features extraction include Fast Fourier Transformation (FFT), wavelet transformation and Hilbert Huang transform (HHT) [8], while the classification step can be achieved using supervised Machine Learning (ML) methods such as Artificial Neural Networks (ANN), Decision Tree, Support Vector Machine (SVM) or using rule

[*] Corresponding author.
E-mail address: grazia.todeschini@kcl.ac.uk (G. Todeschini).

https://doi.org/10.1016/j.epsr.2022.108795
Received 16 May 2022; Received in revised form 16 August 2022; Accepted 4 September 2022
Available online 16 September 2022
0378-7796/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

based systems [8]. A recent paper proposes a method that combines multidomain feature extraction, Self Organising Maps (SOM) and ANN to classify stationary and non-stationary disturbances for wind turbine generators [9]. The model is trained on a synthetic dataset and tested using two different sets of real signals. The accuracy obtained using synthetic data is up to 100%, but the accuracy obtained when using real data is not provided. As argued in [10], the main limitation of feature extraction approaches is that the accuracy of the prediction relies on the selection of hand crafted features, which requires substantial effort. Furthermore, abnormal wave-shapes correspond to a very small fraction of the recordings [11], thus resulting in unbalanced training sets that may not be suitable when employing traditional supervised learning methods. To address this issue, alternative approaches based on waveshape change detection have been proposed [6, 12].

Alternatively, several authors have proposed methods based on Deep learning (DL) [13–18]. DL is a particular class of ML and refers to techniques for learning high-level features from data in a hierarchical manner using stacked layer-wise architectures. The main advantage of DL over more traditional methods is that it uses data representation learning rather than explicit engineered features to perform the classification task, hence removing the need for time-consuming feature engineering. DL has gained popularity in recent years, and it has been successfully applied to various domains [19] and time series classification [20, 21]. Ma et al. [22] proposed a method based on stacked auto-encoders to extract high-level features of PQDs from simulated waveforms. Mohan et al. [16] studied and compared several DL architectures to identify power disturbances using both synthetic and real world data. Results showed that the performance of the classifier is higher when trained with synthetic waveform distortions and voltage fluctuations generated with parametric equation, while lower accuracies were achieved when using real word PQDs data, due to the lack of sufficient training data for some disturbances. In [14], Liu et al. presented an approach that combines singular spectrum analysis, curvlet transformation and deep CNNs to classify PQDs. The method is shown to achieve high accuracy, but the classification task is not completely automated as it still relies on intermediate data pre-processing steps for feature extraction. This limitation was overcome by [13] who proposed a closed-loop framework to detect and classify PQDs, employing deep CNNs for both feature extraction and classification. This model was trained using a simulated dataset with 768,000 samples. It is shown that the method achieves higher accuracy and lower computation time when compared with state of the art approaches. However, the method uses synthetic data obtained from parametric equations. [18] proposes a new CNN architecture, that achieves high accuracies for 29 types of PQDs, and has been tested on synthetic data.

Whilst the above mentioned methods use raw signals as an input, a different approach is proposed by Balauji et al. [15] using image files of the three-phase PQ event data collected for one year from four different regions in Turkey. The raw data is converted into voltage rms graphs, manually labelled. Despite the promising results, not enough details are given regarding the methodology applied for training and testing of the model and there is no discussion about over-fitting problems.

With the exception of [15] and [16], the studies listed above indicate that high accuracies have been obtained when training DL models on synthetic data, but a research gap has been identified in their validation with real world measurements. Large amounts of data are necessary to train DL models, requiring adequate computational resources and complex optimization procedure to avoid over-fitting. Furthermore, when using real world data, the lack of specific instances of PQDs anomalies, only occurring in rare cases, may bias the algorithm and produce low accuracy for under-represented classes if the model is trained with unbalanced data sets. This aspect is addressed in [23] where a CNN is fully trained from scratch and a real world data is used. This network provides promising results, however, it should be noted that seven out of twenty PQDs for each classifiers were used for training, corresponding to 30% of the available data.

In order to address the above challenges in obtaining accurate prediction for real PQDs, this paper proposes a novel end-to-end Deep Transfer Learning (DTL) framework that uses images of voltage waveforms as training samples. Instead of training the classification model from scratch, transfer learning (TL) enables to reuse existing image classification CNNs and, through a process called ‘fine-tuning’, adapt them for a new task: in the case of this paper, classification of voltage waveforms. TL reduces the complexity and time of the training process and requires only a small number of images of anomalies to achieve a high predictive performance. The PQD events considered in this work are voltage sags, voltage swells and interruptions, obtained from power quality monitors installed at various locations. These disturbances were chosen since a sufficient number of events were available. According to the proposed approach, labelled voltage waveform images are fed to several pre-trained CNNs to identify the model that leads to the best results in terms of accuracy and computation time. Because a pre-trained network is used, the proposed approach does not require extensive knowledge of the models. Simultaneously, this approach eliminates the need for hand crafted feature extraction methods and complex model optimisation, facilitating the deployment of DL for automated classification of PQDs in real word applications. Furthermore, the method requires a small number of images for each anomaly, hence further reducing the time necessary for training. To the best of the author’s knowledge, this is the first example of an image-based PQDs classifier trained and evaluated on real-world power quality recording, which obtains high accuracy by leveraging advances in image classification achieved by pre-trained CNNs.

The paper is structured as follows: Section 2 introduces the TL paradigm and the CNNs considered for the study; Section 3 presents the proposed DTL approach; Section 4 presents experimental results and quantifies the accuracy of the proposed approach. Section 5 compares the proposed approach with other ones presented in the literature. Finally, Section 6 draws the conclusions and outlines future work in this area.

2. Background: transfer learning and CNNs

Transfer learning is an emerging paradigm that enables training predictive models by reusing data, models and knowledge from other tasks or domains [24, 25]. TL methods have been applied to time series prediction [26], natural language processing, sentiment analysis and image recognition [24, 27, 28].

DL architectures based on CNNs have emerged as very effective tools for image recognition [29]. CNNs are layered architectures including convolutional, activation and pooling layers [30]; numerous CNNs have been proposed and they are available as pre-trained models. Most of these architectures has been developed using the ImageNet database [31], with advances driven by the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). ImageNet contains millions of images of common objects such as keyboards, coffee mugs, pencils, and animals [31]. CNN models are particularly useful in image classification because they have large learning capability, and they are able to make strong and mostly correct assumptions about images features [32]. Furthermore, in comparison to feedforward neural networks of similar size, they have fewer parameters to train [32].

Since the high level features learned to classify images are generic, the TL paradigm allows to re-use highly optimized and efficient pretrained models for other image classification tasks. Typically, TL is achieved by keeping the upper layers of the CNN unchanged and finetuning the lower layers to specialize the learning towards the new task. Compared to training a CNN from scratch, TL requires less computational resources, resulting in faster training time as well as reducing over-fitting when training with small size datasets. Because the proposed methodology specifically applies TL to DL architectures (CNNs), the acronym DTL will be used in the rest of the paper.

In this study, three networks trained with the ImageNet dataset were considered: SqueezeNet, GoogLeNet and ResNet. These architectures were selected because they represent networks of increasing level of complexity, with the aim of finding the best trade-off between accuracy and training time. These networks are briefly described below, while a detailed review of their performance for generic image classification can be found in [29]. The network selection process for the purpose of PQD classification is described and discussed in Section 4.1.

SqueezeNet is small network that achieves AlexNet-level accuracy on ImageNet with 50x fewer parameters [33]. The number of filters per fire module is gradually increased from the beginning to the end of the network. The advantage of this architecture is that it requires less communication across servers during distributed training and less bandwidth to export a new model from the cloud to an embedded system. Smaller CNNs are more feasible to deploy in hardware with limited memory.

GoogLeNet won the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014 [34]. It implements a novel element called inception module and the architecture consists of a 22 layer deep CNN and the number of parameters are reduced to 4 million [30] (while in AlexNet they are 60 millions).

ResNet is based on a novel architecture with 'skip connections'and features heavy batch normalization [35]. Using this technique, the authors were able to train a NN with 152 layers (instead of 22 in GoogLeNet). ResNet achieved a top-5 error rate of 3.57% which beats human-level performance and was awarded the first place in ILSVRC 2015.

3. Deep transfer learning framework for PQD classification

An overview of the generic end-to-end DTL framework proposed in this study is presented in Fig. 1. The first step consists in selecting a pre-trained network, featuring convolutional layers and final classification layers. Convolutional layers learn low-level generic features such as to recognize blobs, edges and colours from the ImageNet dataset. The final classification layers recognise specific patterns for classification. Domain-specific training images (i.e. voltage waveforms) are used to retrain the network via DTL: as shown in the centre of Fig. 1, the convolutional layers of the pre-trained are kept, while the final layers are replaced with a fully-connected layer and an output layer. These two layers learn details about the new images, and the number of outputs for the new fully connected layer is equal to the number of labels. Following hyperparameter optimization, a new model (optimized network) is obtained. At the deployment stage, the model is fed with new images, and as a final result a label is associated with each image, thus resulting in PQD classification.

Two examples of classification will be presented in this paper: for binary classification, the labels are 'normal' and 'abnormal'. In this context, 'normal' waveforms are the ones that are compliant with the characteristics described by the power quality standards [1, 2]. The remaining waveforms are classified as 'abnormal'. For classification of multiple PQDs (multiclass classification), four labels are possible: 'normal', 'voltage sag', 'voltage swell' and 'interruption'.

3.1. Data acquisition

Power quality disturbance data for this study are extracted from the PQube power quality monitors recordings. These recordings are available on the free cloud-based map by Power Standard Labs [36]. Several PQube monitors have been installed worldwide, and they provide voltage and current readings as a database of events, including various PQDs and snapshots of normal waveforms. The time series for each disturbance can be downloaded in various formats (for examples .txt or . csv files), and the events are accompanied with a label. Additionally, an image for each time series is available. This source has been chosen because it is freely available and continuously updated: as a result, the methodology presented in this paper can be reproduced and new recordings can be retrieved every few weeks for further validation and tuning of the models.

A number of locations were checked for suitability, and it was found that the meters located at the University of Strathclyde (UK), ECOXplore Pte Ltd (Singapore) and Keystone Hatchery (South Africa) provided data which were comparable. More specifically, these meters provided a consistent set of three-phase waveforms for a several months and various PQDs. For each event, both voltage waveform time-series and images were downloaded, and the data was organized into folders by their label. The same dataset was used in [37].

For network training and optimisation (Section 4.1–Section 4.3), power quality data were retrieved from the website from March 2017 until September 2019. 100 waveforms were used for initial training and network selection, and 167 for fine-tuning of the selected model. For out-of-sample testing (Section 4.4), 154 additional data was retrieved between October 2019 and December 2019. The dataset used for this work is relatively small, due to availability of data, and one of the aims of the proposed approach is the development of classification methods that are robust in spite of the size of the dataset.

Other PQDs, such as harmonics, were not considered due to the lack of data. While it would have been possible to generate synthetic data for the missing PQDs, one of the novel contributions of this work consists in demonstrating the suitability of proposed DTL approach applied to PQDs on real-world data. As discussed in the introduction, the use of synthetic waveforms for PQDs classification using CNNs has been already presented in other works such as [16, 22].

3.2. Data pre-processing

A first attempt to classify PQDs was made by using directly the

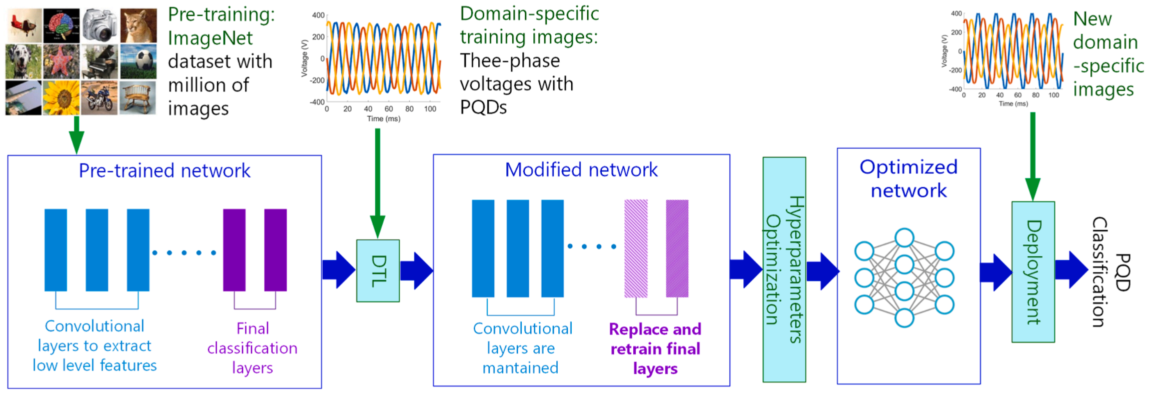

Pre-training: ImageNet dataset with million of images
Voltage (V)
400
200
0
-200
-400
0
20
40
60
80
100
Time (ms)
Domain-specific training images: Thee-phase voltages with PQDs
New domain -specific images
Pre-trained network
Convolutional layers to extract low level features
Final classification layers
DTL
Modified network
Convolutional layers are mantained
Replace and retrain final layers
Hyperparameters Optimization
Optimized network
Deployment
PQD Classification

Fig. 1. The proposed end-to-end DTL framework replaces the two final layers of a pre-trained network with a fully connected layer and an output layer. The model is retrained using appropriately formatted images of voltage waveform. Following hyperparameters optimization, the new model is deployed to label new PQDs.

images retrieved from [36]. An example of such images is shown in Fig. 2: from the top, the line-to-line voltage, line-to-neutral voltage, neutral-to-earth voltage, phase currents and neutral current are shown. These images provided to be unsuitable for the proposed application for various reasons: in the case of long disturbances, some figures include a central cutaway, as shown in Fig. 2; in other cases, images of the same disturbance from different events had varying *x*-axis length, thus making them not comparable.

Therefore, it was decided to write a script to automatically create consistent images suitable for training and testing. The raw data (in .csv format) was retrieved and a MATLAB script was developed to read sampled values, sampling rate and labels from the raw data and generate images such as the one shown in Fig. 3. The images were then saved with unique names and were organized into folders by type of disturbance. The folder structure and the image resolution (224 × 224) were designed to meet the specification of the pretrained network. The proposed approach is versatile as the images can be generated from the .csv files using any software of choice, and source files can be retrieved from different power quality meters with varying resolution. The requirements to meet are: high resolution and contrast, size as stated above, and consistency of the axes.

In this project, the typical PQD duration was three to five cycles, and this is a typical behaviour due to the transients associated to power system operation. The waveforms used here are aligned with the typical recording shown in [2] - even if shorter PQD duration is theoretically possible, they are not common in practical power systems. As a result, the images include six cycles of the three-phase voltage waveform, however, the number of cycles can be easily modified by changing one parameter in the script.

3.3. Model training and testing

To validate and test the proposed DTL methodology, three incremental steps were undertaken as outlined below, with more details and results provided in Section 4. Each step allows reinforcing the confidence in the model and results in further enhancements.

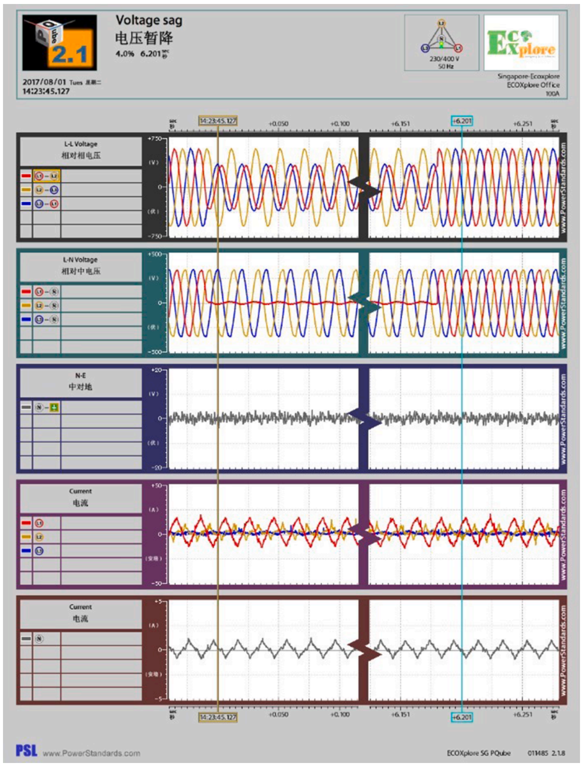


Fig. 2. Screenshot of a voltage sag from PQube, [36].

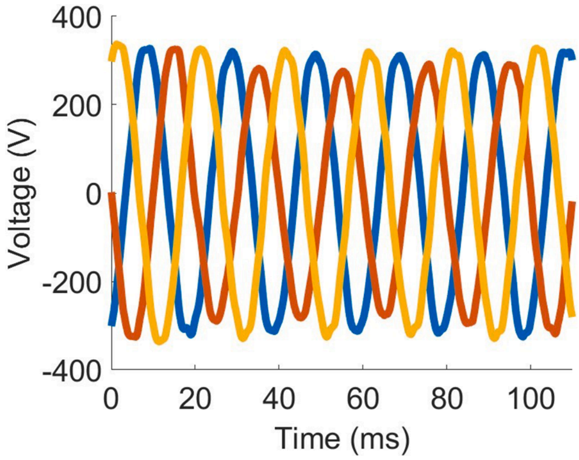


Fig. 3. Voltage sag waveform generated using the MATLAB script developed for this project.

Binary classification and network selection (Section 4.1): three well-researched pretrained networks (SqueezeNet, GoogleNet and ResNet) were trained and tested for suitability of the proposed application, namely their ability to achieve a binary classification (normal vs abnormal). High accuracy was obtained with all three models, and the best network was identified.

Multiclass classification (Section 4.2) and hyperparameter optimization (Section 4.3): for the best network only, the proposed framework was repeated on the more complex multiclass classification problem to recognize individual PQDs. As a result, five model were developed, with different hyperparameters settings.

Out-of-sample test (Section 4.4): using the five developed models, the multiclass classification was repeated using new data collected few months after the initial experiments. This collection of waveforms represents an independent dataset that allows assessing the ability of the five models developed above to classify PQDs from new waveforms. As a result, the 'best network overall' was identified.

In all tests outlined above, the weights of the convolutional layers were frozen by setting their learning rate to zero: as a result, while the networks were training, the parameters of these layers were not modified. This process has been adopted because the available data set was small, thus preventing the earlier layers from overfitting. Overfitting refers to the process where the model fits too well the training data, thus being unable to make correct predictions on unseen datasets [19], and it has been observed when using small datasets. Furthermore, freezing the convolutional layers reduce the time taken by the training process.

Hyperparameters were kept fixed for binary classification. When training the classifier for identifcation of multiple PQDs, hyperparameter optimization was carried out by using a grid search approach, due to ease of implementation. The following training options were used: max *epochs, mini batch size* and *learn rate*. The max *epochs* is how many times the data passes through the network forward and backward, the *mini batch size* is how many images are in a batch to divide the whole data set and the *learn rate* is the hyperparameter that establish how quickly the model adapts. *Learn rate* is set to a non-zero value for the layers which are not frozen, while for the other layers it is equal to zero). Stochastic gradient descent (SGD) method [38] was chosen as optimization method. Accuracy measure and confusion matrix were used to evaluate the model performance following the steps described in Section 4.3.

The advantage of the proposed DTL framework compared to other approaches is that it enables training the model with a small number of images, using the ability of pre-trained models to learn low-level features in an efficient way. Another advantage of this approach is that it significantly curtails the number of hyperparameters, hence reducing the time taken for model optimization.

4. Evaluation and results

4.1. Binary classification and pre-trained network selection

The initial evaluation consisted in testing the ability of three CNNs (i. e. SqueezeNet, GoogleNet and ResNet) to classify abnormal and normal waveforms (binary classification). This set of experiments guided the choice of the best pre-trained model, based on the compromise between training time and accuracy. The networks were trained using a small dataset containing 100 images extracted from PQube. These images were generated using the method described in Section 3.2 and split in two sets, with 70% of the images used for training and 30% for testing.

In this phase the hyperparameters were set to constant values: max *epochs* was equal to 35, *mini batch size* to 30 and the *learn rate* to 3e-4. All other parameters were left as default. The three CNNs were used to classify the test images, and their performance was evaluated using the accuracy measure (percentage of the images successfully classified over the total number of images) and the confusion matrix. This process led to the choice of the best network, and it was repeated three times for repeatability.

The results of the experiments are shown in Table 1. SqueezeNet took the shortest time and provided the lowest accuracy. Both GoogLeNet and ResNet were able to predict with 100% accuracy, however ResNet took approximately twice as long than GoogLeNet. As a result, GoogLeNet was chosen for the remaining experiments for its balance between training time and accuracy prediction.

As part of this initial evaluation, image augmentation (i.e. the modification of the existing dataset to generate additional images) was tested to assess if it would yield any benefits. To this end, further experiments were carried out. It was observed that augmentation causes significant loss of details as well as the introduction of new differences between images, resulting in overall lower accuracy as well as large variation in accuracy, ranging from 66.7% to 100%. Therefore, image augmentation was not used in the remaining experiments.

4.2. Classification of multiple PQDs

In this case, a new dataset with 167 images was created, where 116 were used for training and 51 for testing. More specifically the dataset contained 60 normal waveforms, 46 voltage sags, 39 interruptions and 22 voltage swells. The inequality in the number of waveforms for each of the classes is due to the availability of voltage disturbances on the PQube repository. It is worth noticing that the training set is aligned with realistic distribution of PQDs, as voltage sags are the most common types of events. Additionally, the dataset is small by design to demonstrate the ability of the network to learn from small data samples. In the training process 70% of the images were used for training and 30% for testing.

The majority of the data was retrieved from [36], as described in Section 3.1. A few waveforms were generated analytically to increase the size of the dataset (12 swells were generated by scaling some of the waveforms retrieved from [36]) and to improve the classification of waveforms that are close to the limits provided by the standards (some normal waveforms with a voltage amplitude equal to +105% and +95% of the rated value were created), respectively.

Table 1
Prediction accuracy using three pre-trained CNNs.

Network	Time	Accuracy
SqueezeNet	95.61 s	97%
GoogLeNet	158.5 s	100%
ResNet	373.1 s	100%

4.3. Hyperparameter optimization

As described above, the *mini batch size*, max *epochs* and *learn rate* were considered for hyperparameter optimization and a total of 45 experiments were carried out. In these experiments max *epochs* varied between 5 and 45 with steps of 5, while *mini batch size* varied between 10 and 50 with steps of 10. Initial experiments showed that the optimal learning rate was 0.0003, with lower values showing unstable results and large variations of accuracy - therefore this parameter was fixed in all runs.

To assess stability, every optimization run was performed five times and the mean accuracy and time were recorded. A randomly chosen hold out set containing 30% of the data was used at each run to calculate overall accuracy.

The average accuracy obtained with this method is shown in Table 2, where the lowest accuracy is 94.9%, obtained with max *epochs* of 5 and *mini batch size* of 50. The highest accuracy of 100% was achieved in five cases highlighted in Table 2: the values of mini batch size and max epochs for these five networks are (10,45), (30,10), (30,40), (40,35), and (40,40), respectively. The confusion matrix for one of the five networks is illustrated in Fig. 4, showing that this network is able to classify the 51 waveforms used for testing with 100% accuracy. Two examples of output predictions for a voltage sag and for an interruption are shown in Fig. 5 and Fig. 6. For each example, the model associates a label to the PQD under consideration, and a level of confidence measured in percentage (class probability).

Results of the average training time when varying max epochs and mini batch size are listed in Table 3. A clear trend is identified as increasing the max epochs increases the training time of the network. However, Table 2 shows that increasing max epochs gives higher accuracies. *Mini batch size* has some impact on the accuracy as smaller batch sizes provide marginally higher accuracies, but the effects are more pronounced in training time.

As shown in Table 3, higher batch sizes lead to shorter training times.

4.4. Out-of-sample testing

In order to further test the robustness of the proposed approach, an out-of-sample test dataset was collected four months after the initial experiments consisting of 154 waveforms. The data set contained 88 normal waveforms, 41 voltage sags, 18 voltage swells and 6 interruptions. The five networks that achieved 100% accuracy following model optimization were selected for out-of-sample testing. Accuracy results for out-of-sample testing are summarized in Table 4.

The worst performing network was $\mathit{Net}_{(30,40)}$ achieving 81.7% overall accuracy in this case, 21 instances of normal waveforms are classified as swells, while four swells are classified as sags. The best performing network was $\mathit{Net}_{(35,40)}$ that achieved 99.3% accuracy: according to the confusion matrix shown in Fig. 7, all but one disturbance are classified correctly. The only incorrect classification is a voltage sag being classified as an interruption - however, as it will be described in the next section, this incorrect prediction is due to wrong label associated to the original waveform. $\mathit{Net}_{(40,40)}$ achieved similar results as $\mathit{Net}_{(30,40)}$ with some occurrences of normal waveforms classified as PQDs, achieving 83.7% accuracy.

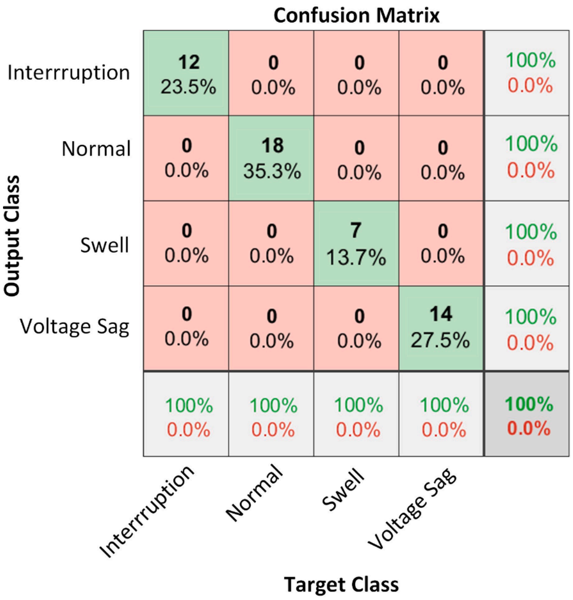
4.5. Incorrect predictions and voltage swell classification

The results described in the previous section show that $\mathit{Net}_{(35,40)}$ provided the best accuracy. The confusion matrix for this model is shown in Fig. 7. For this case, the classification resulted in one mismatch only, caused by associating an initial wrong label to the waveform data.

More specifically, Fig. 8 is classified as 'voltage sag' by the PQube metre, however, all five networks used for out-of-sample testing classified it as interruption: this result is shown in the fourth cell of the top row for the confusion matrix of Fig. 7. For the first cycles, the voltage

Table 2
Average accuracy (as a%with varying max epochs and batch size over five runs). five cases resulted in 100% accuracy (bold font).

Batch Size	Max Epochs								
	5	10	15	20	25	30	35	40	45
10	97.6	97.6	98.8	98.0	98.4	98.4	99.6	99.2	**100**
20	98.0	98.4	99.2	98.4	98.0	99.6	98.8	98.4	99.2
30	98.4	**100**	99.2	98.8	99.2	99.2	99.6	**100**	98.8
40	96.5	96.9	99.2	99.2	99.6	99.2	**100**	**100**	99.2
50	94.9	99.2	99.2	99.6	99.6	98.8	99.6	99.2	99.2


Confusion Matrix

Output Class \ Target Class	Interruption	Normal	Swell	Voltage Sag	
Interrruption	**12** 23.5%	**0** 0.0%	**0** 0.0%	**0** 0.0%	100% 0.0%
Normal	**0** 0.0%	**18** 35.3%	**0** 0.0%	**0** 0.0%	100% 0.0%
Swell	**0** 0.0%	**0** 0.0%	**7** 13.7%	**0** 0.0%	100% 0.0%
Voltage Sag	**0** 0.0%	**0** 0.0%	**0** 0.0%	**14** 27.5%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	**100%** **0.0%**

Output Class
Target Class

Fig. 4. Confusion matrix for the optimized networks identified in Table 2. The networks are able to classify all PQDs with 100% accuracy.

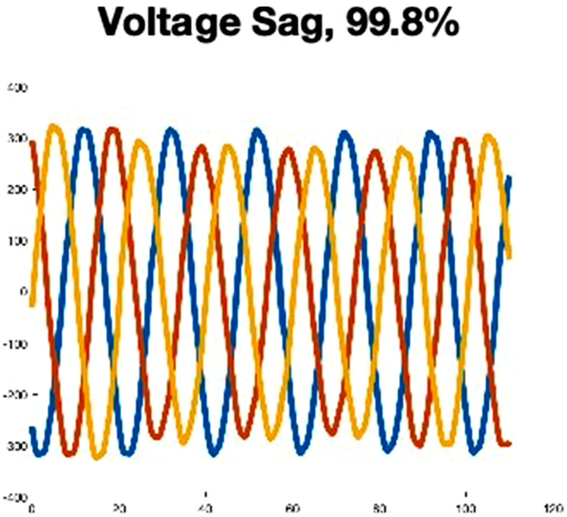


Fig. 5. Output prediction and score for a sample voltage sag.

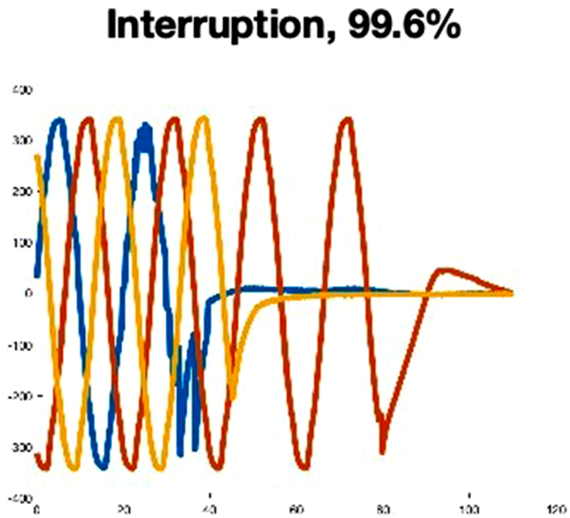


Fig. 6. Output prediction and score for a sample interruption.

waveform is close to 'normal', and around 30 ms a voltage drop takes place on all phases with a linearly decreasing amplitude. Although the snapshot provided does not show the voltage dropping to zero, it seems likely that within a few cycles an interruption will occur. Therefore, the label applied by the optimized CNNs appears appropriate and allows identifying an anomaly thus flagging a waveform for further analysis.

As a second observation, $\mathit{Net}_{(40,35)}$ was able to classify all voltage swells correctly, as shown in Fig. 7, in spite of the low number of voltage swell images available for training. This result is significant because some of the voltage swells used for out-of-sample testing were artificially created, and no similar waveforms were included in the training set. On the contrary, $\mathit{Net}_{(30,40)}$ misclassified some of the swells as 'voltage sag'. This anomaly is worth exploring, as it allows explaining the importance of network optimisation. Fig. 9 was created artificially and labelled 'voltage swell' because the voltage profile of one phase exceeds 105% of the rated value. The other two phases are either close or below the rated value. Therefore, both a voltage sag and a voltage swell are present in the waveform, but the label 'voltage swell' is associated to this image because this is the prominent disturbance. $\mathit{Net}_{(30,40)}$ labels the waveform shown in Fig. 9 as 'voltage sag'. This result is explained as follows: because the training dataset contained a large number of voltage sags, the network $\mathit{Net}_{(30,40)}$ is biased toward this disturbance. The result provided by the model is acceptable because a voltage sag is present in the waveform, although it is not the most prominent disturbance. On the contrary, $\mathit{Net}_{(40,35)}$ classifies correctly all out-of-sample waveform, including the one shown in Fig. 9. This result confirms the importance to carry out the optimization described in Section 3 to identify a network that performs accurately in spite of using an unbalanced training set.

5. Evaluation and comparison with other DL methods

The average time taken to perform the proposed analysis was calculated as 1.1 s per PQD, including the steps of image generation, classification and exporting the results. This approach is therefore promising for applications where a large number of dataset is available, such as continuous monitoring, or real-time analysis. The model may need to be re-trained at intervals as changes in power system configuration, ageing of equipment or installation of new equipment may introduce novel PQDs. The need for re-training may be suggested by the decrease in the accuracy of the predictions for new data, and this process

Table 3
Average time in seconds from varying max epochs and batch size over five runs.

Batch Size	Max Epochs								
	5	10	15	20	25	30	35	40	45
10	139.2	265.4	390.7	522.1	665.1	790.9	928.3	1028.7	1136.4
20	78.1	146.0	213.7	285.6	357.8	419.2	493.3	553.9	618.1
30	56.6	105.0	153.2	203.6	255.7	298.2	351.6	396.8	440.0
40	48.5	86.4	124.7	166.4	206.6	241.4	288.0	318.0	352.1
50	53.7	97.0	139.1	185.5	230.5	273.5	320.9	358.8	400.4

Table 4
Comparison of prediction accuracy for out-of-sample data using five different networks.

Network (batch size, max epochs)	*Net*(10, 45)	*Net*(30, 10)	*Net*(30, 40)	*Net*(40, 35)	*Net*(30, 10)
Accuracy	94.8%	85.0%	81.7%	**99.3%**	83.7%

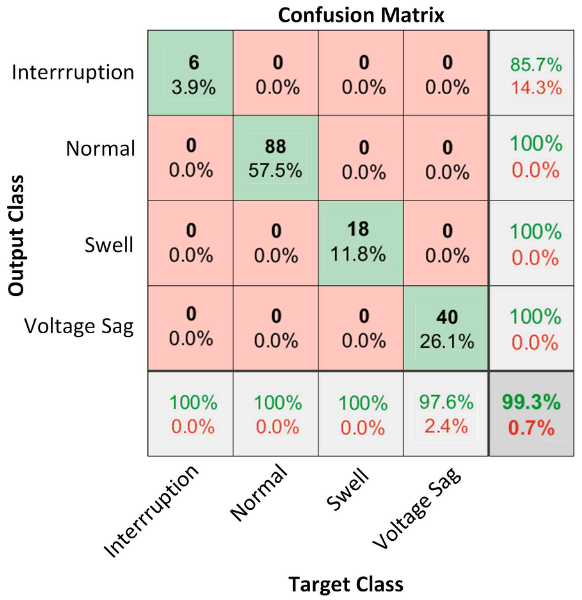


Fig. 7. Confusion matrix for the network model *Net*(40,35), using out-of-sample data.

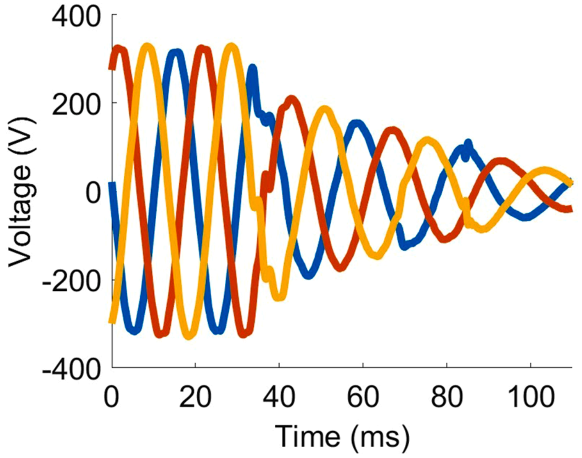


Fig. 8. Mismatch 1 - Voltage sag (PQube) vs interruption *Net*(40,35).

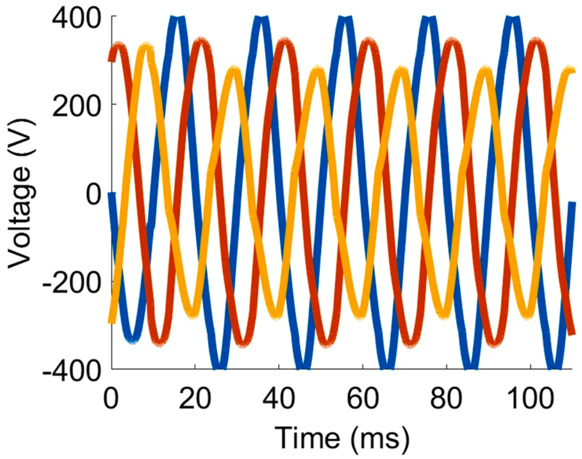


Fig. 9. Mismatch 2 - Voltage swell (synthetic data) vs voltage sag according to *Net*(30,40).

can be compared to calibration of traditional power quality meters.

The sampling time for the raw data may be relatively low, and differ between datasets. In this project, the sampling time varied between 128 and 256 samples per cycle, and was maintained when generating the waveforms used for training and testing. In addition to the results shown in the previous sections, out-of-sample analysis was carried out for 65 voltage sags where the sampling rate was reduced to 16 samples per cycle. The results showed a 100% accuracy with an average prediction speed equal to 1.02 s. While reducing the number of samples per cycle does not lead to a significant decrease of the processing time, it allows for smaller storage requirements and still results in a high accuracy.

The literature review shows only a scarce amount of results using DL for PQDs classifications. Table 5 shows a comparison between these approaches. The sampling rate of the proposed method is significantly lower than all the others. The results presented in [16] show that in general accuracy decreases when using real measurements. [15] shows high accuracy when raw data are used, however, due to lack of details, this method cannot be duplicated. The results in [23] show high accuracy, however, 30% of the test data is used for classification. Additionally, because the network is trained from scratch, the training time is approximately 8 s/image, which is significantly larger than the one obtained with the proposed methodology.

Table 5
Comparison of various DL methods proposed in the literature.

Reference	Sample frequency	Synthetic Data	Accuracy	Raw Data	Accuracy
Proposed	0.8–1.28 kHz	–	–	Y	99.8%
[13]	3.2 kHz	Y	99.96%	–	–
[15]	25.6 kHz	–	–	Y	94% - 100%
[16]	–	Y	98.4%	Y	91.9%
[18]	6.4 kHz	Y	86.9–100%	–	–
[22]	–	Y	99.75%	Y	–
[23]	7.6 kHz	Y	99.8%	Y	98.8%

In general, the use of images for PQD classification poses some concerns in terms of processing time, storage requirements, and training time. The discussion carried out above and the comparison with other methods shows that the proposed methodology mitigates all concerns, due to the effective deployment of DTL, and the use of low-resolution images.

6. Conclusions

This paper proposes a novel framework to build a PQDs classifier based on the modification of existing CNNs using DTL. Real world power quality data available from online repositories were collected, and high-resolution images were generated to fine tune an existing pre-trained network (GoogleNet) to provide an accurate classification. A max epoch equal to 35 and batch size equal to 40 were chosen following hyperparameter optimization.

The novelty of the proposed approach consists in the use of DTL to develop an imagebased CNN classifier with the ability to automatically detect PQDs in three-phase voltage waveforms. The validity of the method is demonstrated by using with real-world data collected from power quality monitors installed at different locations, overcoming limitations of current CNN-based PQDs classifiers, which obtain high accuracy on mostly synthetic data. The disturbances considered for this work included voltage sags, voltage swells and interruptions because these are the anomalies available in the adopted data set. Validation on out of sample data showed that, despite the small size of the training data set, the DTL model achieves high accuracy in the classification of real-word PQDs, when compared to classifiers using synthetic data in the training phase. This work leverage advances of DTL technologies in the field of image classification to pave the way towards real-world deployment of effective image-based PQDs detection.

Future work will address the identification of more complex and subtle power quality disturbances (such as harmonics and voltage fluctuations), in both three-phase and single-phase voltage recordings. Applications of this methodology to real time monitoring will also be considered.

CRediT authorship contribution statement

Grazia Todeschini: Conceptualization, Resources, Writing – review & editing, Supervision. **Karan Kheta:** Data curation, Methodology, Software, Investigation, Writing – original draft. **Cinzia Giannetti:** Conceptualization, Methodology, Software, Validation, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgments

Dr Todeschini and Dr Giannetti are supported by the UK Engineering and Physical Sciences Research Council (EPSRC) (EP/T013206/2; EP/S001387/1; EP/V061798/1). All authors would like to acknowledge the support of the IMPACT project, part-funded by the European Regional Development Fund (ERDF) via the Welsh Government. Dr Giannetti would like to acknowledge the support of AccelerateAI, part-funded by the ERDF via the Welsh Government.

References

[1] IEC, 61000-4-30: electromagnetic compatibility: testing and measurements techniques - Power quality measurement methods (2015).
[2] IEEE Standard Association, IEEE guide for voltage sag indices (2014). doi:10.1109/IEEESTD.2014.6842577.
[3] J. Arrillaga, R. Watson, Power Systems Harmonics, John Wiley & Sons, 2003.
[4] CIGRE Joint Working Group C.29, Power quality aspects of solar power. technical brochure 672, Tech. rep (2016).
[5] W. Xu, Experiences on using gapless waveform data and synchronized harmonic phasors, in: General Meeting, 2015. Technical Report.
[6] A. Bastos, S. Santoso, Universal waveshape-based disturbance detection in power quality data using similarity metrics, IEEE Trans. Pow. Del. 35 (4) (2022) 1–9.
[7] C. Lee, Y. Shen, Optimal feature selection for power-quality disturbances classification, IEEE Trans. Pow. Del. 26 (4) (2022).
[8] P. Khetarpal and M.M. Tripathi, A critical and comprehensive review on power quality disturbance detection and classification, Sustain. Comput. Informa. Syst. 28 (2020).
[9] D.A. Elvira-Ortiz, J.J. Saucedo-Dorantes, R.A. Osornio-Rios, D. Morinigo-Sotelo, J. A. Antonino-Daviu, Power quality monitoring strategy based on an optimized multidomain feature selection for the detection and classification of disturbances in wind generators, Electronics (Basel) 11 (2) (2022).
[10] N. Mohan, K.P. Soman, R. Vinayakumar, Deep power: deep learning architectures for power quality disturbances classification, in: Int. Conf. on Technological Advancements in Power and Energy (TAP Energy), 2017, pp. 1–6.
[11] L. Aguayo, G.A. Barreto, Novelty detection in time series using self-organizing neural networks: a comprehensive evaluation, Neural Process. Lett. 47 (2018).
[12] B. Li, Y. Jing, W. Xu, A generic waveform abnormality detection method for utility equipment condition monitoring, IEEE Trans. Pow. Del. (2017) 32.
[13] S. Wang, H. Chen, A novel deep learning method for the classification of power quality disturbances using deep convolutional neural network, Appl. Energy 235 (2019) 1126–1140. November 2018.
[14] H. Liu, F. Hussain, Y. Shen, S. Arif, A. Nazir, M. Abubakar, Complex power quality disturbances classification via curvelet transform and deep learning, Electr. Power Syst. Res. 163 (2018).
[15] E. Balouji, O. Salor, Classification of power quality events using deep learning on event images, in: 3rd Int. Conf. on Pattern An. and Image An., IPRIA, 2017.
[16] N. Mohan, K.P. Soman, R. Vinayakumar, Deep power: deep learning architectures for power quality disturbances classification, in: IEEE Int. Conf. Techn. Advancements in Pow. and En., TAP Energy, 2017.
[17] S. Omrana, E.M.F. El Houby, Prediction of electrical power disturbances using machine learning techniques, J. Ambient Intell. Human. Comput. (2019).
[18] B.Y. Husodo, K. Ramli, E. Ihsanto, T.S. Gunawan, Real-time power quality disturbance classification using convolutional neural networks. Recent Trends in Mechatronics Towards Industry 4.0, Springer Singapore, Singapore, 2022, pp. 715–724.
[19] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT press, 2016.
[20] A.E. Essien, C. Giannetti, A deep learning framework for univariate time series prediction using convolutional lstm stacked autoencoders, in: IEEE Int. Symp. INnovations in Intelligent SysTems and Applications (INISTA), 2019, pp. 1–6.
[21] A.E. Essien, C. Giannetti, A deep learning model for smart manufacturing using convolutional LSTM neural network autoencoders, in: IEEE Trans. Ind. Informatics 16, 2020.
[22] J. Ma, J. Zhang, L. Xiao, K. Chen, J. Wu, Classification of power quality disturbances via deep learning, IETE Tech. Rev. (4) (2017).
[23] K. Cai, et al., Classifying power quality disturbances based on phase space reconstruction and a convolutional neural network, Appl. Sci. 9 (18) (2022).
[24] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: a survey, Knowl. Based Syst. 80 (2015) 14–23.
[25] S. Ruder, P. Ghaffari and J.G. Breslin, Knowledge adaptation: teaching to adapt (2017). arXiv:1702.02052.
[26] C. Giannetti, A. Essien, Towards scalable and reusable predictive models for cyber twins in manufacturing systems, J. Intell. Manuf. (2021).
[27] K. Weiss, T.M. Khoshgoftaar, D. Wang, A Survey of Transfer Learning, 3, Springer. International Publishing, 2016, https://doi.org/10.1186/s40537-016-0043-6.
[28] J. Flynn, C. Giannetti, Using convolutional neural networks to map houses suitable for electric vehicle home charging, AI 2 (1) (2021) 135–149.
[29] A. Canziani, A. Paszke and E. Culurciello, An analysis of deep neural network models for practical applications (2016) 1–7arXiv:1605.07678.
[30] A. Elhassouny, F. Smarandache, Trends in deep convolutional neural Networks architectures: a review, in: Proc. of 2019 Int. Conf. Comp. Science and Ren. En., ICCSRE, 2019, pp. 1–8, https://doi.org/10.1109/ICCSRE.2019.8807741.
[31] Image Net Project, Image-net.org. 2022.
[32] A. Krizhevsky, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Neural Information Proces. Syst, 2012.

[33] F.N. Iandola, M.W. Moskewicz, K. Ashraf, S. Han, W.J. Dally, K. Keutzer, Squeezenet: alexnet-level accuracy with 50x fewer parameters and ¡1 mb model size, CoRR (2016) abs/1602.07360.

[34] C. Szegedy, et al., Going deeper with convolutions, in: IEEE Conf. on Comp. Vision and Pattern Recognition, 2015.

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conf. on Comp. Vision and Pattern Recogn, 2016.

[36] Power Standards Lab, Live world map of power quality, map.pqube.com/. 2022.

[37] C. O'Donovan, C. Giannetti, G. Todeschini, A novel deep learning power quality disturbance classification method using autoencoders, in: Int. Conf. on Agents and Artificial Intelligence, ICAART, 2021.

[38] W.A. Gardner, Learning characteristics of stochastic-gradient-descent algorithms: a general study, analysis, and critique, Signal Process. 6 (1984).