Shaft Health Assessment

Bing-Hong Hong
Department of Electrical Engineering
National Taipei University of Technology
Taipei, Taiwan
t106318072@ntut.edu.tw

Hsin-Ting Chou
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
justin0010523@gmail.com

Shao-Tse Chien
Department of Computer Science
National Taipei University of Technology
Taipei, Taiwan
t105820049@ntut.org.tw

Chih-Hao Chiang
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
sky41282007@gmail.com

***Abstract*—The feature of shaft was extracted from time domain and frequency domain. Then the Fisher criterion was utilized to find the important features which are 20.8 Hz and 84.6 Hz, which reveal the feature in the frequency domain is more distinguishable than the feature in the time domain. Using these two features, we train the model with logistic regression, self-organizing map(SOM), SOM-MQE, and support vector machine(SVM). Finally, the shaft conditions were able to separate into three different conditions, which are health, unbalanced level 1 and level 2 shaft by logistic regression, SOM, SOM-MQE, and SVM. Finally, through comparison the confusion matrix among logistic regression, SOM, and SOM-MQE, we could find that the result of SOM, SOM-MQE, and SVM is better than logistic regression.**

Index Terms—unbalance shaft, Fisher score, logistic regression, Self-organizing map(SOM), SOM-MQE, SVM

1. INTRODUCTION

An unbalanced shaft was formed by distortion from stress, thermal distortion or deposits and oil buildup. Stress was added on the shaft and gradually making the shaft to be unbalanced with the ongoing manufacturing process. Thermal distortion happened if the shaft was operating in high temperatures, metal or other materials could become more easily to extend, causing shaft distortion. Moreover, deposits and oil buildup happened if the machine was involved in material handling. It would be possible that minerals, dust or dirt began to build up on the shaft, making distortion to occur. This could also happen if the shaft was exposed to oil. Oil gradually accumulated into parts, making the shaft to be unbalanced.

Due to the unbalanced force on the shaft, the shaft vibrated, causing the parts of the machine to lose and even damage other parts. The vibration also caused the machine to generate extra noise. Also with the vibration, additional force on rotation would pressure on the bearings holding up the shaft, making the life of a bearing to be shortened. Work conditions also became unsafe. With the extra vibration, parts of the machine were more easily to break. And with the increased possibilities of the machine to breakdown, maintenance time would also increase, causing loss of money and time [1].

Fortunately, the technology got a lot of advanced such as embedding system and computer; hence, the machine learning algorithm such as Self organizing map and other neural network family can be implemented. Through these technology, the condition of shaft could be detected into three different level, which is healthy, unbalanced level 1, and unbalanced level 2.

This report is organized as follows. Section 2 mentioned the methodology and materials of this report. Section 3 demonstrated the performance by

real case of unbalanced shaft and healthy shaft. Section 4 summarized the conclusions.

2. METHODOLOGY AND MATERIALS

2.1. Fisher score

Fisher criterion is used to measure how suitable a single variable is for separating classes. The formula of fisher criterion is shown in (1).

$$F(X^j) = \frac{\sum_{k=1}^{c} n_k (u_k^j - u^j)^2}{\sum_{k=1}^{c} n_k (\sigma^j)^2} \quad (1)$$

The numerator of fisher criterion is the difference between the mean of classes. Moreover, the denominator of fisher criterion is the sum of the classes' variance. Therefore, the higher the fisher criterion is, the better the feature is. The diagram of fisher criterion is shown as follow[5].

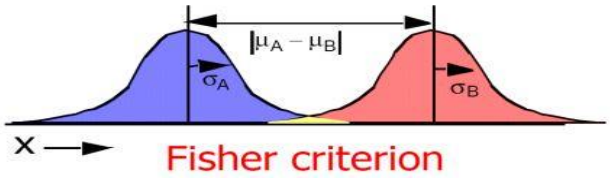

Fig 1. The diagram of Fisher criterion

2.2. multi-class logistic regression

the goal of Multi-class Classification is to classify the features of dataset into many different classes. The difference between Binary Classification and Multi-class Classification is that previous one only has two classes, but the latter one has more than two classes [3]. For example, binary classification just like a yes-no question, whereas multi-class classification resembles multiple choice question.

Logistic regression can be either a kind of binary classification or a kind of multi-class classification while its result is a probability which is between 0 and 1. Through searching the maximal probability, the data can be separate to many classes. The logistic function is defined as:

$$\text{logistic}(\eta) = \frac{1}{1+exp(-\eta)} \quad (2)$$

The steps of logistic regression are similar to linear regression. In the linear regression model, the relationship between outcome and features is modeled as:

$$\eta = \hat{y}^{(i)} = \beta_0 + \beta_1 {x_1}^{(i)} + \cdots + \beta_p {x_p}^{(i)} \quad (3)$$

For logistic regression, the results are probabilities between 0 and 1, so we combine two functions together:

$$\mathrm{P}(y^{(i)} = 1) = \frac{1}{1+exp\left(-\left(\beta_0 + \beta_1 {x_1}^{(i)} + \cdots + \beta_p {x_p}^{(i)}\right)\right)} \quad (4)$$

2.3. SOM

Self-organizing map (SOM) is a type of unsupervised learning algorithm that operates in two section. One is training while the other is mapping. Training section built the map which used input examples. However, mapping section automatically classified a new input vector and updated the weight of neuron during each iteration. The map is usually defined as a finite two-dimensional space which neurons are arranged in a regular hexagonal or rectangular grid. Each neuron is associated with a weight vector which is a position in the input space[4].

SOM Training Algorithm

given training data points: $\mathcal{X} = \{x^k\}$

1. Present vector $x^q \in \mathcal{X}$
2. Determine winning unit i^*
$$i^* = argmin||w_i - x^q||$$
3. Update all weights by Δw_{ij}
$$\Delta w_{ij} = \eta \wedge (i, i^*, t)\left(x_j^q - w_{ij}\right)$$
$$\wedge (i, i^*, t) = \exp\left[\frac{-||r_i - r_{i^*}||^2}{2\sigma^2(t)}\right]$$
$$\sigma(t) = \sigma_0 exp\left(-\frac{t}{\tau}\right)$$
4. Repeat from 1.

2.4. SOM-MQE

MQE (Minimum Quantization Error) is a binary classification method, which could be applied to health assessment. We used healthy data of a shaft as training data to train a self-

organization map. MQE found the shortest distance between new data and the neurons of trained model. The workflow to obtain MQE is shown in Fig 2.

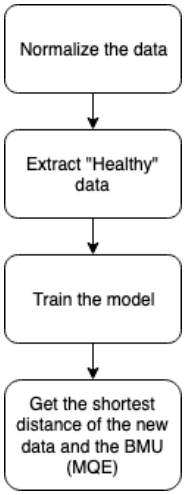

Fig 2. The flowchart of SOM-MQE.

2.5. Support vector machine

Support vector machine is a binary supervised algorithm that can divide data by a decision boundary and its margin [7][8]. For linear SVM, we have a dataset while x is feature one and y is feature two that has n rows of data which forms in:

$$(x_1 + y_1) + (x_2 + y_2) + \cdots + (x_n + y_n) \qquad (5)$$

For linear SVM, decision boundary will be a linear equation that can separate data evenly. The linear equation will form in:

$$\vec{w} * x - b = 0 \qquad (6)$$

There will be two hyperplanes called “margin”, the distance of the margin is $\frac{2}{||\vec{w}||}$, and these two hyperplanes are form in (7), and the diagram of SVM is shown in Fig 3.

$$\vec{w} * x - b = 1 \text{ and } \vec{w} * x - b = -1 \qquad (7)$$

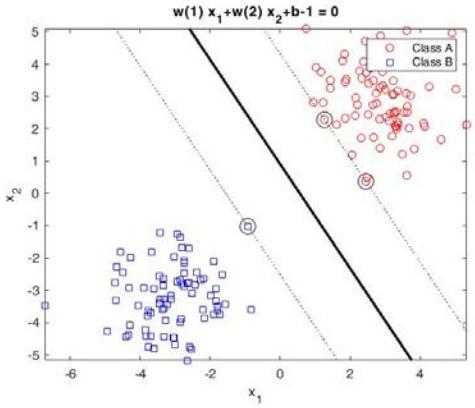

Fig 3. The diagram of SVM [9].

3. EXPERIMENTAL RESULTS

In the experimental results, we applied standardization to normalize our data in the preprocessing part. Then the features in time domain were extracted by some statistic technique such as mean, variance, skewness. Moreover, the features in frequency domain was extracted by Fourier transform as well. After we extracted the features, the feature selection technique, Fisher criterion, was applied to search which feature is important. Finally, we applied logistic regression, SOM, and SOM-MQE to train our machine learning model. Finally, we use confusion matrix to evaluate the performance of each algorithm. The flow chart of experiment is shown in the Fig 4.

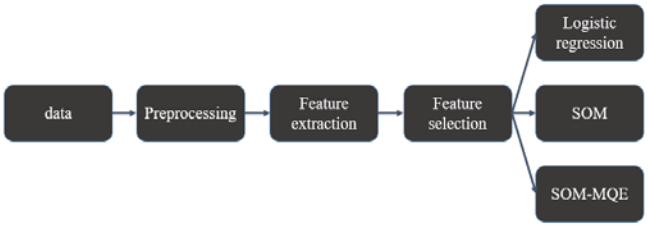

Fig. 4 The flow chart of experiment

3.1. Shaft dataset

In the data we obtained, the shaft is rotating at 20 Hz. The sampling rate of the data is at 2560 Hz. We have two kinds of sets of data. One is training data, and the other is testing data. Training data consists of three parts, healthy, unbalance shaft level 1, and unbalance shaft level 2. Each having 20 observations of vibration data. Testing data consists of 10 observations of healthy, 10

observations of unbalance level 1, and 10 observations of unbalance level 2. Each observation of the shaft vibration contains 38400 records.

3.2. Data preprocessing

In data preprocessing, we applied the standardization to normalize the signal, which mean is zero and standard deviation is one. Then the features of data were extracted from the time domain and the frequency domain. The result of Fourier transform is shown in Fig 5. It should be noted that we can observe that there has some significant different between the healthy shaft and the unbalanced shaft in the first harmonic wave of the frequency domain.

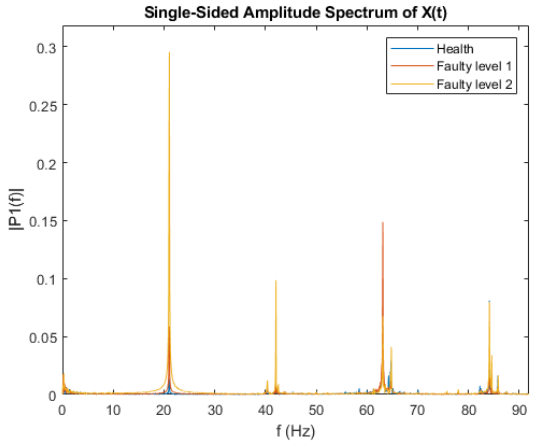

Fig 5. The result of Fourier transform

3.3. Feature extraction and selection

In feature selection, the fisher criterion was applied to find which feature can distinguish among healthy shaft, unbalanced shaft level1, and unbalanced shaft leve2 better. The fisher score of each feature is shown in Fig 6. In our case, the frequency 20.8 Hz and 84.6 Hz were selected by the fisher criterion, which has the higher score than the other features.

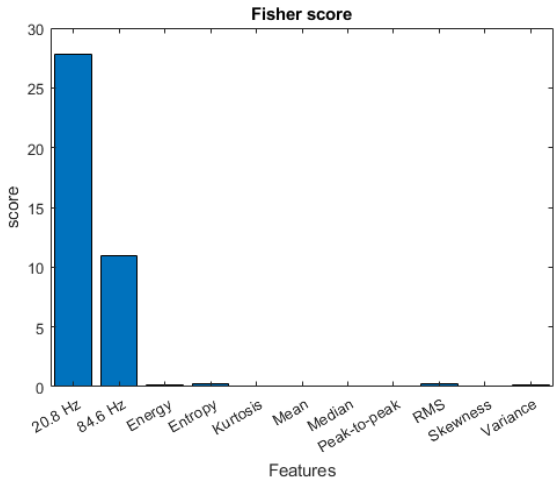

Fig 6. The result of Fisher criterion

3.4. The result of SOM

First of all, we trained three SOM models, which based on three different training data, which are healthy, unbalanced level 1, and unbalanced level 2. Second, through searching the minimum distance, we label the testing data points as the group which is the nearest. Finally, we compare the results of SOM and true labels, and calculate the confusion matrix shown in the Fig 7.

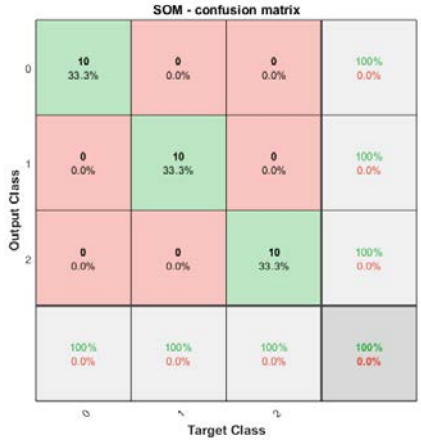

Fig. 7 The confusion matrix of SOM

3.5. The result of SOM-MQE

MQE indicated how far the data is from normal condition to indicate degradation. In our results, we found a large difference between normal condition, unbalanced level 1 and 2. The first 10 samples, the healthy samples, has MQE values lower than 0.003. While other conditions of the shaft have a much higher MQE score. The results of MQE value of each training data and

testing data is shown in Fig 8 and Fig 9. respectively, and confusion matrix is shown in Fig 10.

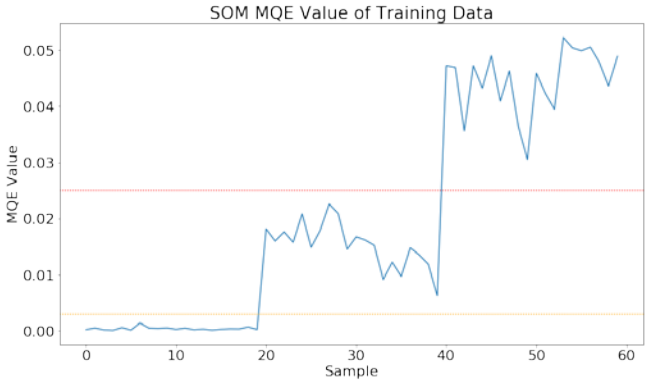

Fig 8. The SOM-MQE value of training data

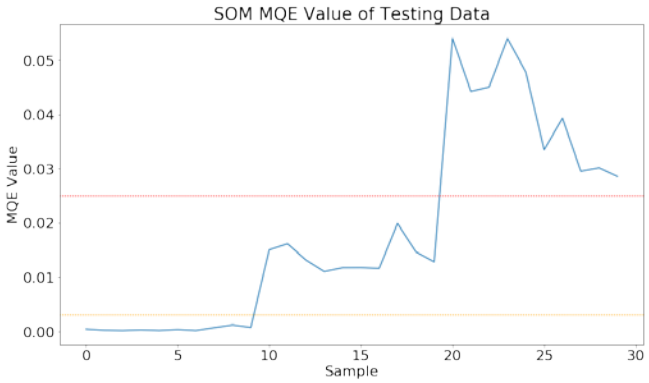

Fig 9. The SOM-MQE value of testing data

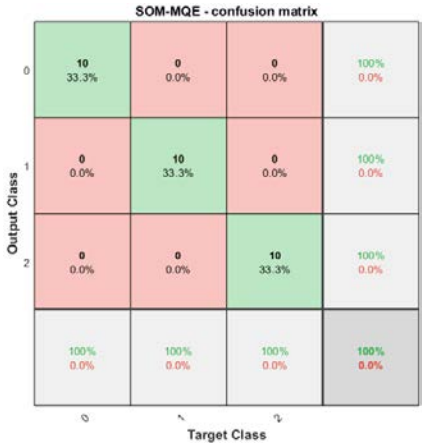

Fig 10. The confusion matrix of SOM-MQE

3.6. The result of logistic regression

After the important feature was determined, the logistic regression was utilized to train the model. Through logistic regression, we could calculate how much probability of each data belongs to each class. Then, we determined that the prediction label of each data can be determined by the highest probability of class. Finally, based on the prediction label and the ground truth of each data, the confusion matrix is shown in Fig 11.

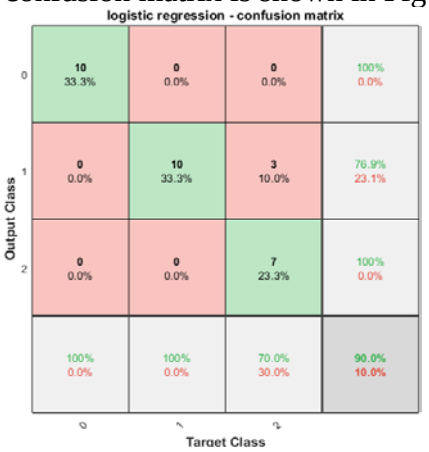

Fig 11. The confusion matrix of logistic regression

3.7. The result of SVM

In the multi-classification problem, a method called one against one was utilized to classify the data into three classes. First of all, we have divided the data into two groups at the beginning, one is healthy, the other is unbalance. In Fig 12, we can see that decision boundary has perfectly divide healthy and unbalance data into two classes.

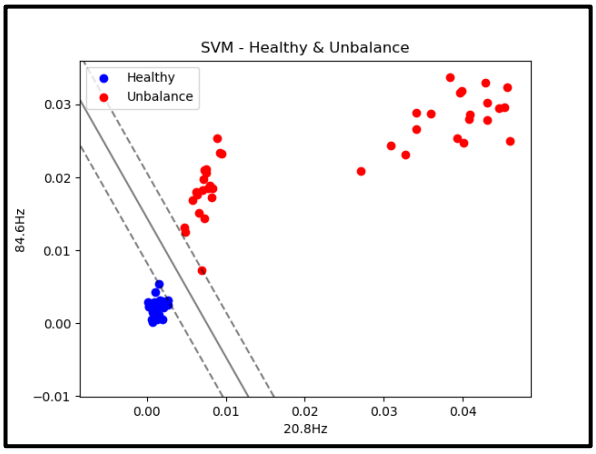

Fig 12. The decision boundary between Healthy and Unbalance

Then we separated unbalance data into unbalance one and unbalance two, and the result is perfect as well. Through this result, we could

observe that the distance between the unbalance 1 and the unbalance 2 data is large enough to separate to two classes. The result of second boundary is shown in Fig 13.

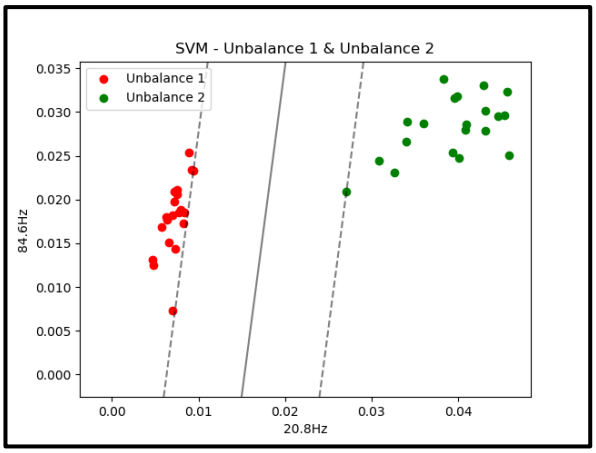

Fig 13. The decision boundary between unbalance level 1 and unbalance level 2

Finally, we combined two model together to get training model and map the testing data on the training model to get the testing result. The final SVM model is shown in Fig 14. According to this result, we can see that the data can be separated to three classes perfectly.

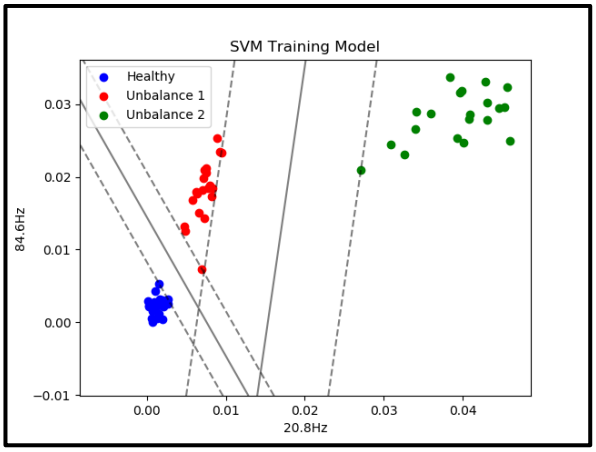

Fig 14. The final decision boundary of SVM

Finally, we use the testing data to test our SVM model. The black solid lines in the figure are the decision boundary, which used to separate three classes, healthy, unbalance 1, and unbalance 2. From Fig 15, we can see that the SVM model can separate to three conditions well.

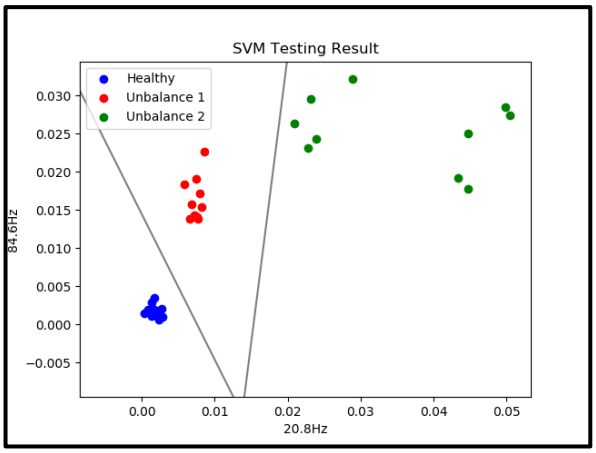

Fig 15. The testing result of SVM

Finally, we compare the results of SOM and true labels, and calculate the confusion matrix shown in the Fig 16., which accuracy is 100%.

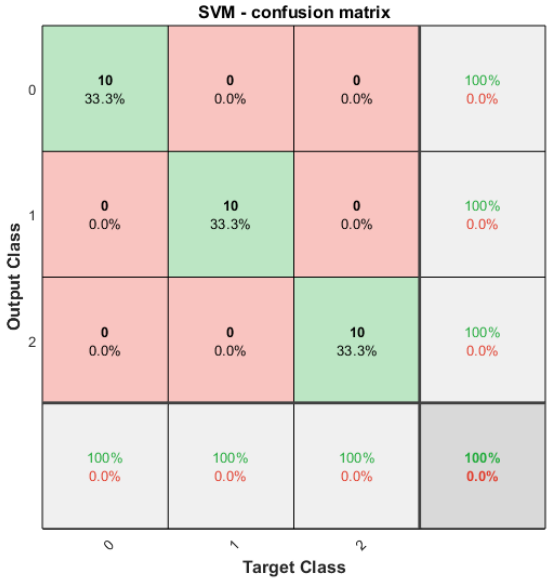

Fig 16. The confusion matrix of SVM

4. CONCLUSION

In conclusion, after we extracted the feature from time domain and frequency domain of the data, we applied the Fisher criterion to search which the important feature is. Then the two frequencies of the data, 20.8 Hz and 84.6 Hz, were selected by the scores of Fischer criterion. Using these two features, we train the model with logistic regression, SOM, SOM-MQE, and SVM, which are able to separate different shaft conditions

(healthy, level 1 and level 2) with a very significant difference. Finally, through comparison the result of logistic regression, SOM, SOM-MQE, and SVM, we could discover that the performance of SOM, SOM-MQE, and SVM is better than the logistic regression. The reason behind this phenomenon is that the SOM and SOM-MQE contain a lot of neurons, which updated every iteration so that the performance might have better than the logistic regression. Last but not least, the SVM algorithm classify the data to three class perfectly since the distance of training data of each class is large enough to separate accurately.

REFERENCES

[1] Test Devices, *Three Major Ways Your Manufactured Rotating Component Can Become Unbalanced*, Test Devices, Aug. 12, 2017. Accessed on: Aug. 14, 2019. [Online]. Available: https://www.testdevices.com/three-major-ways-manufactured-rotating-component-can-become-unbalanced/

[2] Urvashi Jaitley, *Why Data Normalization is necessary for Machine Learning models*, Medium, Oct. 8, 2018. Accessed on: Aug. 14, 2019. [Online]. Available: https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029

[3] Wikipedia, *Multiclass classification*, Wikipedia, Accessed on: Aug. 14, 2019. [Online]. Available: https://en.m.wikipedia.org/wiki/Multiclass_classification

[4] Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464-1480.

[5] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

[6] Lapira, Edzel, et al. "Wind turbine performance assessment using multi-regime modeling approach." Renewable Energy 45 (2012): 86-95.

[7] GeeksforGeeks, *Classifying data using Support Vector Machines(SVMs) in Python*, GeeksforGeeks, Accessed on: Aug. 14, 2019. [Online]. Available: https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-python/

[8] Wikipedia, *Support vector machine*, Wikipedia, Accessed on: Aug. 14, 2019. [Online]. Available: https://en.m.wikipedia.org/wiki/Support-vector_machine

[9] Dhairya Kumar, *Demystifying Support Vector Machines*, Medium, Feb. 26, 2019. Accessed on: Aug. 14, 2019. [Online]. Available: https://towardsdatascience.com/demystifying-support-vector-machines-8453b39f7368