




Comparative Study of Event Prediction in Power Grids using Supervised Machine Learning Methods


Kristian Wang Høiem 
Dept. of Energy Systems
SINTEF Energy Research
Trondheim, Norway
kristian.w.hoiem@sintef.no

Vemund Santi
Dept. of Computer Science
NTNU
Trondheim, Norway
vemund@santi.no

Bendik Nybakk Torsæter 
Dept. of Energy Systems
SINTEF Energy Research
Trondheim, Norway
bendik.torsater@sintef.no

Helge Langseth
Dept. of Computer Science
NTNU
Trondheim, Norway
helge.langseth@ntnu.no

Christian André Andresen 
Dept. of Energy Systems
SINTEF Energy Research
Trondheim, Norway
christian.andresen@sintef.no

Gjert H. Rosenlund 
Dept. of Energy Systems
SINTEF Energy Research
Trondheim, Norway
gjert.rosenlund@sintef.no

Abstract—There is a growing interest in applying machine learning methods on large amounts of data to solve complex problems, such as prediction of events and disturbances in the power system. This paper is a comparative study of the predictive performance of state-of-the-art supervised machine learning methods. The event prediction models are trained and validated using high-resolution power quality data from measuring instruments in the Norwegian power grid. The recorded event categories in the study were voltage dips, ground faults, rapid voltage changes and interruptions. Out of the tested machine learning methods, the Random Forest models indicated a better prediction performance, with an accuracy of 0.602. The results also indicated that rapid voltage changes (accuracy = 0.710) and voltage dips (accuracy = 0.601) are easiest to predict among the tested power quality events.

Index Terms—Machine Learning, Power system, Power Quality Analysis, Fault Prediction, Predict Faults, Predictive Models, Power Quality Measurements

NOMENCLATURE

AHA	Automatic Event Analysis
ANN	Artificial neural network
AUC	Area under the curve
DDG	Dynamic dataset generator
DSO	Distribution system operator
MCC	Matthews correlation coefficient
ML	Machine learning
PQ	Power quality
PQA	Power quality analysers
RF	Random forest
RMS	Root mean square
RNN	Recurrent neural network
ROC	Receiver operating characteristics
RVC	Rapid voltage change
SVM	Support vector machines

The authors would like to thank the Research Council of Norway and industry partners for the support in writing this paper under project 268193/E20 EarlyWarn.

TSO	Transmission system operator
V	Voltage
V12	Line voltage between phase 1 and 2

I. INTRODUCTION

A. Motivation and Background

In a world that is increasingly dependent on electricity, providing a stable power delivery to end-users is of utmost importance. Trends such as the introduction of variable renewable energy sources, changing consumer behavior and loss of rotational inertia leads to a more transient grid operation. In order to maintain a high level of security of supply, a development in the tools used for power systems operations is needed. One such potential tool is the utilisation of machine learning techniques to predict unwanted events in the power grid. Such tools may give sufficient warning horizon for mitigation actions to be taken, thus avoiding the potential detrimental consequences of the unwanted event.

B. Relevant literature

Previous studies conducted in this research area have mostly been focusing on classification of disturbances in the electrical power grid rather than prediction, such as [1]–[4]. Despite this, methods used in classification problems are assumed applicable for prediction problems as well. Some studies that focus on fault prediction have been found in research literature, such as [5]–[8]. However, none of them include a comparison study of the predictive performance of the state-of-the-art machine learning (ML) algorithms. In addition, none of them are using high-resolution power quality (PQ) data from power quality analysers (PQAs) to train and validate the ML model.

C. Contributions and Organisation

The research gap presented in the previous section is addressed in this paper. Thus, in this study, ML methods used for power system event prediction are examined. Specifically, multiple supervised machine learning models that take real

high-resolution voltage measurement data from PQAs as input are developed, and their predictive performance is compared.

There are some central questions addressed in this paper, and they build on top of each other. Firstly, the paper investigates which attributes (often called features) from power quality (PQ) measurements are most suited for predicting events in the power grid. Secondly, different machine-learning methods are compared for their relative predictive performance based on these attributes. Lastly, the paper investigates the predictive capability of the best-performing method on different types of events in the dataset. It is the hope of the authors that the work in this paper will inspire other teams to utilise these methods in their work to bring forth a more robust power operation for the future.

II. DATA

The authors have been granted conditional access to power quality data for the majority of the Norwegian distribution grid by a group of distribution system operators (DSO) and the Norwegian transmission system operator (TSO). The overall database spans the period from January 2009 to early March 2020, and the nominal line voltages at the 49 locations where the measuring instruments were installed varied from 10 to 420 kV. A total of roughly 270 years of PQ time series has been collected from the 49 measurement nodes. This gives an average of 5-6 years of historical data from each node, although the number of years of available data varied significantly from node to node. Of the 49 measurement nodes, 15 were used in this study.

1) *Data sources:* This paper exploits data from power quality analysers in an attempt to quantify the probability for events occurring in the distribution and transmission grids. The data analysed originates from Elspec PQAs, which continuously sample voltage and current waveform at a sampling frequency of up to 50 kHz. Data is being compressed before stored in a centralised database for analysis. The operational PQA devices collect and compress many events and disturbances each year, and some nodes have been online for over 15 years. To properly manage and extract value from such a massive dataset, two software packages have been developed.

The Automatic Event Analysis (AHA) program is used to automatically detect and report lists of events and disturbances in the recorded time series [9]. The tool can identify and classify interruptions, ground faults, voltage dips and rapid voltage changes (RVC). These are annotated with event type, start time and end time for each event. A majority of applications and algorithms within machine learning requires labeled datasets for exploitation of patterns and signals in the data, and these labels are extracted using the AHA software. To detect explanatory signals for predictive purposes, one also need the power quality data leading up to the error. For this purpose, the Dynamic Dataset Generator (DDG) software has been developed [10], [11]. This software takes an event as input and extracts time series of desired resolution and duration for user-specified parameters, such as voltages and current waveforms, harmonics and RMS values, as output.

In combination, the AHA and DDG software provide datasets of labelled time series to be used for training and testing data-driven predictive methods. The method for generation of labelled datasets is illustrated in Fig. 1.

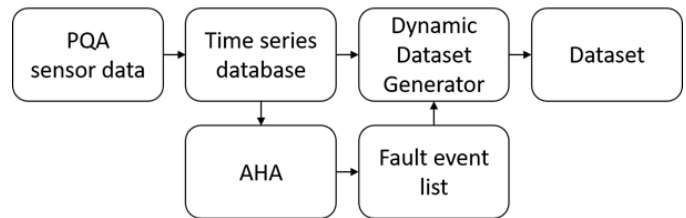


Fig. 1. Method for generation of labelled datasets

2) *Data pre-processing:* When presented with the data, the most appealing approach is to leverage algorithms that are tailored for time series forecasting, and use the raw data as explanatory variables. This approach is investigated in [9]–[12].

Various events, such as voltage dip, ground fault, power interruption, and RVC, were detected from historical time series data. It was assured that the given time before the event did not include any other event. This meant that recurring or sequential events did not give indications of artificially high predictive ability. In total, 2285 events were detected. Of these, 1124 were voltage dips, 1008 were ground faults, 76 power interruptions, and 116 RVC. A balanced time series dataset of all the events and non-events were generated with a duration of 50 minutes, having a resolution of one sample per second.

This paper presents multiple feature engineering methods that aggregate time series of harmonic frequencies in high-resolution voltage data. The aggregation method used were mean-, minimum-, maximum- and standard deviation-values. The resulting dataset contained samples of 386 features each, without any temporal dimension. This was further reduced to 162 features per sample by ranking the most important features. The features were combined with Support Vector Machines, Random Forests and multiple Neural Network architectures, with the aim of predicting events with a 10-minute prediction horizon. The proposed event prediction model, which includes dataset generation, data pre-processing and training/validation of the machine learning model, is illustrated in Fig. 2.

It is important to arrange the data appropriately for any data-driven methods in order to enable the methods to be trained well by the parameters that are presented to them. This needs to be seen in connection with the meta-parameters chosen for each method. Each method is described further in Section III, where such data-arranging is outlined. A variation over different approaches has been made during the development of this paper.

III. METHODS

Although the literature on data-driven methods is vast, and these methods are starting to become commonplace in the industry, a brief introduction to the major methods used for the comparative work in this paper is given. The

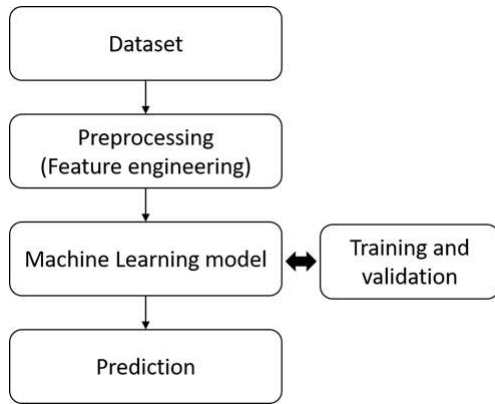


Fig. 2. Illustration of the proposed event prediction model

performance metrics used in the paper are also discussed, as they are crucial to understanding the relative performance of the methods utilised.

A. Machine learning methods

This section introduces the different machine learning methods used in this paper. All models were developed in Python 3.6 using the packages Scikit-Learn and Keras.

1) *SVM*: Support Vector Machines (SVM) is a machine learning method that has been widely used in classification problems. It operates by finding the hyperplane in a multi-dimensional space that separates the given classes with the largest margin between the hyperplane and the nearest sample [13].

2) *Random Forest*: Random Forest (RF) is an extension to Decision Trees, creating an ensemble of Decision Trees on different subsets of the total dataset and aggregating their outputs to get a more robust prediction model. The Decision Tree is a machine learning method for classification problems that works by inferring rules for splitting the dataset into multiple subsets based on the properties of the data [14].

3) *Feed-forward Neural Network*: Feed-forward Neural Networks, also known as artificial neural networks (ANN), consist of layers of nodes and weights combined as a weighted sum, with an added non-linear function applied element-wise for each output of a layer. The so called hidden layers defines the mapping that transforms the input to the output. To optimise the performance of a neural network, a set of labeled samples is run through the network to generate predictions. The output of the network is then compared to the labels of the samples, which is regarded as the ground-truth label of each sample. A loss measure can then be calculated by the use of a comparison function to calculate the total error of the network for the given samples. By using the gradient descent algorithm, the weights in the network are adjusted to reduce the total loss. This is done by propagating the error layer-by-layer backwards through the network and calculating the individual contribution to the loss for each weight in the network [15].

4) *Recurrent Neural Network*: One of the most popular ways to deal with sequential data is Recurrent Neural Networks (RNN). RNNs are neural networks that not only feed the output values forward to the next layer, but also use the values as an input into itself in the next time step. There are multiple implementations of RNNs, among which the Long Short-Term Memory (LSTM) is particular popular. RNNs have internal memory, making it capable of remembering previously seen inputs from a sequence. This helps increasing the accuracy of prediction, as RNNs can utilise the entire input-sequence to generate its output [15].

B. Metrics

There are multiple ways to compare the performance between machine learning models. It is important to choose a metric or set of metrics that measures the predictive ability of the evaluated methods realistically and representatively for the intended use of the methods. This requires insight into the domain where the predictive method is intended to be used.

1) *Matthews Correlation Coefficient*: Matthews Correlation Coefficient (MCC) is a measure of model performance in a binary classification problem [16]. It takes into account the true positive rate, true negative rate, false positive rate and false negative rate, and is regarded as a robust measure for classification problems with an unbalanced dataset [17]. The MCC takes on values in the interval $[-1, 1]$, where a value of -1 means there is perfect negative correlation between the input variable and the dependent output variable, a value of 0 means the two variables are uncorrelated, and a value of 1 means there is a perfect correlation between the two.

2) *Receiver Operating Characteristic curves*: A Receiver Operating Characteristic (ROC) curve is a plot showing the ability of an estimator to discriminate between true positive and false positive outcomes by changing the threshold needed for classifying a sample as positive [18].

The Area Under the Curve (AUC) is the total area under the ROC curve, and is often used as a way to describe and compare ROC curves without drawing the actual curves for comparison. An AUC value of 1 means the model is able to perfectly classify all samples.

IV. METHOD APPLICATION AND RESULTS

In order to answer the questions outlined in the introduction, four tests have been performed. They are briefly summarised below in the sequence they were conducted.

A. Attribute selection

The first test investigated the important features of the dataset created by Feature extraction. This test used a Random Forest model to rank the feature importance of each individual feature. The identified features as illustrated in Table I provided a foundation for model creation in the later tests.

The time step features were selected from the harmonic components of the three-phase voltage signal. The attributes used for each of the line- and phase voltages were described in Section II.

From the result, the aggregation methods of maximum and standard deviation have an overall higher importance, but there is no clear difference between phase- and line voltage harmonic components. All harmonic components up to the 256th component were investigated. Out of the 120 most important features, all were aggregates of harmonic components below the 16th order. It was therefore concluded that no harmonic component above the 15th order would add significant predictive capability to the models.

B. Benchmarking of the models

In this test, the machine learning models were benchmarked against each other. The models' hyperparameters were tuned using a validation set before testing the performance on a test set.

Table II presents the results for all models and the hyperparameters used. The results show that the Random Forest model performs slightly better than the rest, while the recurrent neural network has the lowest overall performance. Figure 3 shows the ROC curves of the performance of the benchmarked machine learning models. Only some variations in the ROC curve can be observed between the models, however the RNN curve is found to be markedly worse than the other.

C. Predictive capability relative to event type

The best-performing model from the benchmarking test was selected to investigate whether a model can give an indication on the predictive capabilities on the different event types. The Random Forest model was trained on the entire training dataset, including all event types as in the benchmarking test, and then the performance was tested on test sets for each event type. From this, the generally trained model's ability to predict a given type of event was compared using the accuracy of prediction for each type of event. Table III summarises the results.

D. Event specific model training

The Random Forest model was trained on a dataset made up of events of the same category, and predicted on a test set of the same event category. This was done in order to investigate the effect of the predictive performance of the models when specifically trained on a subset of the data containing only

TABLE I

THE 10 MOST IMPORTANT FEATURES IN THE DATASET, RANKED BY A RANDOM FOREST

Rank	Harmonic number	Time step feature	Line / Phase	Aggregation method	Importance
1	Harmonic 10	Minimum	V1	Maximum	0.00772
2	Harmonic 14	Minimum	V3	Maximum	0.00692
3	Harmonic 10	Minimum	V31	Maximum	0.00686
4	Harmonic 10	Minimum	V2	Maximum	0.00656
5	Harmonic 10	Minimum	V3	Maximum	0.00631
6	Harmonic 4	Minimum	V31	Standard deviation	0.00611
7	Harmonic 14	Minimum	V1	Maximum	0.00607
8	Harmonic 10	Minimum	V12	Maximum	0.00600
9	Harmonic 14	Minimum	V2	Standard deviation	0.00599
10	Harmonic 14	Minimum	V1	Standard deviation	0.00592

TABLE II

ACCURACY, MATTHEWS CORRELATION COEFFICIENT AND AREA UNDER THE CURVE SCORES FOR THE MODELS TESTED IN SECTION IV-A, DISPLAYED TOGETHER WITH HYPERPARAMETERS USED.

Model	Hyperparameters	Accuracy	MCC	AUC
SVM	Kernel = Radial basis function Penalty multiplier = 10	0.572	0.185	0.627
RF	Max depth = 15 Number of trees = 100 split criterion = entropy	0.602	0.213	0.642
ANN	First layer size = 128 Second layer size = 64 Third layer size = 64 Optimizer = Adam Loss function = Cross-entropy Batch size = 128 Number of epochs = 300	0.600	0.213	0.618
RNN	First GRU state size = 64 Second GRU state size = 32 Attention layer size = 32 Fully-connected layer size = 32 L2-regularization = 0.01 Optimizer = Adam Loss function = Cross-entropy Batch size = 128 Number of epochs = 300	0.585	0.175	0.586

TABLE III

THE ACCURACY OF THE RANDOM FOREST MODEL SELECTED BASED ON THE BENCHMARKING WITHIN EACH EVENT CATEGORY IN SECTION IV-C.

Sample category	Accuracy
Voltage dip	0.580
Ground fault	0.671
RVC	0.526
Interruption	0.444
No events	0.588

TABLE IV

ACCURACY, MATTHEWS CORRELATION COEFFICIENT (MCC) AND AREA UNDER THE CURVE (AUC) SCORES ACHIEVED WHEN TRAINING AND PREDICTING ON SINGLE CLASSES OF EVENTS.

Event type	Accuracy	MCC	AUC
Voltage dip	0.601	0.242	0.690
Ground fault	0.589	0.188	0.630
RVC	0.710	0.393	0.782
Interruption	0.500	0.0	0.531

one type of events (in addition to the non-event data). The test investigated whether machine learning models focusing on a single type of event are better at predicting these specific events, or if the type of event have little impact on overall performance.

Table IV presents the results from predicting single class events. From this, samples with RVC and voltage dip appear to be easier to predict, while samples with power interruptions are more challenging. The ROC curves in Figure 4 show the model's sensitivity to the setting of the threshold. Note that by separating the overall dataset into subsets according to event type results in fewer samples to train and test on. This gives ROC curves that have large jumps, as can be seen for the RVC and power interruption curves.

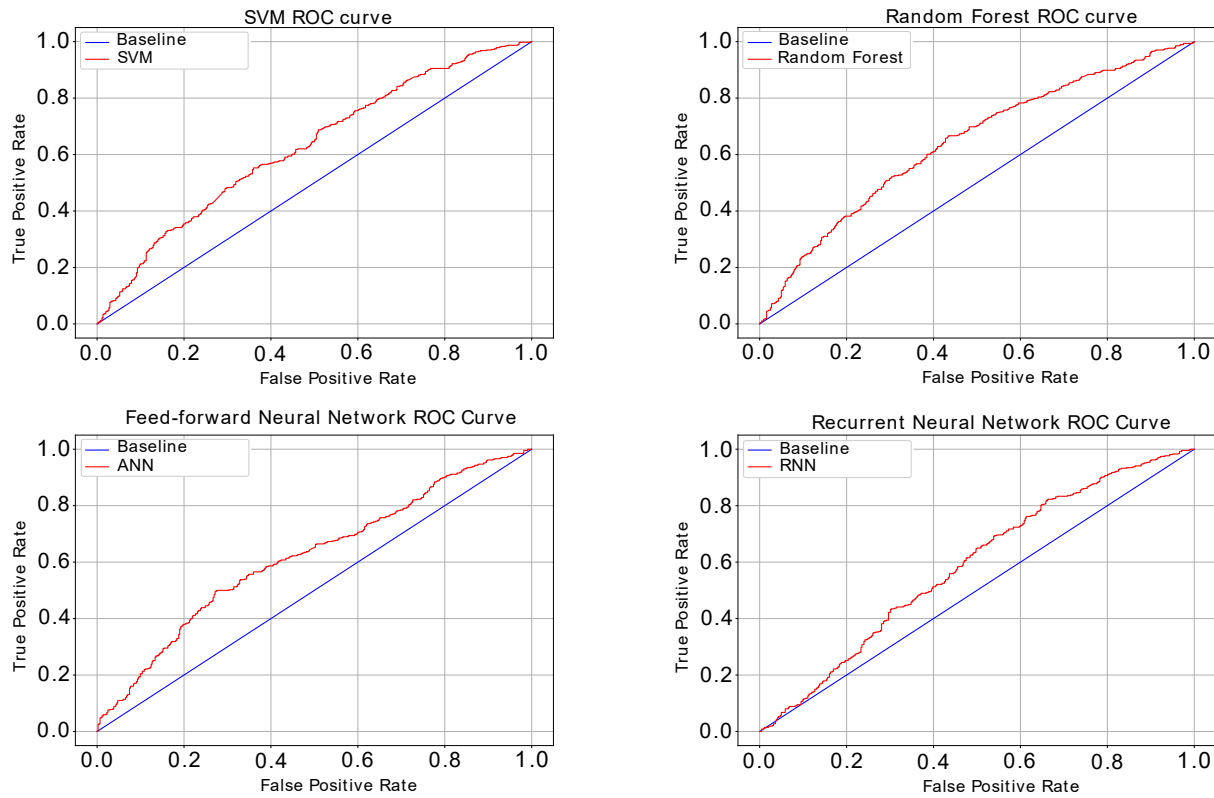


Fig. 3. ROC curves of (top left) SVM, (top right) Random Forest, (bottom left) Feed-forward Neural Network, and (bottom right) Recurrent Neural Network from the benchmarking tests.

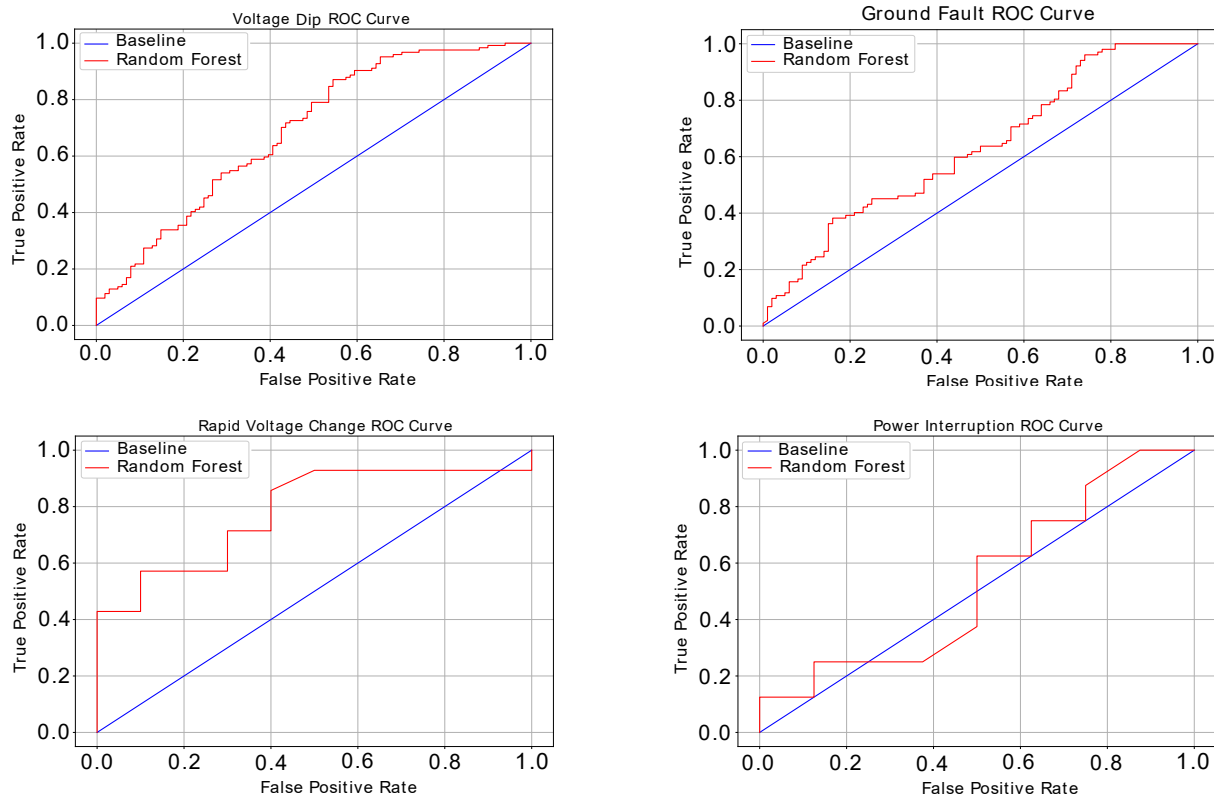


Fig. 4. ROC curves for a Random forest trained and tested on (top left) voltage dips, (top right) ground faults, (bottom left) rapid voltage changes, and (bottom right) power interruptions. These are results from models trained on subsets of the data containing specific event types and not on the combined dataset of all event types.

E. Discussion

As noted above, the sub-setting of the data has caused some subsets to be rather sparse for applying these methods. It is therefore plausible that the high AUC score for the event-specific Random Forest model for RVC in Table IV is artificially high. This is supported by the lack of smoothness in the ROC curve for this model in Figure 4. It is, however, interesting to see that the model for voltage dip and for ground fault show a small but significant vertical initial step in the ROC curve in the event-specifically trained models compared to the models trained on the general dataset. This indicates that these models are capable of predicting a small portion of these events with high confidence and a low level of false positive. This is encouraging for further work.

The fact that these tests have only utilised balanced datasets must be further investigated by the application of the models on more realistic unbalanced datasets. In such an application, the importance of a low level of false positives will become pronounced.

The application of the models are made more challenging by the fact that all events are not preceded by any other event. In an operational setting, events will often cluster in time. Thus, the occurrence of some events, such as ground faults, can be used as a parameter for predicting other events, such as power interruptions. As a consequence, the conditions under which these models have been tested may be too strict to realistically judge their true predictive capabilities in an operational setting.

V. CONCLUSION AND FURTHER WORK

In this paper, a comparison of fault prediction models based on different supervised machine learning methods was conducted. The machine learning models were trained using high-resolution power quality measurements from the Norwegian power grid. The predictive models were trained to predict four different event categories; namely voltage dips, ground faults, rapid voltage changes (RVC) and interruptions. The results from the comparison of the Random Forest model, the Support Vector Machine model, the Feed Forward Neural Network model and the Recurrent Neural Network model show that the Random Forest model is performing marginally better than the other models, with an accuracy of 0.602. However, the performance is so far not at a level where operational deployment would be feasible. Both the false-positive rate and the false-negative rate are too high for such a step. It is also observed that there are differences in performance between the different types of events that the models are trained on. The results indicate that rapid voltage changes, with an accuracy of 0.710, and voltage dips, with an accuracy of 0.601, are easiest to predict among the tested PQ events.

It would be beneficial to the research within this area if this work could be carried on by investigating the effects of application to unbalanced datasets. This would give a more realistic reflection of the potential usefulness of these models in an operational setting. Furthermore, applying the models to time series that contain events in the analysed time window

would also give a more realistic judgement of operational usefulness.

As with all other data-driven methods, performance is expected to improve with data volume. It is expected that having more training data would improve the performance of the models. It is also expected that the combination of PQA data with other data sources, such as weather data, power flow data or systems configuration data, would improve the predictive capability.

REFERENCES

- [1] R. Kumar, B. Singh, D. T. Shahani, A. Chandra, and K. Al-Haddad, "Recognition of Power-Quality Disturbances Using S-Transform-Based ANN Classifier and Rule-Based Decision Tree," *IEEE Transactions on Industry Applications*, 2015.
- [2] E. Balouji, I. Y. Gu, M. H. Bollen, A. Bagheri, and M. Nazari, "A LSTM-based deep learning method with application to voltage dip classification," *Proceedings of International Conference on Harmonics and Quality of Power, ICHQP*, vol. 2018-May, pp. 1–5, 2018.
- [3] K. Manivinnan, C. L. Benner, B. Don Russell, and J. A. Wischkaemper, "Automatic identification, clustering and reporting of recurrent faults in electric distribution feeders," in *2017 19th International Conference on Intelligent System Application to Power Systems, ISAP 2017*, 2017.
- [4] F. L. Grando, A. E. Lazzaretti, M. Moreto, and H. S. Lopes, "Fault Classification in Power Distribution Systems using PMU Data and Machine Learning," *2019 20th International Conference on Intelligent System Application to Power Systems, ISAP 2019*, 2019.
- [5] J. L. Viegas, S. M. Vieira, R. Melicio, H. A. Matos, and J. M. Sousa, "Prediction of events in the smart grid: Interruptions in distribution transformers," *Proceedings - 2016 IEEE International Power Electronics and Motion Control Conference, PEMC 2016*, pp. 436–441, 2016.
- [6] R. Eskandarpour and A. Khodaei, "Machine Learning Based Power Grid Outage Prediction in Response to Extreme Events," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3315–3316, 2017.
- [7] R. Fainti, M. Alamaniotis, and L. H. Tsoukalas, "Three-phase line overloading predictive monitoring utilizing artificial neural networks," in *2017 19th International Conference on Intelligent System Application to Power Systems, ISAP 2017*, 2017.
- [8] R. A. Shuvro, P. Das, M. M. Hayat, and M. Talukder, "Predicting Cascading Failures in Power Grids using Machine Learning Algorithms," *51st North American Power Symposium, NAPS 2019*, no. 1541148, pp. 0–5, 2019.
- [9] V. Hoffmann, K. Michalowska, C. A. Andresen, and B. N. Torsæter, "Incipient Fault Prediction in Power Quality Monitoring," in *25th International Conference on Electricity Distribution, CIRED 2019*, no. June, Madrid, 2019.
- [10] K. W. Høiem, "Predicting Fault Events in the Norwegian Electrical Power System using Deep Learning - A Sequential Approach," *MSc Thesis*, 2019.
- [11] V. M. Santi, "Predicting faults in power grids using machine learning methods," *MSc Thesis*, 2019.
- [12] C. A. Andresen, B. N. Torsæter, H. Haugdal, and K. Uhlen, "Fault Detection and Prediction in Smart Grids," in *9th IEEE International Workshop on Applied Measurements for Power Systems, AMPS 2018 - Proceedings*, 2018.
- [13] A. M. Andrew, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods," 2001.
- [14] A. Liaw and M. Wiener, "Classification and Regression with Random Forest," *R News*, 2002.
- [15] A. C. Ian Goodfellow, Yoshua Bengio, "The Deep Learning Book," *MIT Press*, vol. 521, no. 7553, p. 785, 2017.
- [16] D. M. W. Powers and Ailab, "Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation," *Journal of Machine Learning Technologies*, 2007.
- [17] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS ONE*, 2017.
- [18] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine," *Clinical Chemistry*, vol. 39, no. 4, pp. 561–577, 4 1993. [Online]. Available: <https://academic.oup.com/clinchem/article/39/4/561/5646806>