IX ECCOMAS Thematic Conference on Smart Structures and Materials
SMART 2019
A. Benjeddou, N. Mechbal and J.F. Deü (Eds)

SUPERVISED CLASSIFICATION WITH SCADA DATA FOR CONDITION MONITORING OF WIND TURBINES

ERVIN HOXHA[†], YOLANDA VIDAL[†] AND FRANCESC POZO[†]

[†] Control, Modeling, Identification and Applications (CoDAlab), Department of Mathematics,
Escola d'Enginyeria de Barcelona Est (EEBE), Universitat Politècnica de Catalunya (UPC),
Campus Diagonal-Besòs (CDB), Eduard Maristany, 16, 08019 Barcelona, Spain
e-mail: ervin.hoxha1990@gmail.com Web page: http://https://codalab.upc.edu

Key words: Condition Monitoring, Fault Classification, Wind Turbine, SCADA, Data Based, Machine Learning

Abstract. The reliability requirements of wind turbines have increased significantly in recent years in the search for a lower impact on the cost of energy. In addition, the trend towards larger wind turbines installed in remote locations has significantly increased the cost of repair or replacement of the component. In the wind industry, therefore, condition monitoring is crucial for maximum availability [1]. This contribution makes a review of supervised machine learning classification techniques for wind turbine condition monitoring using only SCADA data already available. That is, without installing extra sensors or costly purpose-built data sensing equipment. Although there has been extensive research into the use of machine learning techniques for wind turbine monitoring, the more recent trend in this type of literature is to focus on a specific WT sub-assembly: the bearings and planetary gearbox [2], the generator and power converter [3], the blades [4], etc. Oil debris systems can detect pitting failures but cannot detect cracking faults. Vibration based systems can detect both pitting and cracking, but most cannot determine the health of components in the planetary section. This work approaches condition monitoring of various wind turbine components (torque actuator, pitch actuator, pitch sensor, and generator speed sensor) with a unique strategy. In particular, for this purpose, a review of supervised machine learning classification techniques is performed and analyzed.

SMART 2019 Conference.

1 INTRODUCTION

Wind energy is a renewable energy, this means that as long as the sun continues to shine we will have wind to generate electricity. Wind energy is also clean and green. At the moment developed countries generate most of their energy from fossil fuels like coal, natural gas and oil. However burning fossil fuels creates green house gases that are through to contribute to the global climate change. In order to make wind energy profitable, the global wind industry has been witnessing an increase in average turbine size over the years. In this regard, offshore wind farms can increase energy production, as bigger wind turbines (WT) can be installed and they take advantage of higher and more uniform wind speeds.

The operation and maintenance of wind turbines is difficult and costly, especially for offshore wind turbines. Thus, condition monitoring is essential to reduce the energy cost. The faults must be anticipated before the break down takes place, and it is crucial to know what type of fault is, and where is it taking place.

Reference [5] reports a wide variety of faults in wind turbines. However, the latest research trend is to target only one specific parts of the wind turbine. For example, reference [6] target conditiom monitoring approaches of WTs with focus on gearbox, generator, blade, braking system, and rotor. The latest tendency in this type of literature review is to focus on a specific WT sub-assembly: the bearings and planetary gearbox [7], [8]. Reference [9] focuses on the generator and power converter, reference [10] targets the blades, etc. These type of localized strategies require almost always to install extra (an costly) sensors. However, there is already a huge amount of data from the existing sensors in the wind turbine (SCADA-Supervisory Control and Data Acquisition) that could be used for condition monitoring. The reason it is not actually used is that SCADA data is typically recorded at 10-minute intervals to reduce transmitted data bandwidth and storage. This low-frequency resolution negatively affects the diagnosis capabilities, and may hide short-lived events. On the other hand, high-resolution (but feasible) SCADA data should allow the dynamic turbine behavior to be identified with higher fidelity and thus improve detection efficiency. Following what is stated in [1], in this work a research framework is proposed that takes SCADA data with an additional high but feasible (1 s) frequency from the sensors. That is, the only requirement is to increase the frequency rate in the SCADA data from the already available sensors. Thus, this work approaches condition monitoring of various wind turbine components (torque actuator, pitch actuator, pitch sensor, and generator speed sensor) with a unique strategy and using only SCADA data.

Machine learning has been essential in the condition monitoring research arena. For example, application of machine learning method in bridge health monitoring[11], health monitoring of aeroplane structural component based on K-means clustering [12], etc . However, machine learning for condition monitoring of WT is still an incipient research area. In this work different machine learning strategies are used to monitor the condition of a 5 MW wind turbine using only SCADA data and taking into account faults in various WT components (torque actuator, pitch actuator, pitch sensor, and generator speed sensor).

This work is organized as follows. In Section 2, we describe the fault detection and classification methodologies. We discussed and analyzed the results in Section 3. In Section 4, some conclusions about the current challenges and future work are given.

2 FAULT DETECTION AND CLASSIFICATION METHODOLOGIES

2.1 Model overview

The simulated model stated in [13] is utilized in this work. It integrates a 5 MW wind turbine modelled using the FAST software (National Renewable Energy Laboratory, Golden, Colorado, USA), see [14]. The model proposes to simulate the sensors in the block diagram environment Simulink by adding signals from band-limited white noise blocks that are parameterized by noise power to the actual variables provided by the FAST software, see Table 1.

The fault scenarios, that comprehend sensors and actuators, are displayed in Table 2. The interested reader can find a comprehensive description of these faults and their importance in [15].

2.2 Data collection

A total of 260 simulations were conducted in this work: 100 with a healthy WT, and 20 simulations for each studied fault. The simulations have a duration of 400 seconds. Observe that the wind sequence is not used as a known measurement. It is also noteworthy that the data used by the fault detection strategy has a sampling period of 1 second. As noted in the introduction, following what is stated in [1], in this work a research framework is proposed that takes SCADA data with an additional high but feasible (1

Table 1: Available sensors (measured data).

Number	Sensor Type	Symbol	Unit	Noise Power
S1	Generated electrical power	$P_{e,m}$	W	10
S2	Rotor speed	$\omega_{r,m}$	rad/s	104
S3	Generator speed	$\omega_{g,m}$	rad/s	$2 \cdot 104$
S4	Generator torque	$\tau_{c,m}$	Nm	0.9
S5	Pitch angle of first blade	$\beta_{1,m}$	deg	$1.5 \cdot 103$
S6	Pitch angle of second blade	$\beta_{2,m}$	deg	$1.5 \cdot 103$
S7	Pitch angle of third blade	$\beta_{3,m}$	deg	$1. \cdot 103$
S8	Tower top fore-aft acceleration	$a_{fa,m}$	m/s^2	$5 \cdot 104$
S9	ower top side-to-side acceleration	$a_{ss,m}$	m/s^2	$5 \cdot 104$

Table 2: Fault scenarios.

Number	Fault	Type
F1	Pitch actuator (High air content in oil)	Change in system dynamics
F2	Pitch actuator (Pump wear)	Change in system dynamics
F3	Pitch actuator (Hydraulic leakage)	Change in system dynamics
F4	Generator speed sensor	Gain factor (1.2)
F5	Pitch sensor	Stuck value ($\beta_{3,m} = 5$ deg)
F6	Pitch sensor	Stuck value ($\beta_{3,m} = 10$ deg)
F7	Pitch sensor	Gain factor (1.2)
F8	Torque actuator	Offset value (2000 Nm)

second) frequency from the sensors.

2.3 Data reshape

Our goal is to minimize the detection time (T_d) and at the same time to obtain a high accuracy. T_d specifically means the time from when the fault occurs till it is detected. It is proposed to organize the available data from the simulations in samples of only three time steps (this will lead to a detection time of approximately three seconds). In reference [13] the faults detection requirements are given in terms of the sampling time. Fault 8, related to the torque actuator, requires to achieve $T_d < 3s$. This is the fault that requires the fastest detection time. The other faults have a slower dynamic. Fault 1 requires to be diagnosed in less than $8s$ and Faults 4 to 7 require a T_d less than 10 seconds. We will organize our data in 3 time steps. As mentioned before, we used only 400 seconds duration of 260 simulations, from 9 available sensors. Since we will organize the data in 3 time steps, and 400 is not divisible by 3, we will use only 399 seconds. Recall that initially the data was stored as below:

$$\begin{pmatrix} x_{1,1}^{(k)} & x_{1,2}^{(k)} & \dots & x_{1,399}^{(k)} \\ x_{2,1}^{(k)} & x_{2,2}^{(k)} & \dots & x_{2,399}^{(k)} \\ \dots & \dots & \dots & \dots \\ x_{260,1}^{(k)} & x_{260,2}^{(k)} & \dots & x_{260,399}^{(k)} \end{pmatrix} \in M_{260\times 399}^{(k)}(\mathbb{R}) \tag{1}$$

where k is linked to different sensors $k = 1,2,3...9$, so there is one matrix associated to each sensor. There are 260 simulations, therefore the matrix has 260 rows. In order to minimize the detection time, instead of using this matrix, we reshape the data in a matrix with only 3 columns ($J = 3$):

$$\begin{pmatrix} x_{1,1}^{(k)} & x_{1,2}^{(k)} & x_{1,J}^{(k)} \\ x_{1,J+1}^{(k)} & x_{1,J+2}^{(k)} & x_{1,2J}^{(k)} \\ \vdots & \vdots & \vdots \\ x_{1,400-J+1}^{(k)} & x_{1,400-J+2}^{(k)} & x_{1,399}^{(k)} \\ x_{2,1}^{(k)} & x_{2,2}^{(k)} & x_{2,J}^{(k)} \\ x_{2,J+1}^{(k)} & x_{2,J+2}^{(k)} & x_{2,2J}^{(k)} \\ \vdots & \vdots & \vdots \\ x_{2,400-J+1}^{(k)} & x_{2,400-J+2}^{(k)} & x_{2,399}^{(k)} \\ \vdots & \vdots & \vdots \\ x_{260,1}^{(k)} & x_{260,2}^{(k)} & x_{260,J}^{(k)} \\ x_{260,J+1}^{(k)} & x_{260,J+2}^{(k)} & x_{260,2J}^{(k)} \\ \vdots & \vdots & \vdots \\ x_{260,400-J+1}^{(k)} & x_{260,400-J+2}^{(k)} & x_{260,399}^{(k)} \end{pmatrix} \in \mathcal{M}_{260\cdot\frac{399}{3}\times 3}^{(k)}(\mathbb{R}), \tag{2}$$

As each sample has 3 seconds, the number of total samples is $I = 260 \cdot \frac{399}{3} = 34580$. Finally, the

matrices coming from all sensors ($k = 1, 2, ...9$) are concatenated the obtain the data matrix X as follows:

$$X = \left(\mathcal{M}^{(1)}|\mathcal{M}^{(2)}|\cdots|\mathcal{M}^{(9)}\right)$$

2.4 Preprocess

As the data comes from different sensors and has different magnitudes, it is columnwise normalized. Then, principal component analysis is used to reduce the dimensionality of the data. In this work, the number of principal components is selected based on keeping 99.98% of the variance. In particular, from a total of 27 components, 99.98% of the variance is accomplished by the first 16 components.

2.5 Supervised classifiers

In this work, two well known classifiers have been studied: k-nearest neighbour and support vector machines. It is not the purpose of this work to review these techniques, however the interested reader can find a comprehensive description of these methods in [17] and [18]. In this work, the parameters of both classifiers have been optimized using the 5-fold cross-validation technique. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data is to be split into. The choice of k is usually 5 or 10, but there is no formal rule. As k gets larger, the difference in size between the training set and the resampling subsets gets smaller. As this difference decreases, the bias of the technique becomes smaller [19]. In this paper we used 5-fold cross-validation.

3 RESULTS, ANALYSIS, AND DISCUSSION

3.1 k-NN

In order to choose the number of neighbours in k-NN, we test the overall accuracy for different numbers of neighbours. As shown in the Fig. 1, we get the highest accuracy when the number of neighbours is 1. Finally, the results for $k = 1$ and $k = 5$ are shown. Confusion matrices show us a comprehensive decomposition of the error between the true classes and predicted classes, see Figures 2 and 3 (an empty blank square means 0 %). In these matrices, each row represents the instances in a true class while each column represents the instances in a predicted class (by the classifier). In particular, the first row (and first column) is labeled as 0 and corresponds to the healthy case. The next labels (for rows and columns) correspond to each fault (from Fault 1 to Fault 8). From the confusion matrices, the following issues can be highlighted.

When the number of neighbours is $k = 1$ (Fig. 2), the overall accuracy is 95.7 %. In this case, the healthy class has a true positive rate (TPR, the percentage of correctly classified instances) higher than 99% and a false negative rate (FNR, the percentage of incorrectly classified instances) smaller than 1%. Fault 1 (related to the pitch actuator fault with high dynamics) has a TPR of 91% and an FNR of 9%. This FNR percentage was mainly obtained from 2% missing faults, 1% misclassified as Fault 3, 5, 6, 7 or 8, and 6% confusion with Fault 2, which is also a fault located in the pitch actuator. Fault 2, related to pitch actuator (pump wear), has a TPR of 89% and an FNR of 11% . It was misclassified as healthy 4% of the time, 2% of the time it was confused with Fault 1, and 5% of the time it was confused as Fault 3, 5, 6 ,7 or 8. Fault 6 (related to the pitch sensor) was the most difficult to be classified, it had a TPR of 88%, 4% was misclassified as Fault 5 (related also to pitch sensor), 3% missing fault, and 5% misclassified as Fault 1,2,3,4,7 or 8. We can see from the confussion matrix that all the Faults, except Fault 2 and Fault 6, have a TPR higher than 91%.

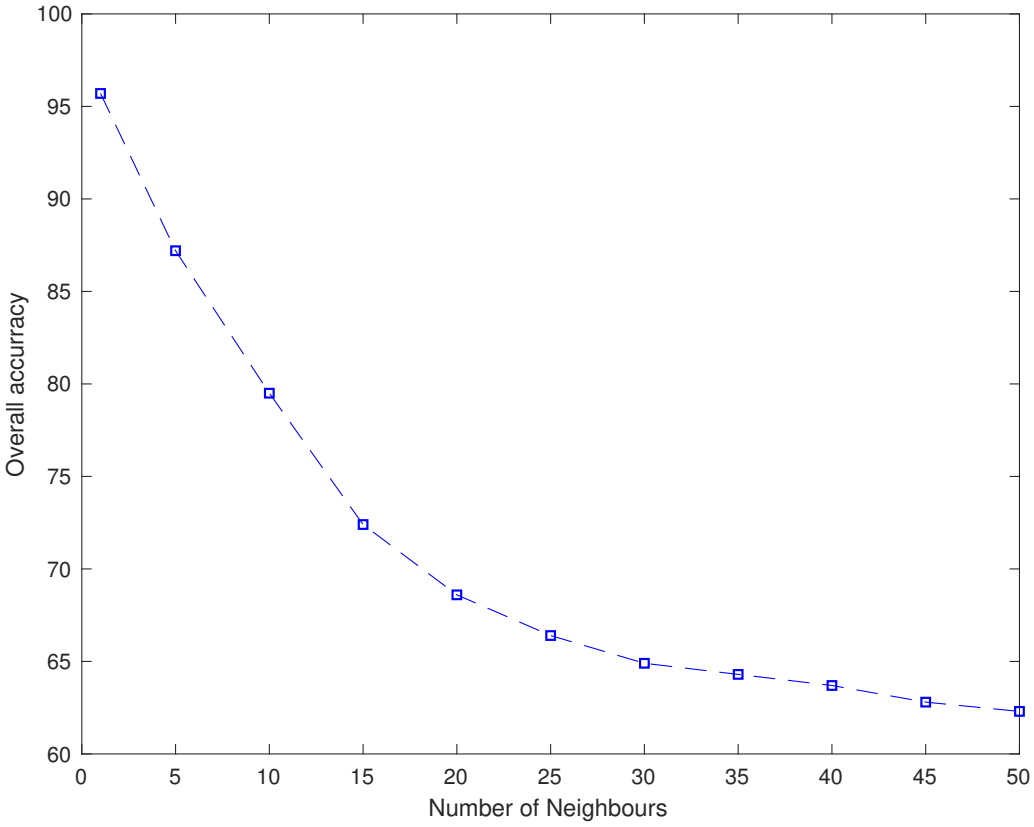

Figure 1: Number of neighbours with respect to overall accuracy.

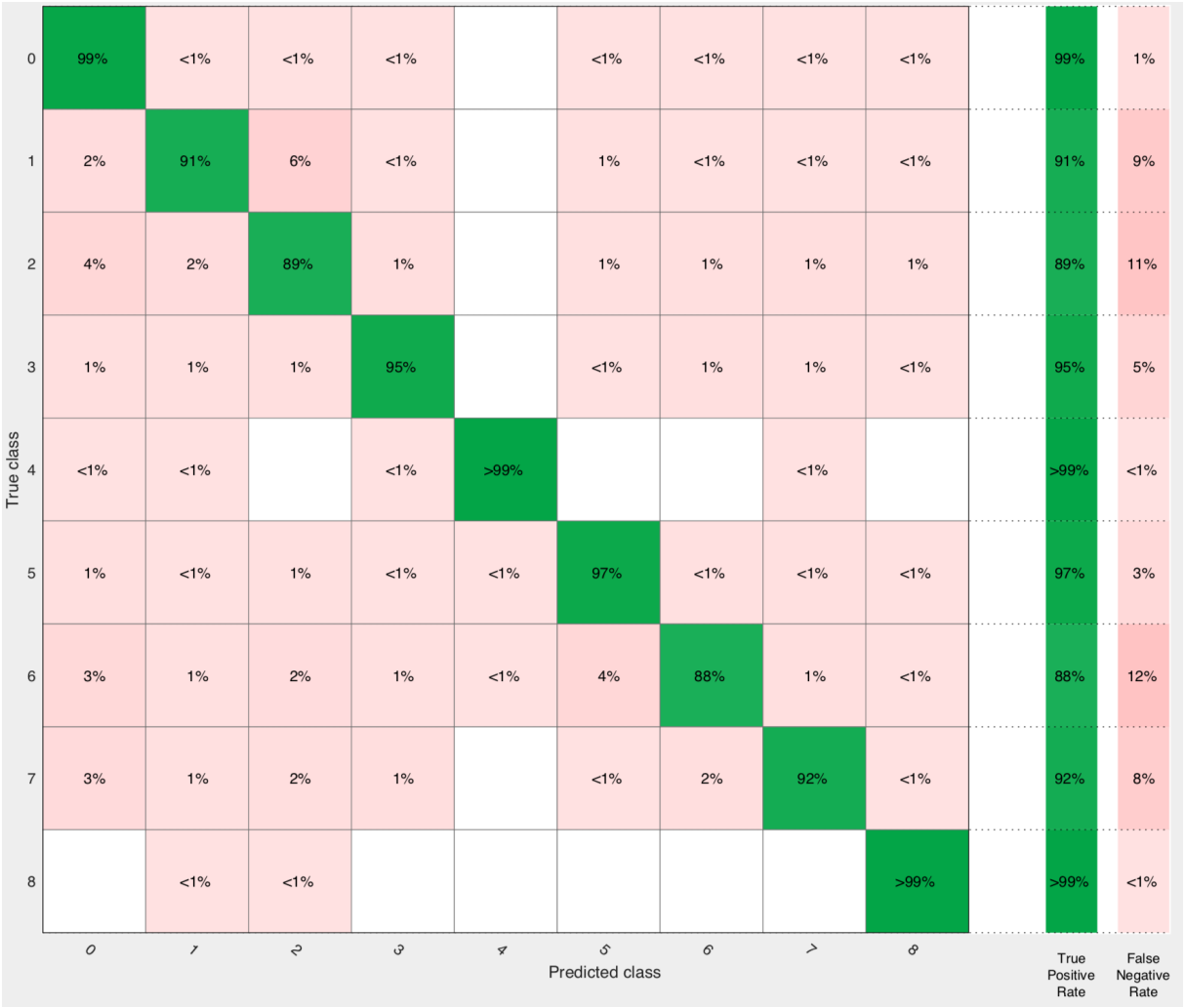

Figure 2: Confusion matrix when the number of neighbours is selected 1

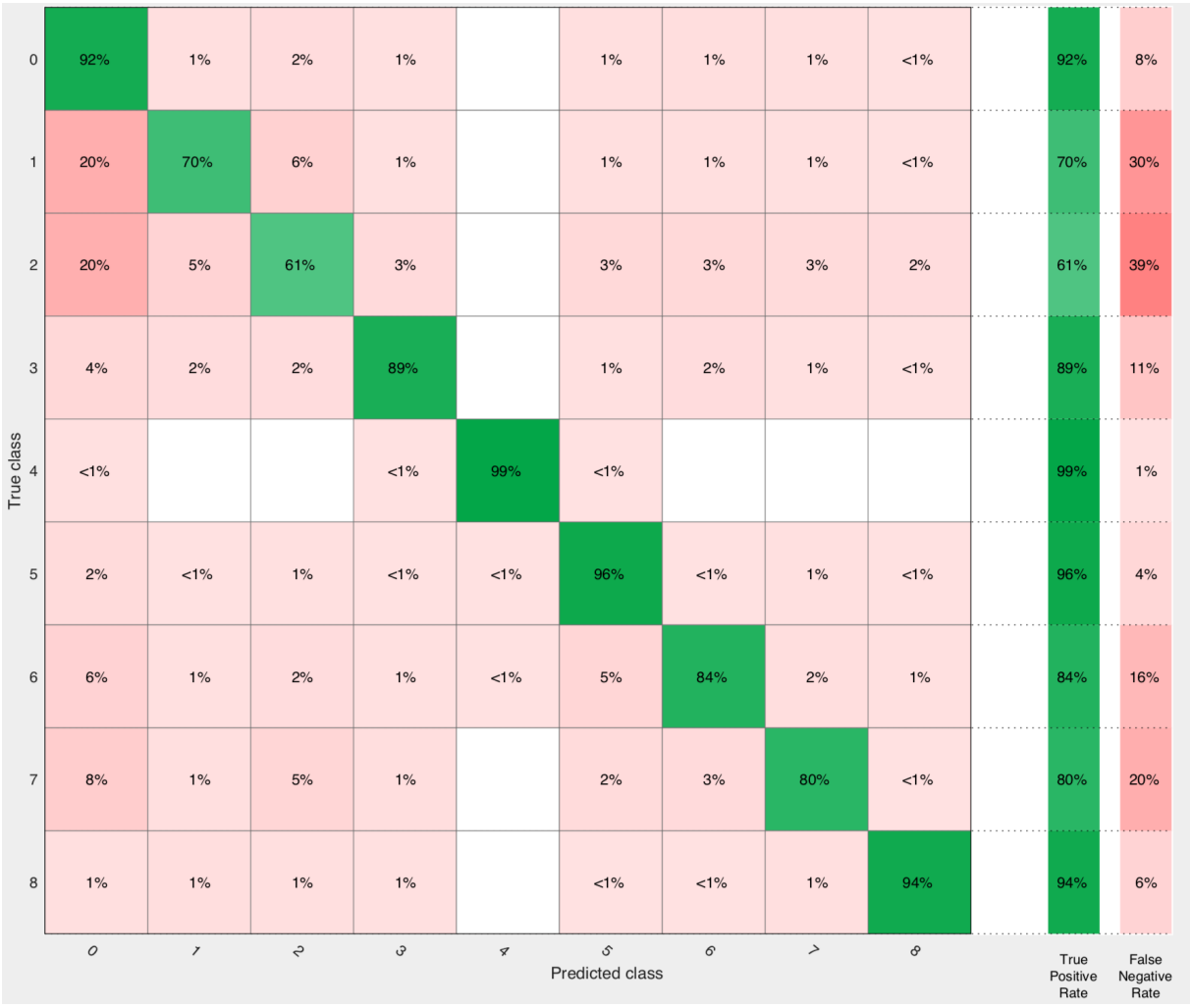

Figure 3: Confusion matrix when the number of neighbours is selected 5

When the number of neighbours is $k = 5$ (Fig. 3), the overall accuracy is 87.2 %. Fault 4 (related to the generator speed sensor) as in the previous case has a TPR of 99%. All the other classes decrease their TPR. The most affected classes are Fault 1 and Fault 2, in particular, Fault 1 decreases its TPR to 70% and Fault 2 to 61%.

3.2 SVM

The 5-fold cross-validation technique is used to select the following parameters: box constraint and kernel scale of the Gaussian kernel, see [20]. The results are presented in Fig. . The highest accuracy is 95% and it is achieved when the kernel scale is equal to $1/4$ and the box constraint is 75 or 100.

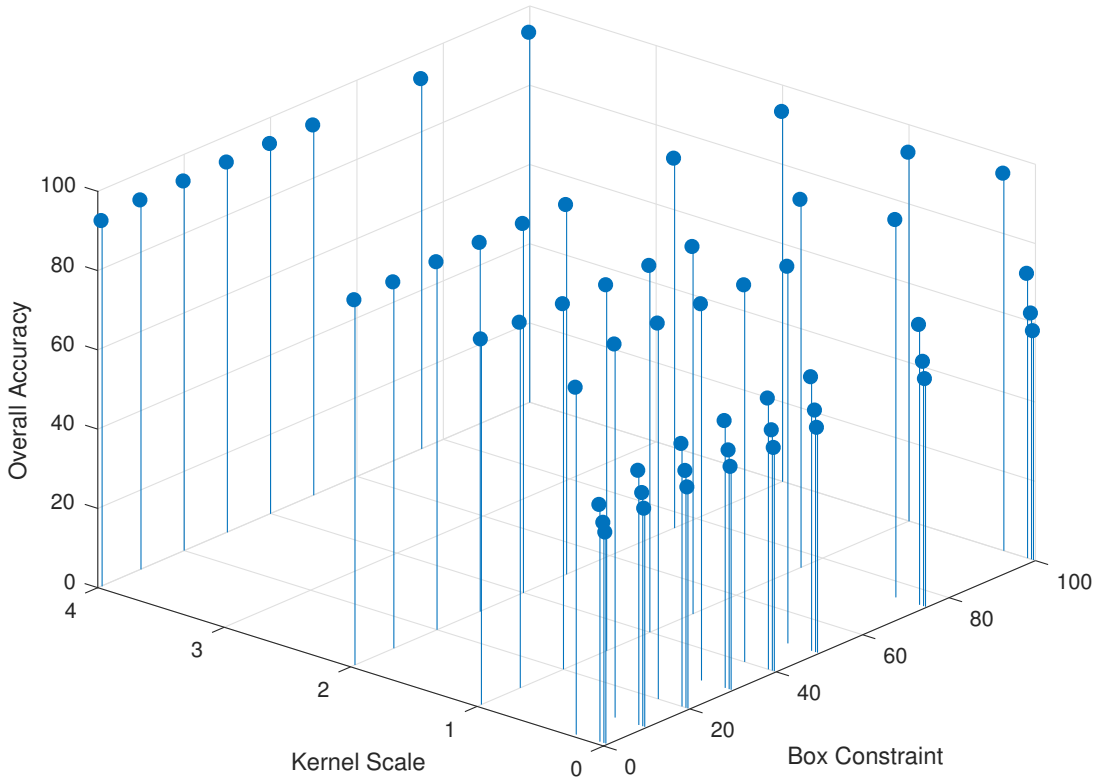

Figure 4: Box constraint value and kernel scale with respect to overall accuracy

Fig. 5 shows the confusion matrix when the box constraint is equal to 75 and the kernels scale is 1/4. In this case Fault 1 (the most difficult to classify) has a TPR of 77%, 17% of the times is misclassified as no fault and 6% is misclassified as Fault 2. Fault 6 has a TPR of 89% and an FNR of 11%. The FNR is mainly obtained from 4% missing fault, 2% misclassified as Fault 2, 3% as Fault 5 and remaining percentage 2% is a misclassification to Fault 1,3,4 or 7. The rest of the faults has a TPR above 90%. Note that Fault 8 has a TPR of 100%.

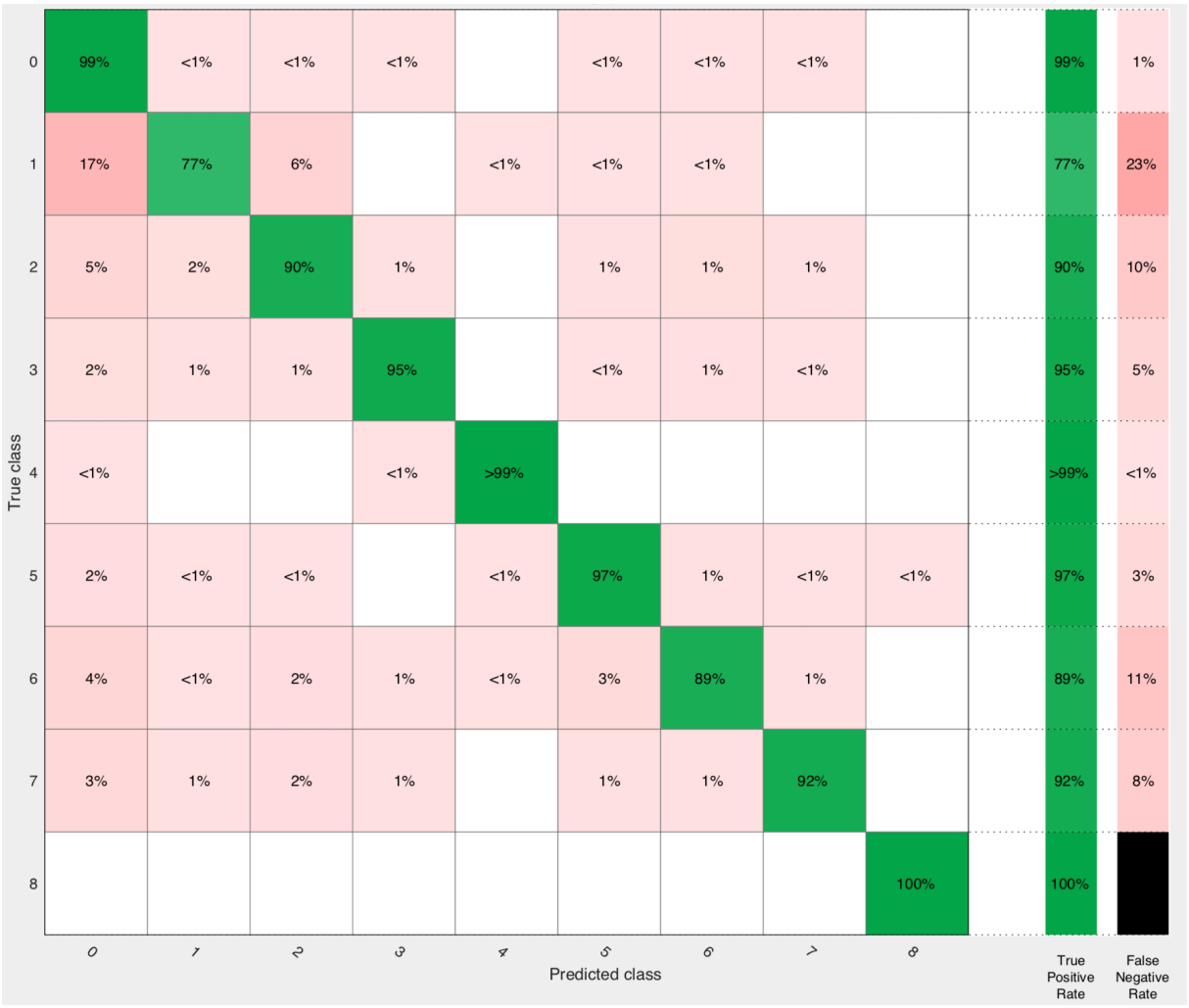

Figure 5: Confusion matrix, SVM with box constraint value 75 and kernel scale 1/4.

4 CONCLUSIONS

In this paper, it was presented a strategy to monitor the condition of a wind turbine.Eight different faults were studied and the model was able to detect and classify the type of the fault. That is, without installing extra sensors or costly purpose-built data sensing equipment. Supervised machine learning classification techniques were essential for wind turbine condition monitoring using only SCADA data already available. Two different classifiers were used, KNN and SVM. It was achieved an overall accuracy of 95.7% when KNN was used and 95% with SVM.

As a future work, we will try to optimise the model, and also include other faults.

REFERENCES

[1] Y. Vidal, F. Pozo and C. Tutiven, *Wind Turbine Multi-Fault Detection and Classification Based on SCADA Data, Energies, Vol. 11, 3018, (2018).*

[2] S.T. Kandukuri, A. Klausen, H.R. Karimi, H.R. K. G. Robbersmyr, *A review of diagnostics and prognostics of low-speed machinery towards wind turbine farm-level health management Renew. Sustain. Energy Rev., Vol. 53, 697708, (2016).*

[3] Huang, X. Wu, X. Liu, J. Gao, Y. He, *AOverview of condition monitoring and operation control of electric power conversion systems in direct-drive wind turbines under faults. Front. Mech. Eng., Vol. 12, 281302, (2017).*

[4] F.X. Ochieng, C.M. Hancock, G.W. Roberts, J. Le Kernec, *A review of ground-based radar as a noncontact sensor for structural health monitoring of in-field wind turbines blades Wind Energy, Vol. 21, 12, (2018).*

[5] Hossain, M.L.; Abu-Siada, A.; Muyeen, S.M. *Methods for AdvancedWind Turbine Condition Monitoring and Early Diagnosis: A Literature Review. Energies 2018, 11, 1309..*

[6] Ahadi, A. *MWind turbine fault diagnosis techniques and related algorithms. Int. J. Renew. Energy Res. (IJRER) 2016, 6, 8089..*

[7] De Azevedo, H.D.M.; Arajo, A.M.; Bouchonneau, N. *A review of wind turbine bearing condition monitoring: State of the art and challenges. Renew. Sustain. Energy Rev. 2016, 56, 368379.*

[8] Kandukuri, S.T.; Klausen, A.; Karimi, H.R.; Robbersmyr, K.G. *A review of diagnostics and prognostics of low-speed machinery towards wind turbine farm-level health management. Renew. Sustain. Energy Rev. 2016, 53, 697708.*

[9] Huang, S.; Wu, X.; Liu, X.; Gao, J.; He, Y. *Overview of condition monitoring and operation control of electric power conversion systems in direct-drive wind turbines under faults. Front. Mech. Eng. 2017, 12, 281302..*

[10] Ochieng, F.X.; Hancock, C.M.; Roberts, G.W.; Le Kernec, J. *AA review of ground-based radar as a noncontact sensor for structural health monitoring of in-field wind turbines blades. Wind Energy 2018..*

[11] Peng, J.; Zhang, S.; Peng, D.;Liang, K. *AApplication of machine learning method in bridge health monitoring, 2017.*

[12] Jianguo, C.; Yingyu, W.;Zhonghai, L. ;Liqiu L.; Yun Z.;Guangyan X. *Health monitoring of aeroplane structural component based on K-means clustering. 2010.*

[13] Odgaard, P.; Johnson, K. *Wind Turbine Fault Diagnosis and Fault Tolerant ControlAn Enhanced Benchmark Challenge. In Proceedings of the American Control Conference, Washington, DC, USA, 1719 June 2013; pp. 16. .*

[14] *https://nwtc.nrel.gov/FAST7.*

[15] Leahy, K.; Hu, R.L.; Konstantakopoulos, I.C.; Spanos, C.J.; Agogino, A.M.; OSullivan, D.T.J. *Diagnosing and predicting wind turbine faults from SCADA data using support vector machines. Int. J. Progn. Health Manag. 2018, 9, 111 .*

[16] Hong, X.; Xu, Y.; Zhao, G. *LBP-TOP: A Tensor Unfolding Revisit. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 2024 November; pp. 513527.*

[17] Cunningham, P.; Delany, S. J. *k-Nearest neighbour classifiers 2007.*

[18] Evgeniou, T.; Pontil, M. *Support Vector Machines: Theory and Applications, 2001*

[19] Mulak, P.; Talhar, N. *Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset 2013.*

[20] Laouti, N.; Othman, S.; Alamir, M.; Sheibat-Othman, N. *Combination of model-based observer and support vector machines for fault detection of wind turbines. Int. J. Autom. Comput. 2014, 11, 274287.*