ABSTRACT

In this paper, we use various regression models and Artificial Neural Network (ANN) to predict the centrifugal compressor performance map. Particularly, we study the accuracy and efficiency of Gaussian Process Regression (GPR) and Artificial Neural Networks in modelling the pressure ratio, given the mass flow rate and rotational speed of a centrifugal compressor. Preliminary results show that both GPR and ANN can predict the compressor performance map well, for both interpolation and extrapolation. We also study the data augmentation and data minimization effects using the GPR. Due to the inherent pressure ratio data distribution in mass-flow-rate and rotational-speed space, data augmentation in the rotational speed is more effective to improve the ANN performance than the mass flow rate data augmentation.

Keywords: Centrifugal Compressor, Gaussian Process Regression, Neural Network, Artificial Intelligence, Machine Learning

INTRODUCTION

The design and off-design performance of a centrifugal compressor can be estimated from map-based calculations or physics-based one-dimensional models [1]. In the physics-based one-dimensional models, enthalpy losses calculated from analytical equations and empirical correlations is used to estimate the outlet pressure and temperature rise. A dynamic vapor cycle modeling toolset ATTMO was developed to address two-phase flow systems. The performance of the centrifugal compressors was modelled using first-principle enthalpy based calculations [2]. Computational Fluid Dynamics (CFD) simulations on centrifugal compressors are used to understand the complex flow characteristics, and to further improve first-principle modeling. For instance, Sundstrom et al. [3] numerically investigated the flow instability in the compressor flow near surge. The unsteady features of the flow field are quantified by means of Fourier transformation analysis, proper orthogonal decomposition and dynamic mode decomposition. Using ANSYS CFX, Sausse et al. [4] conducted CFD simulation on centrifugal compressors with R134a as the fluid medium. Wan et al. [5] conducted both 1D modeling using Vista Centrifugal Compressor Design (VCCD) and CFD simulations to study the sizing and performance of air-based and refrigerant-based centrifugal compressors.

Recently, due to its robustness and flexibility in model development, machine learning techniques are also used to modelling the performance of centrifugal compressors. In modeling, the operating conditions of a centrifugal compressor can be specified by two input parameters, i.e., the impeller rotational speed, and the mass flow rate. The corresponding performance of the centrifugal compressor can be described by pressure ratio and efficiency. Fei et al. [6] proposed an artificial neural network integrating feed-forward back-propagation neural network with
Gaussian kernel function to predict the compressor pressure ratio. Li et al. [7] developed regression model to predict both the pressure ratio and the efficiency of a centrifugal compressor using partial least squares.

In this article, CCD simulations are conducted to generate compressor data, based on which typical machine learning techniques are used to model the centrifugal compressor performance. We believe that the same machine learning framework can also work on compressor experimental data, and other type of centrifugal compressors with different sizes and working fluid medium. More specifically, we first use various regression models to predict the compressor performance map, focusing on the interpolation of pressure ratio. Further, we study the extrapolation performance of regression model, such as GPR. An ANN is also developed for both interpolation and extrapolation of pressure ratio.

TEST PROCEDURE AND DATA ANALYSIS

Using Ansys Vista CCD, we generated the centrifugal compressor data, consisting of the pressure ratio at given mass flow rates and impeller rotational speeds. The latter ranges from 10,000 rpm to 90,000 rpm, with a step of 10,000 rpm. At each fixed speed, there are 25 mass flow rate samples. Hence, in a preliminary study, 3 parameters (rotational speed, mass flow rate and pressure ratio) and 225 samples are used. Figure 1 plots the the pressure ratio as a function of rotational speed and mass flow rate. We also conducted data augmentation study, in which more data are provided to investigate the effects of increased number speeds and increased number of mass flow rate. The number of data used can be found in Table 1.

As shown in Table 1, three types of tests are conducted to understand the capability of machine learning in the prediction of the centrifugal compressor performance map. In Test 1, we use various regression models, and ANN to fit the compressor pressure ratio. All data will be used, with 67% for training and 33% for test. We iteratively adjust the percentage of data used for training and test and compare the obtained errors. It shows that the variation in errors is negligible once the training data is above 67%. In test 2 and test 3, we conduct extrapolation fitting. Data at rotational speeds 20,000 rpm, 40,000 rpm, 60,000 rpm are used for test, and data at other speeds are used for training. In Test 3, the same procedure is taken, but more data of pressure ratio at various rotational speeds and mass flow rates are provided to see the effects of data augmentation on predictions.

TABLE 1. Tests conducted in current study

<table>
<thead>
<tr>
<th>Tests</th>
<th>Type of prediction</th>
<th>Models used</th>
<th># samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1.1</td>
<td>Interpolation</td>
<td>Various</td>
<td>9x25</td>
</tr>
<tr>
<td>Test 1.2</td>
<td>Interpolation</td>
<td>GPR</td>
<td>9x25</td>
</tr>
<tr>
<td>Test 1.3</td>
<td>Interpolation</td>
<td>ANN</td>
<td>9x25</td>
</tr>
<tr>
<td>Test 2.1</td>
<td>Extrapolation</td>
<td>GPR</td>
<td>9x25</td>
</tr>
<tr>
<td>Test 2.2</td>
<td>Extrapolation</td>
<td>ANN</td>
<td>9x25</td>
</tr>
<tr>
<td>Test 3.1</td>
<td>Extrapolation</td>
<td>GPR</td>
<td>18x25</td>
</tr>
<tr>
<td>Test 3.2</td>
<td>Extrapolation</td>
<td>GPR</td>
<td>9x50</td>
</tr>
</tbody>
</table>

EVALUATION METRICS

Coefficient of determination ($R^2$)

The coefficient of determination is the the square of the correlation between predicted values and the actual values, as defined in equation 1. Also, it is the proportion of the variance in the dependent variable that is predictable from the independent variable(s) [8].

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \bar{y})^2}$$  \hspace{1cm} (1)

where $\bar{y}$ is:

$$\bar{y} = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} y_i$$  \hspace{1cm} (2)
The coefficient of determination is used to test hypotheses and characterize the prediction accuracy of the regression function.

**Mean squared error (MSE)**

As defined in equation (3), MSE is the average of the squares of the errors i.e., the average squared difference between the estimated values and the real values [8]. MSE has the same units as the data it measured. Consequently, it evaluates the quality of an estimator or set of predictions in terms of its variation and degree of bias.

\[
MSE(y, \hat{y}) = \frac{1}{n_{\text{samples}}-1} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2
\]

(3)

It is also noteworthy that a small MSE does not necessarily mean a good fit when the actual values are close to zero.

**Mean absolute percentage error (MAPE)**

The MAPE is widely used as a loss function in regression problems of machine learning. In this case, it is used as an evaluation metric, giving the percentage deviation between the predicted and the real value. Thus, it clearly shows how close the prediction to the real data is.

\[
\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]

(4)

**GAUSSIAN PROCESS REGRESSION (GPR)**

GPR is a distribution of functions, and is defined by a mean and covariance function. The inference takes place directly in the space of functions from the function-space view [9]. Using Gaussian Processes (GP) to predict the compressor performance map is justified due to its capacity of finding the posterior distribution over the possible functions \( f(x) \) using observed data. This non-parametric characteristic allows us to transform the prior distribution into a posterior distribution. Thus, it is possible to identify all the functions that match our data.

Using equation (5), the prior distribution can be transformed into the posterior distribution.

\[
\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right)
\]

(5)

in which \( N \) is a normal distribution with mean \( \mu \) and variance \( \sigma^2 \).

Equation (6) is the joint normal distribution followed by \( f \), the joint probability of the observed values, and \( f' \), the prediction of new input values.

\[
\begin{pmatrix} f \\ f' \end{pmatrix} \sim N\left( \begin{pmatrix} \mu \\ \mu' \end{pmatrix}, \begin{pmatrix} KK' \\ K'K'' \end{pmatrix} \right)
\]

(6)

in which \( K \) is the covariance matrix between all observed samples, and it also called the kernel function of the observed values. \( K' \) is the similarity between the known values (training) and the new inputs. \( K'' \) indicates the resemblance of testing values among each other. It is interesting to note that if the vectors \( x \) and \( x' \) are similar, \( f \) is basically same as \( f' \).

It is interesting to analyze the influence of the covariance on the results. As a versatile model, GPR can adapt the model by putting some constraints. One of them is the covariance matrix, which can modify the smoothness of the function. This performance is observed because the magnitude of values decrease as the points are further apart in the covariance matrix and also, because it is symmetric. It means that \( x \) and \( x' \) are practically equal. Thus, due to the proximity of these input values, the output ones will have the same behaviour. As a summary, the covariance function measures the similarity between \( x \) and \( x' \), and it determines the behaviour of our model.

The properties pointed above will determine the distribution over functions. The equation (7) indicates \( f \) is distributed as a Gaussian process, which is defined by the mean function \( (m(x)) \) and the covariance function \( (K(x, x')) \).

\[
f \sim GP(m(x), K(x, x'))
\]

(7)

We can just take certain number of data points and compute the mean function \( m \) and the covariance function \( K \) for those data points \( (X = [x_1, \ldots, x_n]) \), and then sample from the respective joint Gauss distribution \( (f(x)) \).

\[
f(X) \sim N(m(X), K(X, X))
\]

(8)

In this study, we used Matern kernel to perform Gaussian process regression. Matern covariance is given by equation (9) for two observations with distance \( r = |x - x'| \),

\[
k_{\text{Matern}}(r) = \frac{2^{1-v}}{\Gamma(v)} \left( \frac{\sqrt{2v}r}{l} \right)^v K_v \left( \frac{\sqrt{2v}r}{l} \right)
\]

(9)

in which \( \Gamma \) is the gamma function, \( K_v \) is the modified Bessel function of the second kind. Both \( v \) and \( l \) are non-negative covariance parameters. The length-scale \( l \) defines the smoothness.
and \( \nu \) determines the periodicity of the kernel. These kernel hyper-parameters control the model structure.

**DATA NORMALIZATION**

In our data, the range of the rotational speed is one million times higher than the range of the mass flow rate. The euclidean distance used in many models will be dominated by the rotational speed. Therefore data normalization is needed to scale the range of the data features. Three scaling methods are used: Standardization (10), mean normalization (11), and min-max normalization (12), as shown in the follow expressions.

\[
\begin{align*}
    x' &= \frac{x - \bar{x}}{\sigma} \\
    x' &= \frac{x - \text{mean}(x)}{\text{max}(x) - \text{min}(x)} \\
    x' &= \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)}
\end{align*}
\]

**PREDICTION USING ALL ROTATIONAL SPEEDS**

In this section, Test 1, i.e., the prediction using all the rotational speeds is conducted. In this case, 150 random samples (67\%) are used for training and 75 samples (33\%) for testing. The data have been scaled and shuffled before training.

**Interpolation using different regressors**

The models used in this study are: K-Neighbors Regressor (KNR), Bayesian Ridge Regressor (BRR), Decision Tree Regressor (DTR), Gradient Boosting Regressor (GBR), Kernel Ridge Regression (KRR), Support Vector Regressor (SVR), and Ada-Boost Regressor (ABR). Each model was optimized through the GridSearchCV function. The optimal configuration of each model is shown in Appendix. Also, using GPR, we are able to choose the kernel hyper-parameters without a need to perform a grid search on a cross-validated loss function, making a model more flexible and accurate.

Table 3 shows the results from SVR using linear and RBF (Radial Basis Function) kernels. Each kernel represents a method to parameterize the data. From Table 3, we can see that RBF fit our data better. This is expected since the data in figure 1 has a polynomial degree higher than quadratic.

Table 3 shows the results from SVR using linear and RBF (Radial Basis Function) kernels. Each kernel represents a method to parameterize the data. From Table 3, we can see that RBF fit our data better. This is expected since the data in figure 1 has a polynomial degree higher than quadratic.

Using all the rotational speeds, high accuracy interpolation of the performance map can be obtained within milliseconds of computational processing time.

**Interpolation using GPR**

In this section, GPR is used for interpolation. GPR uses the observed training data to define a posterior distribution over target functions, and its mean is used for prediction. Also, using GPR, we are able to choose the kernel hyper-parameters without a need to perform a grid search on a cross-validated loss function, making a model more flexible and accurate.

Figure 2 shows the predicted compressor performance map using GPR. It can be seen that the predicted values fit the real pressure ratio very well. Table 4 shows that less than 1\% error can be obtained using this model. Also, the high R2 value indicates that the predictions are full correlated with the model fit.

**Interpolation using ANN**

In this section, we developed a centrifugal compressor ANN with two inputs (rotational speed and mass flow rate) and one output (pressure ratio), using 150 training samples and 75 testing samples. The activation function used in the hidden layers is the Rectified Linear Unit (ReLU), and the optimizer is Adam. Both metrics and loss are controlled by MSE. The EarlyStopping callback function was used to avoid the model over-fitting. A complexity study was conducted to find the best performance of the model, using an iterative process of increasing the number of neurons and the hidden layers. Complexity indicates the
### TABLE 2. Test 1.1 Comparison the predictions of different models.

<table>
<thead>
<tr>
<th>Model</th>
<th>KNR 99.96</th>
<th>BRR 90.67</th>
<th>DTR 99.94</th>
<th>GBR 99.95</th>
<th>KRR 100</th>
<th>SVR 99.69</th>
<th>ABR 99.42</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ [%]</td>
<td>99.96</td>
<td>90.67</td>
<td>99.94</td>
<td>99.95</td>
<td>100</td>
<td>99.69</td>
<td>99.42</td>
</tr>
<tr>
<td>MSE [$10^{-3}$]</td>
<td>4.84</td>
<td>105.48</td>
<td>7.44</td>
<td>6.70</td>
<td>0.05</td>
<td>3.82</td>
<td>7.23</td>
</tr>
<tr>
<td>MAPE [%]</td>
<td>0.62</td>
<td>15.79</td>
<td>1.03</td>
<td>0.92</td>
<td>0.28</td>
<td>3.32</td>
<td>4.18</td>
</tr>
<tr>
<td>Time [ms]</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>50</td>
<td>7</td>
<td>3</td>
<td>54</td>
</tr>
</tbody>
</table>

### TABLE 3. Test 1.1 Results using different SVR kernels.

<table>
<thead>
<tr>
<th>SVR</th>
<th>Linear</th>
<th>RBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ [%]</td>
<td>90.47</td>
<td>99.69</td>
</tr>
<tr>
<td>MSE</td>
<td>0.107</td>
<td>0.0038</td>
</tr>
<tr>
<td>MAPE [%]</td>
<td>16.34</td>
<td>3.32</td>
</tr>
<tr>
<td>Time [ms]</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

### TABLE 4. Test 1.2 Results using GPR.

<table>
<thead>
<tr>
<th>Test 1.2</th>
<th>GPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ [%]</td>
<td>100</td>
</tr>
<tr>
<td>MSE [$10^{-3}$]</td>
<td>0.478</td>
</tr>
<tr>
<td>MAPE [%]</td>
<td>0.16</td>
</tr>
<tr>
<td>Time [ms]</td>
<td>3</td>
</tr>
</tbody>
</table>

### TABLE 5. Test 1.3 Results using ANN.

<table>
<thead>
<tr>
<th>Test 1.3</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ [%]</td>
<td>99.34</td>
</tr>
<tr>
<td>MSE [$10^{-3}$]</td>
<td>1.07747</td>
</tr>
<tr>
<td>MAPE [%]</td>
<td>2.74</td>
</tr>
<tr>
<td>Time [s]</td>
<td>1</td>
</tr>
</tbody>
</table>

### Extrapolation of the Rotational Speed

As shown in Table 1, test 2 and test 3 are developed to evaluate the extrapolating performance of our models at other rotational speeds.

### Extrapolation using GPR

The best model parameters can be found by using the function GridSearchCV. The hyper-parameters in the kernel Matern are $\alpha = 1e-08$, length-scale $l = 1$, and $\nu = 1.5$, in which $\alpha$ defines the noise of the system. The value of $\alpha$ used in this study is higher than the default one ($1e-10$) to ensure the calculated values forming a positive definite matrix. The Matern kernel can be obtained using the code below, where $gp$ is the GPR function in Sklearn, and $X$ denotes the data points.
Figure 4 presents the extrapolated pressure ratio at rotational speeds 20,000 rpm, 40,000 rpm and 60,000 rpm. It can be seen that the predicted values are reasonably well, especially at the low and high mass flow regions at each speed line. The higher difference between the predication and the real values occurs at the intermediate mass flow rate, and this difference gets bigger when rotational speed increases.

Table 6. Test 2.1 Results using GPR.

<table>
<thead>
<tr>
<th>Test 2.1</th>
<th>GPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2 [%]</td>
<td>99.68</td>
</tr>
<tr>
<td>MSE [10^{-3}]</td>
<td>2.06578</td>
</tr>
<tr>
<td>MAPE [%]</td>
<td>1.34</td>
</tr>
<tr>
<td>Time [ms]</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 5 shows the normalized standard deviation of predicted pressure ratio at each speed. This test is conducted using 25 samples. Since the mass flow rates are different for each rotational speed, they are clearly separated to each other. Thus, we can see that a higher error around the middle mass flow rate at each rotational speed. For example, at speed 40,000 rpm, the error is around 3.9% near mass flow rate 0.3; while it is around 0.1% and 1% when the mass flow rate is 0.2 and 0.33, respectively. The same trend is also shown in Figure 4. Also, the prediction error gets higher when the rotational speed increases. For instance, the peak value of the error is around 5% when the speed is 60,000 rpm, higher than the peak error obtained at 40,000 rpm.

Figure 6 shows the influence of the three scaling methods in comparison with the non-scaled data. In this case, the MAPE from non-scaled data is around 100% and the scaled one around 2%. The three methods basically give very similar MAPE. We used min-max normalization in our following study, to make the normalization data ranges from 0 to 1.
Table 7 shows the results leaving one speed out as test and the others as training.

**Extrapolation using Artificial Neural Network (ANN)**

Using the same data as before at speeds of 20,000 rpm, 40,000 rpm, 60,000 rpm, an artificial neural network was developed. The structure is same as the one used in the interpolation (explained in the section ). RBF kernel [10] is used in our study. The RBF kernel is expressed in equation (13).

$$k(x,y) = \exp(-\gamma \|x-y\|^2)$$  \hspace{1cm} (13)

in which $\gamma$ is a free parameter of the kernel. From Table 8, it can be seen that small values of MAPE and MSE are achieved, despite more computational time is needed.

As in previous section, complexity analysis is conducted through an iterative process to find the optimal structure of our model. The obtained optimal configuration is $R_2^1$, $R_{12364}^2$, $R_{596}^3$, $R_{64}^4$, and $R_1^1$ in the input layer, three hidden layers, and the output layer, respectively.

**DATA AUGMENTATION AND DATA MINIMIZATION**

In this section, we study the influence of data augmentation, i.e., increasing of data samples. Mean absolute percentage error (MAPE) is used for prediction evaluation.

1. **Mass flow rate augmentation using analytical method:**
   To study the effects of augmented mass flow rate, we increased the number of mass flow rate and keep the number of speed lines to be 9. With the Python `polyfit` function, an analytical polynomial with fourth degree was developed to describe the pressure ratio on all speed lines. More data at various mass flow rate can be generated using this analytical formula. These values can be used as training data. Figure 9 shows the generated data using the analytical formula. In figure 10, it is labelled as DA samples.

2. **Rotational speed augmentation:**
   To study the effects of augmented rotational speed, we increased the number of speed lines with the number of mass flow rate fixed at 25. The rotational speeds ranges from 10,000 rpm to 95,000 rpm, with the step of 5,000 rpm. Thus, we have 18 rotational speeds in total. In figure 10, it is labeled as DA speeds.

It can be seen from Figure 8 that both GPR and ANN give accurate predictions indicated by the magnitude of the MSE. GPR shows better prediction than ANN when the rotational speed is below 60,000 rpm. ANN shows better performance when the rotational speed is above that.
From Figure 10, it can be seen that as the number of rotational speeds increases, the value of MAPE is reduced more than a half. It shows that increasing data of speed lines of an centrifugal compressor plays a central role in the improvement of the model accuracy. On the other hand, as the number of samples (mass flow rate) on a fixed speed line increases, the performance indicated by MAPE is pretty much the same, especially at high rotational speed. It indicates that 25 mass flow rate data points on each speed line is more than enough to give a good prediction on the pressure ratio. The further question is then, what is the minimum number mass flow rate data needed to develop a model and through which predict reasonably well on the pressure ratio. Therefore, we reduce the number of mass flow rate data and explore the minimum number needed to develop a model. Figure 11 shows how MAPE changes at various speed lines when less number of mass flow rate (samples) is used. Thus 10 mass flow rate data on each speed line is enough in predicting the pressure ratio of the centrifugal compressor. When the original data is used, as the number of mass flow rate reduces from 25 to 10, comparable MAPE value are obtained. The error is around 2% from 20,000 rpm to 60,000 rpm, and increases to 7% (using data augmentation) at 80,000 rpm. As the number of mass flow rate is further reduced to 7, the predicted of pressure ratio at higher rotational speed such as 70,000 rpm deteriorates. However, using the developed analytical polynomial (as shown in the lower figure), seven mass flow rates can still give good prediction, since the selected seven mass flow rates are optimized to achieve good fitting to the data of pressure ratio. It can be seen that the error at 70,000 rpm is higher than that of 80,000 rpm. This can be partially ascribed to the non-linear nature of the centrifugal compressor data. It is worth to explore more fundamental reasons in future work. To understand why the data augmentation on rotational speed is more effective than augmentation on mass flow rate, we shuffled the original data and conducted Principal Component Analysis (PCA) to extract some of the data features, indicated by the components of the PCA. Figure 12 shows that nine group of data are outlined by the the dominant feature, which corresponds to the rotational speed. This suggests that data augmentation in the dominate feature is more effective than the non-dominant ones in order to improve the predication accuracy.

**SUMMARY**

In this paper, we predicted centrifugal compressor performance map using machine learning techniques. Specifically, we study the accuracy and efficiency of various regression models and Artificial Neural Networks in modelling the compressor pressure ratio, given the mass flow rate and rotational speed of the centrifugal compressor. Among the various regression models, we focused on the Gaussian Process Regression, which gives well fitting of pressure ratio, in both interpolation and extrapolation cases. We also developed an ANN for the centrifugal compressor. It is shown
that ANN perform well in interpolation and extrapolation tests. High accuracy can be achieved in a few milliseconds of computational time. We also studied the data augmentation effects using the developed ANN. Due to the inherent nature of pressure ratio data distribution in mass-flow rate and rotational-speed space, augmentation in number of rotational speed is more effective than mass flow rate data augmentation to improve the ANN performance. The ineffectiveness of data augmentation in the number of mass flow rate actually indicates that less data of mass flow rate is needed. We therefore performed mass flow rate data minimization. It is found that only 10 or 7 mass flow rate data on each rotational speed line are needed in the development of the model. We further performed principal component analysis and confirmed that the dominate data feature is related to the compressor speed line. Therefore, it suggests that to improve the model accuracy, data augmentation should be performed on the data related to the dominate feature. For the data related to the non-dominate feature, very little data is needed in the development of the model.

ACKNOWLEDGMENT
Pau CV thanks Dr. Samir Kanaan-Izquierdo and Dr. Gerhard Escudero Bakx for their cooperation in this project. Pau also would like to thank Sr. Pete J. Balsells, the University of Colorado and the Generalitat de Catalunya for the Balsells research fellowship. Hui Wan thanks the support from 2018 Air Force Research Lab Summer Faculty Fellowship Program. Hui Wan also thanks the computational resource from XSEDE program DMS180004.

This work is partially supported by U.S. Government. Approved for public release; distribution is unlimited.

REFERENCES
FIGURE 12. PCA of the shuffled data. Each color denotes an identified group, which physically corresponds to a rotational speed.


Appendix A
Optimized models and their hyper-parameters:

KNeighborsRegressor(n_neighbors=3, weights='distance')
BayesianRidge(alpha_1=1e-07, lambda_1=1e-05, n_iter=100)
DecisionTreeRegressor(criterion='mae', max_depth=7)
GradientBoostingRegressor(max_depth=5, n_estimators=120)
KernelRidge(alpha=1e-05, degree=2, gamma=10, kernel='rbf')
GaussianProcessRegressor(alpha=0.0001, kernel=RBF())
SVR(C=1000, gamma=0.1, kernel='rbf')
AdaBoostRegressor(loss='square', n_estimators=20)