FAULT ANALYSIS OF SHIP MACHINERY USING MACHINE LEARNING TECHNIQUES

Reference NO. IJME769, DOI No: 10.5750/ijme.v164i1.769

A Ak and **F Kaya İnceişçi,** University of Marmara, Turkey

KEY DATES: Submitted: 05/08/22; Final acceptance: 01/03/22; Published 15/06/22

SUMMARY

Estimating the probability of failure of ship systems has become an important issue. The rapid development of information technologies in recent years has made it possible to integrate machine learning techniques into the design of ship systems. Predicting and preventing the failures that may occur in the ship machinery systems can extend the life of ship machinery. The study presented aimed to predict the turbine failure of a ship engine. For this purpose, the data obtained from an LM-2500 type ship engine were analysed using Artificial Neural Networks (ANN) algorithms. These results were compared with the following techniques: linear regression; decision tree regression; "k" nearest neighbours' regression; random forest regression; Bayesian ridge regression; extra tree regression; and linear Support Vector Regression (SVR). The study showed that the ANN method determined the failure prediction of ship machinery with a higher accuracy than the regression methods.

KEYWORDS
Fault analysis, Ship machines, Machine Learning

NOMENCLATURE

b	Bias
N	Number of data
y_i	Prediction value
$\overline{y}$	Average of actual values
$\hat{y}$	Estimated values
$\in$	Element
ϵ_i	Residual value
ξ	Insensitive loss function
Π	Pi
$\sum$	Sum
%	Percent
°C	Temperature

1. INTRODUCTION

The operational needs of military and civilian ships have increased in recent years with the developing technology. In this context, it is expected that the time spent on the ships will increase, but it is desired that the time allocated to maintenance and repair will decrease. This has made it necessary to estimate the downtime of ship machinery. Accordingly, it has become critical to analyse the failure of ship machinery.

Today, the researches on fault analysis have become one of the important research topics in the field of engineering (Tang and Yao, 2005), (Xie et al., 2020). Applications called predictive maintenance in the literature are among the current trends in the maritime field to control the physical condition of ships. Fault analysis is applied in important systems in ships by monitoring the situation and processes. The purpose of condition and process monitoring is to use information from specific ship systems to evaluate the condition of the system under examination (malfunctioning, degraded, etc.) and to detect further fault analysis (Kobbacy and Murthy, 2008), (Mohanty, 2014). Successful implementation of the maintenance process depends on the quality of the data, the system situation assessment and the applicability of various algorithms that can be used for fault detection (Cheliotis, Lazakis and Theotokatos, 2020). On the other hand, ship maintenance affects the reliability and availability levels of the shipping industry and accounts for 20-30% of a ship's operational costs. It is also an important ship life cycle factor that can minimize downtime and reduce operating costs (Lazakis et al., 2018). There are studies that use different methods to predict the failures of ship machinery.

In recent years, there have been very few different studies on machine learning algorithms and fault prediction applications. Ateş et al. used machine learning for fault analysis on ships. In the study, the stress concentrations occurring in welded areas such as transverse stiffener joints or T-joints in ship structures were examined (Ateş, 2020). Cipollini et al. predicted the compressor distortion coefficient from the data obtained from ship machinery using linear regression models. The study proved that it is possible to predict the condition-based maintenance problem in a controlled manner by using regression techniques (Cipollini et al., 2018). Lazakis et al. have studied on detecting incipient failures of ship machinery by modelling the normal behaviour of diesel generators with support vector machines. The methodology developed has demonstrated its applicability in determining normal and abnormal ship engine conditions (Lazakis et al., 2019). Dua et al examined the issue of estimating the failure probability of the compressor degradation coefficient using data from ship machinery research. They used machine learning methods to accurately determine when ship machinery, one of the critical systems for ships, needed maintenance. Data were limited in the research due to factors such as cost and time (Dua, Graff, 2017).

The changes that may indicate the malfunction over the parameters expressing the machine condition of the ship are interpreted in the study of Raptodimos and Lazakis. The data obtained were clustered using a two-step approach. In the first stage, the data were clustered using an ANN, then using the Euclidean metric. The results were compared with data representing the motor operating state and fault detection was interpreted (Raptodimos and Lazakis, 2018). Zhongling et al. were studied on the problem of fault diagnosis in an air-handling unit using ANN (Zhongling, Ning and Yan, 2005). In Zhou and Xu's paper, ANNs were used to identify trends for the marine engine cooling system that were too complex to be noticed by humans or other computer techniques (Zhou and Xu, 2010).

In this study, it is aimed to use ANNs for the detection of faults in ship machinery and to obtain a comparative result with different machine learning results. For this purpose, pressure, speed, temperature, torque, flow and revolution data obtained from LM-2500 type marine engine were used. With these data, the turbine decay probability of the ship engine was estimated using a multilayer feed forward ANN. The dataset was also analysed with linear regression, decision tree regression, nearest neighbour regression, random forest regression, bayesian ridge regression, extra tree regression, linear Support Vector Regression (SVR). The mentioned methods were tested with different parameters and comparative results were presented.

The remainder of this article is organized as follows. In Section II, information about the system is given. The methods used are described in Section III. The results obtained by ANN and regression algorithms are given and compared in section IV. Section V concludes the paper.

2. SHIP MODEL

A diagram of the system is shown in Figure 1. The main engine is the machines that generate the power that keeps the ship moving. All systems other than this purpose are defined as auxiliary. In the past, the systems that provide propulsion by driving propellers or jets and today can be a steam engine, steam turbine, internal combustion engine or gas turbine, depending on the ship type and size. All systems that assist in the operation of the main engine and perform functions other than the advancement of the ship are called auxiliary engines. Examples of auxiliary systems are generators and compressors. The generators that produce the ship's electricity can be stand-alone machines or shaft generators, which usually use power transmitted to the ship's propeller shaft. It is generally a system consisting of a diesel engine and an alternator. A compressor is a machine generally used to compress air or other gases to pressures higher than atmospheric pressure. Here, power is generated, transmitted to the propeller, and the ship is moved by the movement of the propeller.

The database used in the study was collected from the system implemented by a digital simulator of a marine ship (Frigate), which is characterized by a Gas Turbine (GT) propulsion facility. In other words, the data is simulated using data from a real ship. The database used contains 589,223 observations and was developed, tested, and verified in MATLAB Simulink (Therrio, 2018). The different blocks (propeller, casing, GT, gearbox and controller) formed on the simulator have been developed, fine-tuned, and tested at several similar real drive facilities throughout the year (Dua, Graff, 2017). The variables and degradation parameters used in the study are presented in Table 1.

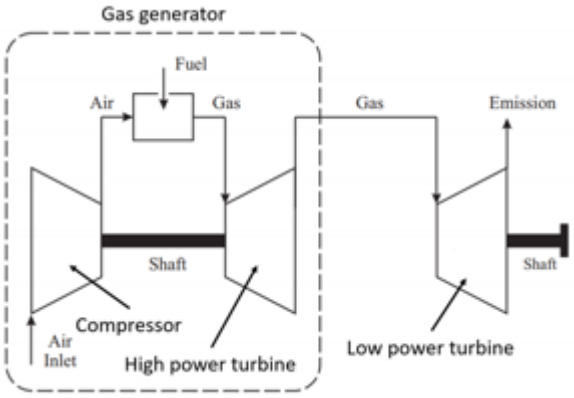


Figure 1: Ship Propeller Systems (Therrio, 2018)

Table 1: Database parameters

Number	Parameters	Units
1	Lever position	[lp]
2	Ship speed	[knot]
3	Gas Turbine shaft torque (GGT)	[kN m]
4	Gas Turbine rate of revolutions (GTn)	[rpm]
5	Gas Generator rate of revolutions (GGn)	[rpm]
6	Starboard Propeller Torque (Ts)	[kN]
7	Port Propeller Torque (Tp)	[kN]
8	HP Turbine exit temperature (T48)	[C]
9	GT Compressor inlet air temperature	[C]
10	GT Compressor outlet air temperature	[C]
11	HP Turbine exit pressure (P48)	[bar]
12	GT Compressor inlet air pressure (P1)	[bar]
13	GT Compressor outlet air pressure (P2)	[bar]
14	Gas Turbine exhaust gas pressure (Pexh)	[bar]
15	Turbine Injecton Control (TIC)	-
16	Fuel flow (mf)	[kg/s]
17	GT Compressor decay state coefficient	-
18	Turbine decay state coefficient	-

In the database, lever position (parameter 1) controls the speed of the ship (parameter 2). TIC is the percent of fuel flow for the propulsion engine (parameter 15). In the engine used, the air first enters the GT in the compressor where the air is pressurized (parameter 12) and warms up (parameter 13). Fuel is added to compressed air (parameter 16) and combustion creates even higher pressure flowing through the high power (HP) turbine. The HP turbine powers (rotates) the gas generator shaft measured in revolutions per minute (RPM) (parameter 5). The gases leaving the HP turbine flow into the low power turbine where the torque is (parameter 3) and speed (parameter 4) is monitored (Therrio, 2018). Parameter 17 and parameter 18 are target parameters. Compressor decay state coefficient and turbine decay state coefficient are predicted by analysing 16 feature vectors. The coefficient for the compressor decay state is optimized between 0.95-1 and the turbine decay state coefficient optimized between 0.975-1.

In this paper, turbine decay state coefficient is used as target parameter. This parameter displays the failure rate in the engine. According to the Free Lunch theory, the properties of the data used should be examined before deciding on the method to be chosen. For this reason, the data were analysed in detail with histograms and graphics. Figure 2 shows the maximum and minimum values of the input components in colours. 16 properties are given in the heat matrix in Figure 2. These 16 features are the 16 features used as inputs in algorithms. In the heat matrix, the dark colour indicates the strong negative relationship between the variables, while the light colour shows the strong positive relationship. In this matrix, the relationship strength of the variables is shown. Using a colour scheme, high values of data are shown in warm colours and low values in cool colours.

In this study, the collected data were divided into two parts as training set and test set to create predictive models. The training set was used to produce the result, and the test set was used to demonstrate the accuracy of the model in predicting the error rate.

3. PREDICTION MODELS

3.1 MACHINE LEARNING METHODS

Machine learning, a branch of artificial intelligence, uses statistics and the computing power of computers to identify complex patterns from data and to make rational decisions. Machine learning techniques are used successfully in classification and regression problems (Carbonell, Mitchell, 1983), (Michie et al., 1994), (Nguyen and Armitage, 2008), (Sebastiani, 2002). While some of the machine learning systems try to completely eliminate the need for human intuition, some take a collaborative approach between human and machine. However, the complete dominance of the person who designed the system over the way data is encoded makes it impossible to completely eliminate human intuition. The machine learning techniques used in this study are explained in the following sections.

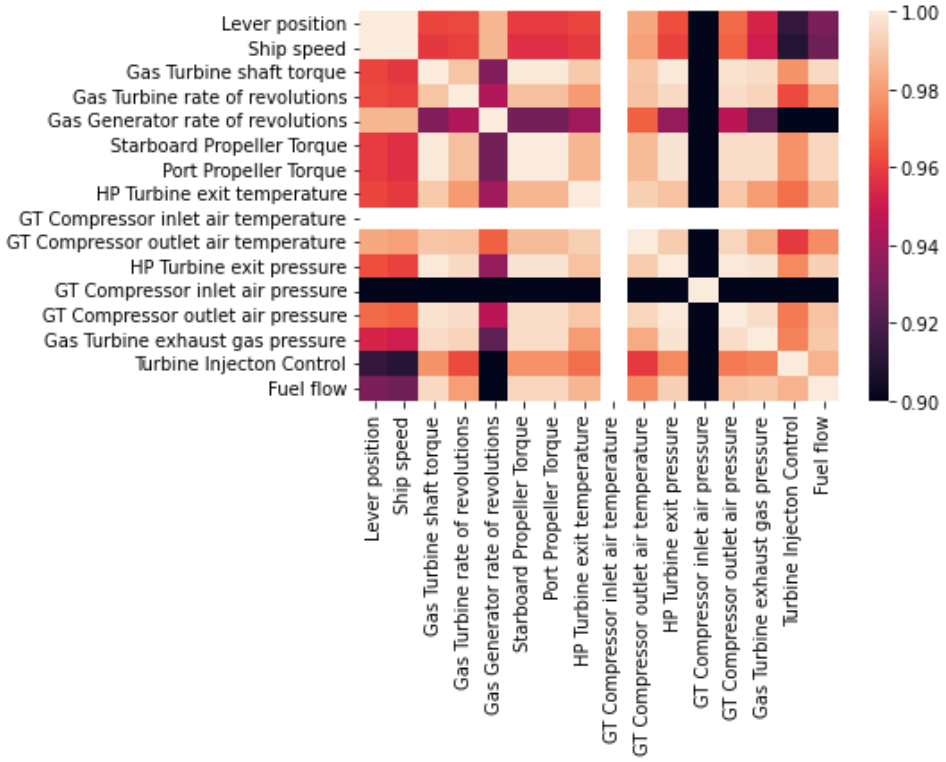


Figure 2: Input parameters received from the ship

3.1.1 Artificial Neural Networks

ANNs can easily address the modelling problems that are analytically difficult and impractical of traditional approaches, including nonlinear, highly ordered, and complex physical processes with time-varying dynamics. Therefore, they can be used for predictive maintenance strategies that can assist decision makers in choosing appropriate maintenance actions for critical ship machinery.

ANN is a machine learning model inspired by the information processing technique of the human brain. There are many nerve cells (neurons) in the human brain connected to each other by synapses. Synapses provide the connection of axons of cells with other dendrites. The nucleus provides periodic reproduction of signals along the axon. The signal carried on the axon is transmitted to the synapses by chemical transporters. A neuron model is shown in Figure 3. There are inputs of a unit, a transfer function processing these inputs, cost function and output value (s) produced (Kamber and Pei, 2006). The neuron is the basic processing unit of artificial neural networks with inputs and outputs. Equation 1 calculates the value of the inputs and outputs of a neuron. Next, the activation function takes place. The activation function is used here to control the y value, that is, to decide whether a neuron will be active or not.

$$\hat{y} = \sum_{i=1}^{n} x_i w_i + b \quad (1)$$

where b is the bias, x is the input to neuron, w is the weights, n is the number of inputs from the incoming layer and i is a counter from 1 to n.

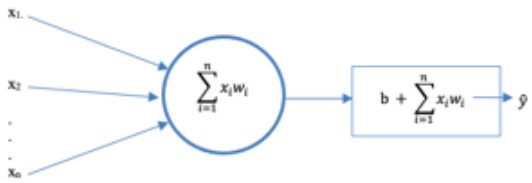


Figure 3: Artificial Neuron Model

Each neuron in ANNs processes the information transmitted to it by other neurons and transmits the result to other neurons. ANN consists of three main layers: input layer, hidden layers and output layer. Information is transmitted to the network via the input layer, they are processed in the hidden layers, and from there they are sent to the output layer. The intention of processing information is to transform the information coming to the network into output by using the weight values of the network. The weights must have the correct values in order for the network to produce the correct outputs for the inputs. ANNs store the information obtained during learning as the connection weights between nerve cells. These weight values contain the information required for ANNs to successfully process the data (Şen, 2004).

There are two approaches for ANNs learning; supervised and unsupervised learning. The most used learning method in ANNs is supervised learning. Within the scope of the supervised learning approach, the dataset to be used for training is determined and applied to the ANN. When the network starts to learn, the learning rate and error analysis are controlled by various control mechanisms, and at a certain point the process is completed. Control mechanisms consist of elements such as network iteration number, learning rate, fault threshold. In the number of iterations, it can be determined how many times the feed forward algorithm will be applied on the current data set. The learning rate indicates at what speed the network will learn and it is generally recommended to take a value between 0.2-0.4 (Tebelskis, 1995), (Kamber and Pei, 2006). After the ANN is trained, it has the ability to generate the relevant response for the examples that it did not encounter during the training. Even if data is entered into a trained ANN in a way that is incomplete, corrupt, or never encountered before, the network will produce the most appropriate output. This feature is the generalization feature of the network (Tebelskis, 1995).

3.1.2 Regression Analysis

Regression analysis is a method used to measure the relationship between two or more variables. If an analysis is made using a single variable, it is called univariate regression, if more than one variable is used, it is called multivariate regression analysis. With regression analysis, information is obtained about the existence of the relationship between variables, and the strength of this relationship, if the relationship exists.

3.1.2 (a) Linear Regression

Linear regression is a method used to model the link between one or more (independent) variables and another (dependent) variable. Linear regression can be formulated as follows:

$$\hat{y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

for each observation $i = 1, ..., n$. Where, $\hat{y}_i$ is the predictor, β_0 predictor of the cut point, β_1 slope predictor of the line and ϵ is the residual value. X matrix of row-vectors or of n-dimensional column-vectors.

3.1.2 (b) Decision Tree Regression

Decision trees are one of the most widely used machine learning methods, and they create classification and regression models in the form of a tree structure according to the structure of the data sets (Kavzoğlu et al., 2012). Decision trees perform a simple decision-making process by transforming the complex data into phases with a multi-stage and sequential approach in solving the problem (Safavian and Landgrebe, 1991). Decision trees consist of 3 parts as root node (Dataset), intermediate node (Tree 1, Tree 2, ...) and the last node (Prediction 1, Prediction 2, ...), as seen in Figure 4.

The basic principle in creating a decision tree structure is to establish decision rules by asking a series of questions about the data. For this process, questions are started to be asked at the root node, which is the basic element of the tree structure, and the growth or branching of the tree continues until the leaves, the last element of the tree structure, are reached (Pal and Mather, 2003).

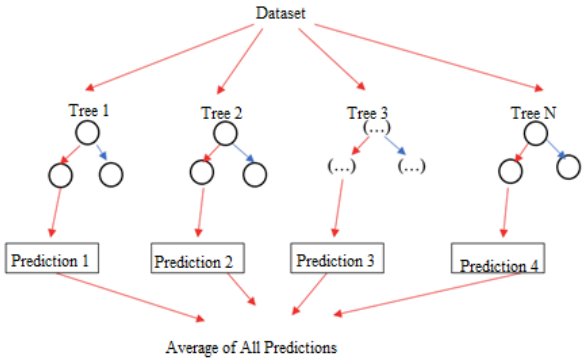


Figure 4: Decision Tree

3.1.2 (c) K Nearest Neighbours Regression

K Nearest Neighbours Regression (KNN) algorithm were proposed in 1967 by T. M. Cover and P. E. Hart. The algorithm is used by making use of data from a sample set whose classes are known. In the KNN algorithm, the samples in the training set are specified with n-dimensional numerical properties. When an unknown sample is encountered, k closest samples are determined from the training set and the class label of the new sample is assigned according to the majority vote of the class labels of the k closest neighbours (Kamber and Pei, 2006). The KNN algorithm keeps the training data and the position of the classes of these data. When it is necessary to decide in which class the new incoming data will be, the distance between the query data and the training samples is calculated (Moosavian et al., 2013). The nearest neighbours from the relevant distances are considered. It is assigned to the class of k neighbour or neighbours according to the attribute values. The selected class is considered as the class of the observation value expected to be predicted. In other words, the new data is labelled. The Fig. 5 shows a case with two classes (circles) and two properties (triangle and circle). The class to which the data belongs is found by looking at the distance from the centre of the sphere.

The KNN algorithm, is a simple and easy-to-implement method to calculate distances in a multi-dimensional input space, which can yield competitive results even when compared to the most complex machine learning methods (Moosavian et al., 2013). The distance value in the KNN algorithm is calculated with the following formula:

$$d(A,B) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

where A and B are the points to measure the distance. The vectors A and B can be expressed by $A = x_1, x_2, \ldots, x_k$ and $B = y_1, y_2, \ldots, y_k$ where k is the dimensionality of the feature space.

Figure 5: K Nearest Neighbours (Moosavian et al., 2013)

3.1.2 (d) Random Forest Regression

Random Forest Regression (RFR) (Fig 6) structure is derived from decision tree classifiers and involves classification and regression trees. Forest structures are obtained by bringing random trees together (Yang, Di and Han, 2008). First step includes a tree built by nodes with recursive splitting and many different decision trees are formed to calculate a response variable. All terminal nodes are assigned to certain classes and all responses are evaluated from all trees. The most predicted class is assigned for the object. Then second random sampling determines the splitting of each node in the tree. Predictor variables, which can be chosen by user or RFR algorithm itself, are randomly selected to create binary rule at each node. Therefore, randomization provides less correlations and lower fault analysis among trees (Breiman, 2001), (Horning et al., 2010).

The advantage of the RFR algorithm is the "pruning", which refers to the removal of terminal nodes where the growth is large and with very small subsets of training data, possibly noise, to simplify the tree. Interpretation is also easy and classification is fast when the rules are established. According to Liaw & Wiener (Liaw et al., 2002) the most common parameters in RFR applications are input training data, number of trees, the number of predictor variables, and parameters to calculate fault and variable significance.

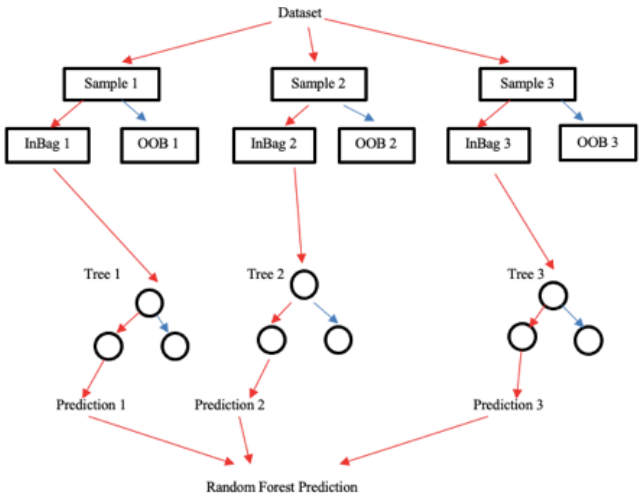


Figure 6: Random Forest Regression (Ahmad, Reynolds and Rezgui, 2018)

3.1.2 (e) Bayesian Ridge Regression

Ridge regression is an alternative to subset selection. It is used to predict the coefficients more accurately compared
to the least squares method for some cases. Ridge regression allows reducing the number of predictors in the model (Uyanık et al., 2020). The coefficient predictions are shifted closer (or equal) to zero (Whittaker et al., 2000). The Bayesian interpretation of ridge regression predictor is given below (Şamkar et al., 2011), (Uyanık et al., 2020):

$$y = X\beta + \epsilon \tag{4}$$

where Here y is the variable vector, X is the variable matrix, β is the parameter vector and ϵ is the error vector.

Due to multicollinearity among independent variables, there will be interference in parameter predictions in regression model. To overcome this drawback and yield more precise regression parameter predictions, ridge regression will induce small bias (Efendi, Effrihan, 2017), (Shi et al., 2016).

3.1.2 (f) Extra Tree Regression

Extra trees algorithm which is also called extremely randomised trees is a recent machine learning technique and developed by Geurts et al. as an extension of RFR (Geurts et al., 2006). Extra tree uses random subset to train predictors as RFR. On contrary, the optimal cutting point is also achieved by randomization in extra tree regression. Extra tree uses whole training dataset to train each tree while RFR uses bootstrap replica to train the model (Ahmad, Reynolds and Rezgui, 2018), (Aláiz-Moretón vd., 2019). The prediction in extra tree is achieved by averaging all probability classes and choosing the class with the highest probability, which is termed "majority voting". The advantages of the process are better results for complex problems and reduction of computational burden (Patil and Phalle, 2018). The parameters can be adapted to different problems manually by default setting to maximize computational advantages and autonomy (Nistane and Harsha, 2018).

3.1.2 (g) Linear SVR Regression

SVR is a common regression method and extended from support vector classification (Vapnik, 2013). It provides flexibility to models for defining how much error can be tolerated and for determining appropriate line to fit the data. SVR minimizes structural risk solve the overfitting problem effectively (Lin et al., 2007). In SVR, kernel function is employed to map the data to a higher dimensional space. Linear SVR finds a model (w) where $w^T x_i$ is close to minimized value y_i. Considering a set of training pairs $\{(x_i, y_i)\}$, $x_i \in R^n, y_i \in R,\ i = 1, \dots, l$ SVR solves regularized optimization problem, which can be shown as:

$$\min_w f(w),\ where\ f(w) \equiv \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_\epsilon(w; x_i, y_i)$$

where C > 0 is the regularization parameter, and ξ_ϵ, is the ϵ insensitive loss function associated with (x_i, y_i) (Ho and Lin, 2012):

$$\xi_\epsilon(w; x_i, y_i) = \begin{cases} max(|w^T x_i - y_i| - \epsilon, 0)\ \ or \\ max(|w^T x_i - y_i| - \epsilon, 0)^2 \end{cases} \tag{6}$$

SVR is referred to use both equations as L1-loss and L2-loss SVR, respectively where L2-loss is differentiable. Once the first equation is minimized, the prediction function is $w^T x$ (Ho and Lin, 2012).

There is also a bias (b) involves in SVR and the prediction function becomes $w^T x + b$ where b can be obtained by substitution of a support vector (Lin et al., 2007). Researches on large-scale linear classification generally omit the b due to negligible effects on the performance for most data (Ho and Lin, 2012).

3.2 ERROR METRICS

In this study, different error criteria as Mean Square Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Coefficient of determination R-Squares (R^2) were used to evaluate the predictive performance of models. MAE and MSE can take values from 0 to ∞. It was determined that the lower the values obtained from the calculations of these ratios, the more successful the prediction algorithm. When the accuracy of the models is evaluated with R^2, the closer the R^2 ratio is to 1, the higher the success of the prediction model (Uyanık et al., 2020). In the MSE, MAE, MAPE and R^2 formulas given below, $\hat{y}$ is the predicted value of y, y_i is the real values and n is the number of instances.

3.2.1 Determination of Coefficient R-Squares

R^2 is a statistical calculation that how close the data is to the regression line. It is computed a ratio of the variance between independent variables and dependent variables:

$$R^2 = - \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{(\overline{y} - y_i)^2}$$

3.2.2 Mean Absolute Error (MAE)

The absolute mean distance between the predicted data and the actual data is called the MAE.

$$MAE = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i|$$

3.2.3 Mean Absolute Percentage Error (MAPE)

MAPE is called to measure the accuracy of predictions in regression and time series models.

$$MAPE = \frac{100}{n} \sum_{i=0}^{n} \frac{|y_i - \hat{y}_i|}{|\hat{y}_i|} \quad (9)$$

3.2.4 Mean Square Error (MSE)

The mean square distance between real data and predicted data and real data is called MSE.

$$MSE = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2 \quad (10)$$

4. RESULTS AND DISCUSSION

ANN with different numbers of layers and neurons have been tried and it has been decided to use a model that includes two hidden layers and uses the RELU activation function. There are 16 cells in the input layer and a single cell in the output layer of the ANN model. The visual of the model is shown in Figure 7. When the number of layers and neurons was increased too much, the response accuracy of the system was not affected much, but the training time of the ANN was prolonged. In order to determine the best performer among the models, Linear Regression, Decision Tree Regression, Knn Regression, Random Forest Regression, Bayesian Ridge Regression, Extra Tree Regression, Linear SVR Regression methods were also used for failure analysis. Thus, prioritization between models is made according to the success and accuracy of the prediction. Structural differences are taken into account when determining different machine learning methods.

The studies were carried out with Google Colab software. Scikitlearn library of Python programming language is used to train the model. The adjustment of the parameters of the used methods is provided by computer simulations.

For the correct training of the models, the data set is divided into 75% as training and 25% as testing. Thus, approximately 9000 samples were randomly determined from 11934 samples in order to learn the properties of the data set, and approximately 3000 samples were used to test the models.

ANN model was also tested for different activation functions. Firstly relu activation function is used for input and output layer. Secondly, sigmoid activation function is used in the output layer and the relu activation function in the input layer. The results obtained by training the model were compared with the results obtained by regression models.

The values for the linear regression model are given in Table 2. Here it is seen how close the predicted values are to the real values.

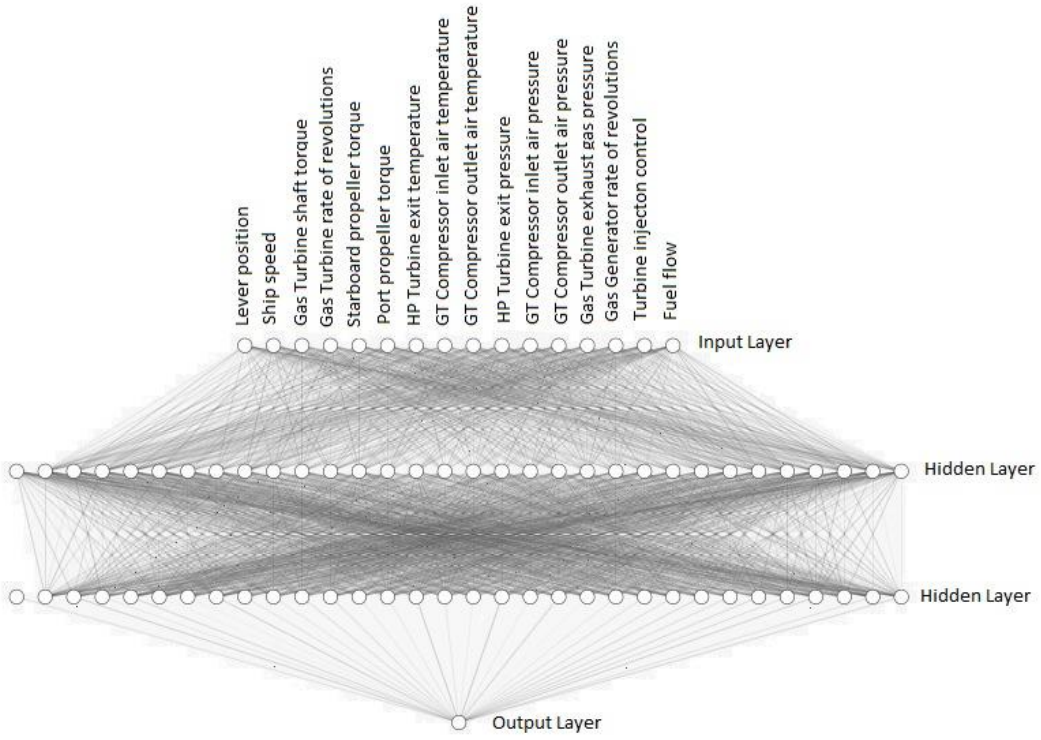


Figure 7: Artificial Neural Network Model

Table 2: Linear Regression Parameters

Formula	Value
$\hat{y}$	0.98
β_0X	2.57173464e-05
β_1X	1.66765328e-01
β_2X	1.01776267e+00
β_3X	2.33896897e-01
β_4X	3.09837366e-01
β_5X	-5.14176470e-01
β_6X	-5.14176470e-01
β_7X	-7.66944202e-01
β_8X	-3.05311332e-16
β_9X	1.50267883e-01
β_{10}X	3.65410084e-01
β_{11}X	-2.22044605e-16
β_{12}X	-1.24237071e+00
β_{13}X	9.47344529e-01
β_{14}X	-2.16719124e-02
β_{15}X + ϵ	6.58664494e-01

Table 3: Bayesian Ridge Parameters

Formula	Value
$\hat{y}$	0.98
$X\beta_0$	-7.19823104e-04
$X\beta_1$	1.67353585e-01
$X\beta_2$	1.01766177e+00
$X\beta_3$	2.33678689e-01
$X\beta_4$	3.09520073e-01
$X\beta_5$	-5.14014555e-01
$X\beta_6$	-5.14014555e-01
$X\beta_7$	-7.66799753e-01
$X\beta_8$	7.15993278e-13
$X\beta_9$	1.50195437e-01
$X\beta_{10}$	3.66946499e-01
$X\beta_{11}$	1.19271441e-12
$X\beta_{12}$	-1.24231073e+00
$X\beta_{13}$	9.31432544e-01
$X\beta_{14}$	-2.16327018e-02
$X\beta_{15}$ + ϵ	6.57945556e-01

In the decision tree algorithm, the random state value is chosen as 0, 10 and 20. Trials have been made for these values. Max depth value is not used for decision trees algorithm in the study. The parameter used in the KNN algorithm is the n neighbour parameter, which determines the number of neighbours, and 3 is taken as the value that gives the best k value. The number of trees in the model is taken as 20 for RFR. The predictive values obtained in the bayesian ridge algorithm are given in Table 3. Random state value 0, 20 and 40 are taken for extra tree algorithm. An ensemble model generates completely original learning sample in many times.

Hyperparameter 1, 2 and 3 were tested in the study. Hyperparameter 1 gave better results. So, hyperparameter 1 values used while training the model (Table 4). The created models are used to obtain the estimated value of each data of the investigated fault. The results showing the accuracy of the performed models and the accuracy of all eight methods are given in Table 5. When the learning rate in ANN was taken as 0.001, the results in Table 5 were obtained. While the ANN algorithm gave the highest accuracy with a rate of 98%, the linear SVR algorithm gave the lowest accuracy with a rate of 86%. Decision Tree, Random Forest Regressions and extra tree regression algorithms gave the best results after ANN with a score of 97%.

Table 4: Models and Hyperparameters

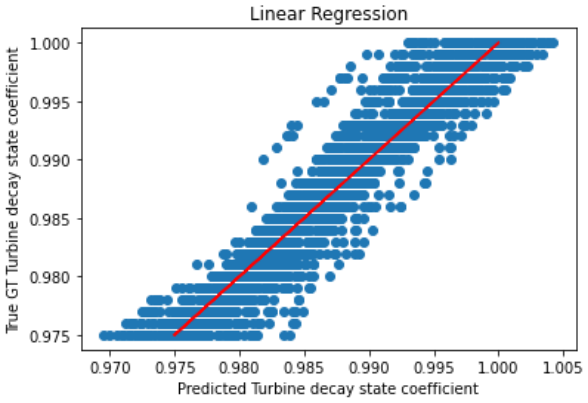
Model	Hyperparameters 1	Hyperparameters 2	Hyperparameters 3
Linear Regression	-	-	-
Decision Tree Regression	random state=0	random state=20	random state=30
KNN Regression	n neighbours=3	n neighbours=10	n neighbours=20
Random Forest Regression	n predictors=1	n predictors=2	n predictors=3
Bayesian Ridge Regression	compute score=True	compute score=False	-
Extra Tree Regression	random state=0	random state=20	random state=40
Linear SVR Regression	random state=10	random state=20	random state=30
ANNs Model	input dim=16	input dim=16	input dim=16
	activation='relu'	activation='sigmoid'	activation='swish'

Table 5: Models and error metrics

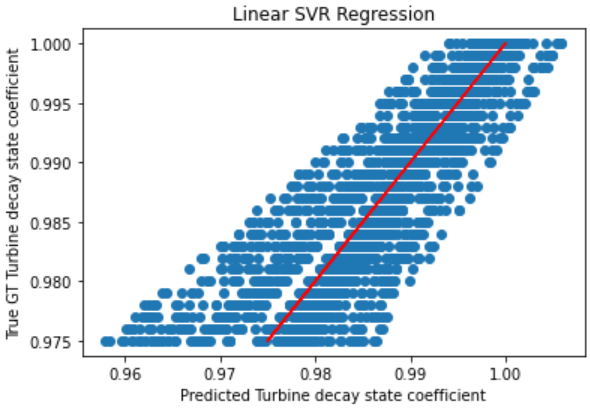
Model	MAE	MSE	R^2	MAPE
Linear Regression	0.0017	5.09	0.91	0.16
Decision Tree Regression	0.00052	1.075	0.97	0.053
KNN Regression	0.0010	5.09	0.91	0.011
Random Forest Regression	0.00038	5.27	0.97	0.038
Bayesian Ridge Regression	0.0016	5.09	0.91	0.16
Extra Tree Regression	0.00058	1.67	0.97	0.05
Linear SVR Regression	0.0021	7.54	0.86	0.21
ANNs Model	0.0008	9.59	0.98	0.87

Tests were also conducted for different values of the learning rate. Accuracy did not change much in increments of the ratio up to 0.07. When the learning rate was increased to 0.07, the accuracy decreased to 97.66%. Since the weights in the ANN were randomly assigned at the beginning, different results were obtained for the same learning rate in the experiments. However, in no case did the accuracy go below 97.66%.

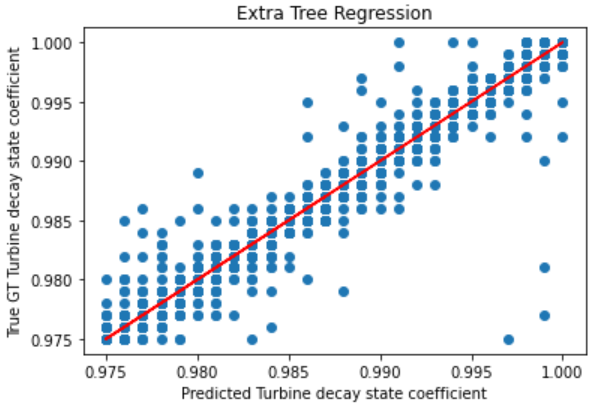
Data set and prediction modelling were performed successfully according to randomly determined samples. Accuracy graphs of these models are given in Figure 8. The actual and predicted fault analysis of the models are visualized in the graphics. The points shown on the y-axis in the graph show the actual fault analysis values and the points shown on the x-axis show the predicted fault analysis values. The line shown in red is a straight line. It is drawn by determining the minimum distance between the predicted values and the actual values. If the actual values and predicted values in this row are close to each other, this indicates that the accuracy is high. The best predictive models are ANN, Decision Tree and Random Forest Regressions.



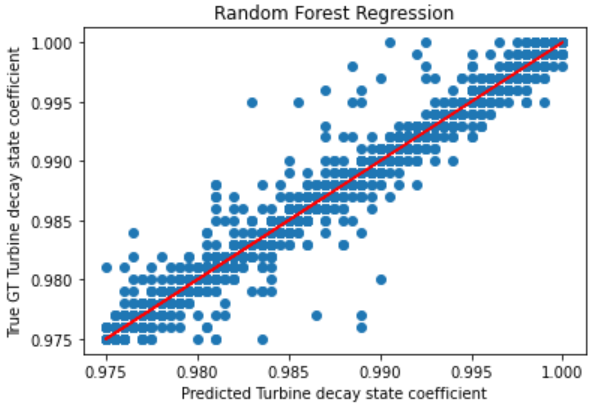
(a) Linear Regression



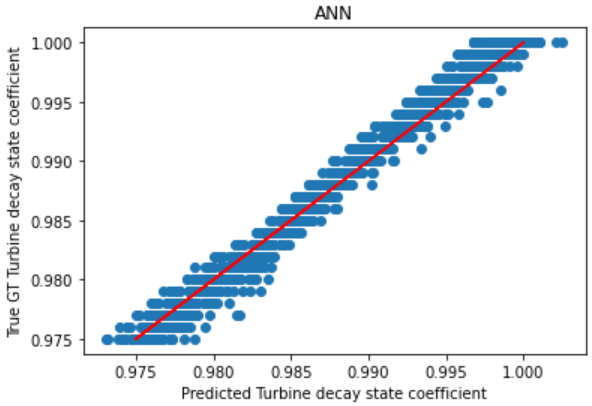
(b) Linear SVR Regression



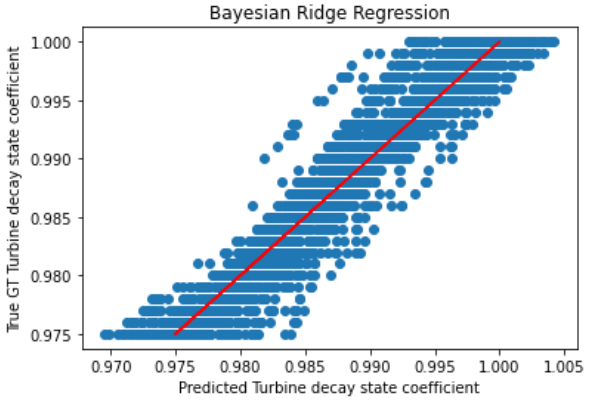
(c) Extra Tree Regression



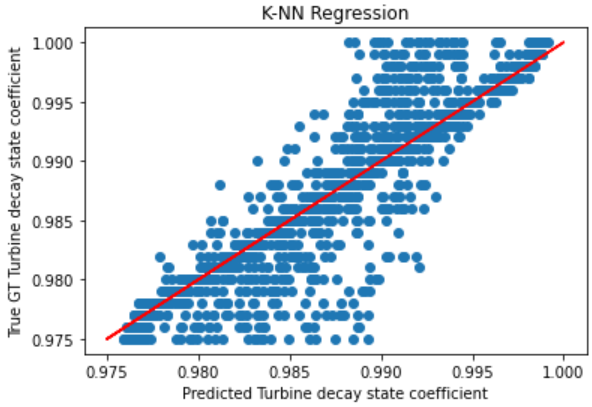
(d) Random Forest Regression



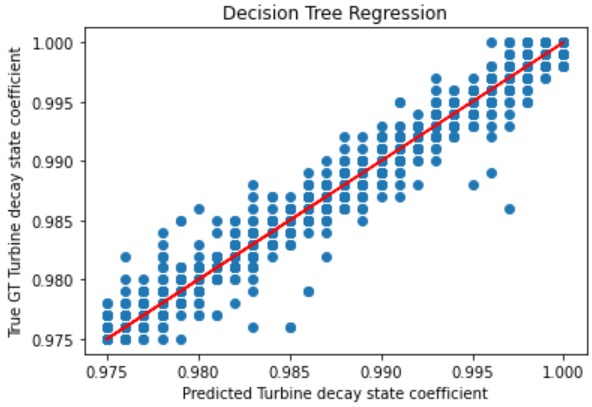
(e) ANNs Regression



(f) Bayesian Ridge Regression



(g) KNN Regression



(h) Decision Tree Regression

Figure 8: Accuracy Graphs

5. CONCLUSION

In this study ANNs and regression analysis methods were used to prediction the turbine decay state coefficient with the data obtained from ship machinery. The data collected from the ship engine system were used to determine the inputs of ANN models and regression. So, the most appropriate method for failure analysis of ship machinery determined. The number of data used in the methods studied and the meaningful relationship of the data with each other greatly affected the accuracy of the results of the methods. Results obtained with different models of ANN algorithm were more successful than regression analysis results. Among the prepared models, as the results of the study suggest ANNs have been the most successful model.

From these results, it is understood that ANNs give more successful results than regression analysis in prediction models prepared with high amount and partially related data. ANNs are preferred because they give more reliable results compared to other algorithms. In the study, the extent to which the performance will change was investigated by using different models. Network parameters such as the number of hidden layers and the number of neurons in the hidden layer were determined as a result of a number of experiments. The ANN algorithm, which is estimated by the failure coefficient obtained from the simulated gas turbine engine using real data, has been successfully modeled.

With some changes in the network topology, such as increasing the number of hidden layers, changing the number of neurons in the layers, and changing the transfer functions used in the network, the predictions obtained with ANNs can be further improved. As a result, it can be seen that the model prepared with ANNs can successfully model the non-linear ship-machine fault relationship, allowing a rapid modelling without the need for intensive data and calibration processes compared to conceptual and physically based models.

6. REFERENCES

1. AHMAD, M.W., REYNOLDS, J., REZGUI, Y. (2018). *Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees.* Journal of cleaner production, 203, 810-821.
2. ALÁIZ-MORETÓN, H., CASTEJÓN-LIMAS, M., CASTELEIRO-ROCA, J.L., JOVE, E., FERNÁNDEZ ROBLES, L., CALVO-ROLLE, J.L. (2019). *A fault detection system for a geothermal heat exchanger sensor based on intelligent techniques.*
3. ATEŞ, B., (2020). *Realization of stress concentration predictions in ship structures with coarse mesh structure and machine learning.* Institute of Science and Technology.
4. BREIMAN, L. (2001). *Random forests.* Machine learning, 45(1), 5-32.
5. MİCHALSKİ, R. S., CARBONELL, J. G. and LEARNİNG, T. M. M. M. (1983). *An artificial intelligence approach.* Understanding the Nature of Learning.
6. CHELIOTIS, M., LAZAKIS, I., THEOTOKATOS, G. (2020). *Machine learning and data-driven fault detection for ship systems operations.* Ocean Engineering, 216, 107968.
7. CIPOLLINI, F., ONETO, L., CORADDU, A., MURPHY, A.J., ANGUITA, D. (2018). *Condition-based maintenance of naval propulsion systems with supervised data analysis.* Ocean Engineering, 149, 268-278.
8. DUA, D., GRAFF, C. (2017). *Uci machine learning repository. URL: http://archive.ics.uci.edu/ml.*
9. EFENDI, A., EFFRIHAN. (2017). *A simulation study on Bayesian Ridge regression models for several collinearity levels.* In AIP Conference Proceedings, 1, 020031. AIP Publishing LLC.
10. GEURTS, P., ERNST, D., WEHENKEL, L. (2006). *Extremely randomized trees.* Machine learning, 63(1), 3-42.

11. HO, C.H., LIN, C.J. (2012). *Large-scale linear support vector regression.* The Journal of Machine Learning Research, 13(1), 3323-3348.
12. HORNING, N., et al. (2010). *Random forests: An algorithm for image classification and generation of continuous fields data sets*. Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan.
13. KAMBER, M., PEI, J. (2006). *Data Mining.* Morgan kaufmann.
14. KAVZOĞLU, T., ŞAHIN, E.K., ÇÖLKESEN, İ. (2012). The use of regression trees in the analysis of landslide susceptibility: The case of Trabzon. Map Magazine, 147(3), 21-33.
15. KOBBACY, K.A.H., MURTHY, D.P. (2008). *Complex system maintenance handbook.* London: Springer.
16. LAZAKIS, I., GKEREKOS, C., THEOTOKATOS, G. (2019). *Investigating an svm-driven, one-class approach to estimating ship systems condition.* Ships and Offshore Structures, 14(5), 432-441.
17. LAZAKIS, I., RAPTODIMOS, Y., VARELAS, T. (2018). *Predicting ship machinery system condition through analytical reliability tools and artificial neural networks.* Ocean Engineering, 152, 404-415.
18. LIAW, A., WIENER, M., et al. (2002). *Classification and regression by randomforest.* R news, 2(3), 18-22
19. LIN, K., LIN, Q., ZHOU, C., YAO, J. (2007). *Time series prediction based on linear regression and svr.* Third International Conference on Natural Computation (ICNC 2007), IEEE.
20. MICHIE, D., SPIEGELHALTER, D.J., TAYLOR, C.C. (1994). *Machine learning, neural and statistical classification.*
21. MOHANTY, A.R. (2014). *Machinery condition monitoring: Principles and practices.* CRC Press.
22. MOOSAVIAN, A., AHMADI, H., TABATABAEEFAR, A., KHAZAEE, M. (2013). *Comparison of two classifiers; k-nearest neighbor and artificial neural network, for fault diagnosis on a main engine journal-bearing.* Shock and Vibration 20.
23. NGUYEN, T.T., ARMITAGE, G. (2008). *A survey of techniques for internet traffic classification using machine learning.* IEEE communications surveys & tutorials, 10(4), 56-76.
24. NISTANE, V., HARSHA, S. (2018). *Performance evaluation of bearing degradation based on stationary wavelet decomposition and extra trees regression.* World Journal of Engineering.
25. PAL, M., MATHER, P.M. (2003). *An assessment of the effectiveness of decision tree methods for land cover classification.* Remote sensing of environment, 86(4), 554-565.
26. PATIL, S., PHALLE, V. (2018). *Fault detection of anti-friction bearing using ensemble machine learning methods.* International Journal of Engineering, 31(11), 1972-1981.
27. RAPTODIMOS, Y., LAZAKIS, I. (2018). *Using artificial neural networkself-organising map for data clustering of marine engine condition monitoring applications.* Ships and Offshore Structures, 13(6), 649-656.
28. SAFAVIAN, S.R., LANDGREBE, D. (1991). *A survey of decision tree classifier methodology.* IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.
29. SAMKAR, H., ALPU, Ö., & ALTAN, E. (2011). AN APPLICATION ON USING M FORECASTERS IN RIDGE REGRESSION. Journal of Dokuz Eylul University Faculty of Economics and Administrative Sciences, 26(1).
30. SEBASTIANI, F. (2002). *Machine learning in automated text categorization.* ACM computing surveys (CSUR), 34(1), 1-47.
31. ŞEN, Z., 2004.
32. SHI, Q., ABDEL-ATY, M., LEE, J. (2016). *A bayesian ridge regression analysis of congestion's impact on urban expressway safety.* Accident Analysis & Prevention, 88, 124-137.
33. TANG, T., YAO, G. (2005). *A fault-tolerant control method based on adaptive fnn for ship control system.* In 2005 International Conference on Control and Automation (Vol. 2, pp. 930-935). IEEE.
34. TEBELSKIS, J. (1995). *Speech recognition using neural networks.* Ph.D. thesis. Carnegie Mellon University.
35. THERRIO, E.A. (2018). *Machine Learning Techniques for Development of a Condition-Based Maintenance Program for Naval Propulsion Plants. Technical Report.* Naval Postgraduate School Monterey United States.
36. UYANIK, T., KARATUĞ, Ç., ARSLANOĞLU, Y. (2020). *Machine learning approach to ship fuel consumption: A case of container vessel.* Transportation Research Part D: Transport and Environment, 84, 102389.
37. VAPNIK, V. (2013). *The nature of statistical learning theory*. Springer science & business media.
38. WHITTAKER, J.C., THOMPSON, R., DENHAM, M.C. (2000). *Markerassisted selection using ridge regression.* Genetics Research, 75(2), 249-252.
39. XIE, T., WANG, T., HE, Q., DIALLO, D., CLARAMUNT, C. (2020). *A review of current issues of marine current turbine blade fault detection.* Ocean Engineering, 218, 108194.

40. YANG, B.S., DI, X., HAN, T. (2008). *Random forests classifier for machine fault diagnosis*. Journal of mechanical science and technology, 22(9), 1716-1725.
41. ZHONGLING, L., NING, J., YAN, Z. (2005). *The study on application of ann in fault diagnosis of air conditioning system.* Refrigeration and Air-conditioning.
42. ZHOU, J., XU, L. (2010). *The fault diagnosis of marine engine cooling system based on artificial neural network (ANN)*. The 2nd International Conference on Computer and Automation Engineering (ICCAE), IEEE.