INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Environment, Workplace and Employment

ISSN online: 1741-8445 - ISSN print: 1741-8437
https://www.inderscience.com/ijewe

Comparison of machine learning methods using time series data: focusing on inverter data

Sang-Ha Sung, Chang Sung Seo, Min Ho Ryu, Sangjin Kim

DOI: 10.1504/IJEWE.2023.10057599

Article History:

Received:	08 March 2022
Last revised:	20 March 2022
Accepted:	12 January 2023
Published online:	19 July 2023

Comparison of machine learning methods using time series data: focusing on inverter data

Sang-Ha Sung

Department of Management Information Systems,
Dong-A University,
Busan 49236, South Korea
Email: sangha@donga.ac.kr

Chang Sung Seo

SCT,
Busan 48059, South Korea
Email: sct@esct.co.kr

Min Ho Ryu and Sangjin Kim*

Department of Management Information Systems,
Dong-A University,
Busan 49236, South Korea
Email: ryumh12@dau.ac.kr
Email: skim10@dau.ac.kr
*Corresponding author

Abstract: In this study, we use inverter data to understand the inverter status and present a predictive model for the future status. Data was collected from the inverter through the sensor, and was collected for about two months from July to August 2020, for a total of 8,954,665 time points. The used data consists of frequency and leak level, and when the value of data increases significantly, it is classified as having an abnormality in the inverter. In this study, we present a time series prediction model that can predict inverter status abnormalities by comparing various machine learning techniques. In this study, the inverter state was predicted using the boosting method, the tree method, the SVR method, and the deep learning method. As a result of the experiment, the error rate of the deep learning technique was the lowest.

Keywords: time series data; machine learning; boosting methods; tree methods; SVR methods; deep learning; regression; inverter data; predict; failure predict.

Reference to this paper should be made as follows: Sung, S-H., Seo, C.S., Ryu, M.H. and Kim, S. (2023) 'Comparison of machine learning methods using time series data: focusing on inverter data', *Int. J. Environment, Workplace and Employment*, Vol. 7, No. 1, pp.13–33.

Biographical notes: Sang-Ha Sung is a PhD candidate at Dong-A University. He is majoring in management information systems and has a Master's degree from Dong-A university. His research interests are data analytics, deep learning and machine learning modelling.

Chang Sung Seo is the CEO of SCT. He was a PhD candidate in Technology Business Policy, Pusan National University. His research interests are IoT and smart factory.

Min Ho Ryu is currently a Professor in the Department of Management Information Systems, Dong-A University. He served as a Professor at the Graduate School of Technology Management at Hoseo University and as a Postdoctoral Fellow at Michigan State University, USA. His research interests are big data and IT management.

Sangjin Kim is currently a Professor in the Department of Management Information Systems, Dong-A University. He served as an Assistant Professor at the Department of Mathematical Sciences at the University of Texas at El Paso and as a Postdoctoral at the Department of Biostatistics and Bioinformatics at Duke University. His research interests are the development of algorithms and application with machine learning and deep learning.

1 Introduction

As the era of the 4th industrial revolution has recently emerged, many companies are showing interest in cutting-edge technologies such as the internet of thing (IOT), artificial intelligence (AI) and big data analysis for process productivity and stability. In a smart factory to which these advanced technologies are applied, processes are automated by intelligent robots, and various data are collected through sensors (Lee, 2019). Smart factories require a new management method that is different from the existing ones. In the past, appropriate measures were taken by the discretion of the manager to prevent failure of factory equipment in advance. However, if the condition of the equipment is wrongly judged, a serious failure of the equipment may occur or excessive maintenance costs may occur. This situation has a fatal impact on productivity and quality. To prevent it in advance, data-based methodologies for machine state management are being studied (Lee et al., 2008). In a manufacturing environment, mechanical equipment has been operated for a long time, and failure can occur due to various factors (Oh and Huh, 2020). In addition, the more complex the process, the more difficult the prediction becomes, so it is important to select an appropriate methodology (Cheon and Yang, 2020).

In this paper, we use inverter data to figure out the inverter status and present a predictive model for the future status. The inverter data used in this study was collected from the actual process and collected from July to August 2020. About 40 to 70 data per minute are being sensed, and preprocessing for the analysis was performed. The data generated by the inverter consists of frequency and leak level, and when the value of the generated data increases significantly, the inverter is classified as having an abnormality. If an inverter malfunction or failure cannot be predicted in advance, an emergency failure occurs, which leads to an increase in maintenance costs. The maintenance period may also be increased due to an emergency failure. In order to solve the problem, it is very important to predict equipment failure in advance. Therefore, the goal of the study builds an optimal prediction model that can predicts inverter abnormalities by comparing various machine learning methodologies. The machine learning technique used in this study can be divided into boosting methods, tree methods, SVR methods, and deep learning methods. The prediction model of inverter failure is generated by using the

model typically used in each analysis technique. Twelve models were used to predict the inverter data, and then the performance of the models was compared through RMSE and MAE and prediction of 12 points from the test sets.

The structure of the study is as follows. Section 1 describes the necessity and purpose of the study. Section 2 describes failure prediction and related research. Section 3 describes the methodology used in the study. Section 4 describes the data collection and preprocessing process. Section 5 describes the experimental results. Section 6 derived the discussion and conclusion of the study.

2 Related work

As the complexity of the process increases and various factors become data, various machine learning techniques are being studied to analyse it. Representative machine learning methods include boosting methods, tree methods, support vector machine (SVM) methods, and deep learning methods. In particular, many studies have been conducted to predict machine and equipment failures using the tree method and the SVM method (Sapankevych and Sankar, 2009; Tang et al., 2020), and recently, studies using the deep learning methods together have been also appearing (Raj and Ananthi, 2019). In the study, we examine the recent research trends for each technique and select an appropriate model to conduct research.

Wu et al. (2010) analysed electronic health record (EHR) using boosting and SVM methods. In the case of the study, medical analysis was performed, but the results of the study confirm that patients can be accurately classified through the boosting technique. In addition, extreme gradient boost (XGBoost), and light gradient boost model (LGBM), which are one of the boosting techniques, have been recently widely used in various data analysis competitions and turned out to have better performance in terms of speed and accuracy (Chen and Giestrin, 2016).

In addition, a lot of research using the tree methods has been also in progress. Wu et al. (2017) predicted tool wear through data analysis provided in the PHM 2010 challenge. Among the machine learning methods such as artificial neural network (ANN), support vector regression (SVR), and random forest (RF) were used, and as a result of the analysis, RF showed the highest accuracy. Lee et al. (2019) conducted a study to predict bearing failure through several machine learning models. In the study, various methods such as SVM, K-nearest neighbour (KNN), and deep neural network (DNN) were used, and the optimal model were selected in terms of accuracy, specificity, and F1 score. The model with the highest F1 score was selected as the bagged tree model using the bagging technique (Lee et al., 2019). Mathew et al. (2018) compared various methodologies to predict the lifespan of a turbofan engine, and showed that RF had the highest accuracy. In the study of Tran et al. (2008), the prediction of failure rate was performed through the regression tree, and the error rate was very low.

Liu et al. (2019) analysed 14 acoustic data generated by the turbine using the SVR technique. Moura et al. (2011) predicted the time of failure through SVR, and compared the performance with ARIMA, a conventional time series analysis model, and recurrent neural network (RNN) based on deep learning. As a result, it was confirmed that failure point prediction was possible through SVR, and it has better performance than that of ARIMA.

Sampaio et al. (2019) used ANN, regression tree, RF, and SVM to predict motor failure time. As a result, it was found that the error rate of ANN was the lowest. In addition, Guo et al. (2017) and Ke et al. (2017) used RNN among deep learning techniques. In particular, Ke et al. (2017) used RNN and long short-term memory (LSTM) together with linear regression and SVR. Using PHM 2010 data, a new machine monitoring system called convolutional bi-directional LSTM was proposed (Zhao et al, 2017).

In the study, a total of 12 prediction models were constructed by reflecting the techniques used in previous studies. The frequency and leak level data of the actual inverter are analysed. We propose an optimal failure prediction model through the performance comparison among the several models.

3 Methodology

In the study, a total of 12 prediction models were used to construct a time series-based failure prediction model. Since the prediction error rate is different depending on the models used, it is necessary to search for the most suitable prediction model. Among the various prediction methodologies, the boosting method, the tree method, the SVR method, and the deep learning method which are representative methodologies were selected as the main prediction methodologies, and the research was conducted with the representative models of each method.

3.1 Boosting methods

Boosting is a machine learning technique to learn how to adjust the weights for the training data of the next classifier based on the learning outcomes of the previous classifier. The boosting-based learning models used in the study are Ada boost, gradient boost, XGBoost and LGBM.

Ada boost performs step-by-step learning in a way that weak classifiers complement each other (Viola and Jones, 2004). When learning weak classifiers sequentially, the results of the previous classifier's misclassification are used to train the next classifier. That is, the weight of the classifier is adaptively modified to accurately classify the sample misclassified in the previous step so that it can focus more on the misclassified data. Although classification performance is improved through this learning method, overfitting may easily occur.

Gradient boost also generates a weak classifier and builds a strong classifier through learning. Graident boost utilises a gradient descent algorithm to reduce the residual error in the learning process (Bentéjac et al., 2019; Rahman et al., 2020). Through this algorithm, the loss function is trained to be minimised.

XGBoost is an algorithm designed to compensate for the disadvantages of gradient boost (Rahman et al., 2020). XGBoost is an algorithm that efficiently solves computational problems that occur when using gradient boost through parallelisation, and has a faster computational speed (Anju and Sharma, 2017). In addition, the complexity of the tree can be adjusted by adjusting detailed parameters. The overfitting problem, which is a limitation of the machine learning model, was improved by using the random subsampling technique of each individual tree.

LGBM is a boosting model that utilises leaf-wise tree growth (Ke et al., 2017). A typical boost model utilises a level-wise method to effectively reduce the depth of a tree, but LGBM expands the tree vertically (Omar and Belkhayat, 2018). Through this learning method, LGBM uses a small amount of memory and the algorithm operation speed is fast. However, LGBM is very sensitive to overfitting. When the size of the data is small, it may indicate a result biased to the training data.

3.2 Tree methods

The tree method creates a single independent model and ultimately predicts the result value through voting on the results of each model. The tree-based learning models used in this study are decision tree and RF (Anju and Sharma, 2017).

A decision tree is a model that classifies data according to certain conditions. Each division divides the variable area into two (Safavian and Landgrebe, 1991). The concept of impurity is used to select the branching criterion of the decision tree, and the tree is formed in a direction in which the value of impurity decreases. Decision trees are easy to use and easy to interpret, but overfitting can easily occur (Song and Ying, 2015). Also, it is difficult to understand the interaction between variables.

RF was proposed to overcome the problem of decision trees. It is a model that creates multiple decision trees on the same data and synthesises the results to make predictions. Some trees may be overfitted, but by creating multiple trees, we reduce the impact of the overfitted tree. In addition, RF is easier to generalise than decision tree because it uses bagging technique.

3.3 SVR methods

The SVR technique is a predictive model used when the variable to be predicted is continuous (Awad and Khanna, 2015). Find the line with the largest gap between observations belonging to different classifications and return it as a continuous number. The SVR models used in this study are linear SVR, Nu SVR, and radial basis function (RBF) SVR. Each model is distinguished by the kernel used when configuring the SVR. Linear SVR is used when creating a linear division boundary. NuSVR uses the parameter 'Nu' to control the number of support vectors (Chang and Lin, 2011). Finally, RBF is a method of classifying support vectors after mapping the given data into a high-dimensional space. It can handle data distributions that are usually difficult to classify (Scholkopf et al., 1997). There are many other kernels, but in this study, the model was constructed through three kernels.

3.4 Deep learning methods

The deep learning technique is a methodology that uses multiple layers to identify the core contents of a large amount of data, adjusts weights through comparison with correct labels, and improves predictive power. The deep learning models used in this study are RNN, LSTM and gated recurrent unit (GRU).

RNN is a representative time series analysis algorithm (Connor et al., 1994). RNNs are mainly used when learning sequential data. Through a circular structure, previous data influences current data. This structure shows good results when processing sequential data compared to other ANN techniques. In the case of general RNN, if the layer is too deep, the information of the hidden state cannot be conveyed well, so a vanishing gradient problem may occur.

An algorithm designed to solve this vanishing gradient problem is LSTM (Sepp and Schmidhuber, 1997). LSTM is designed to transmit information well through the cell state. The cell state is managed through the input gate and the forget gate, and appropriate information is transmitted to the next stage through the output gate. This allows the LSTM to handle data from the past time as well.

GRU is a simplified model of LSTM. It shows a simplified algorithm structure compared to LSTM (Cho et al., 2014). Compared to LSTM, it takes less time to learn because there are fewer weights to learn.

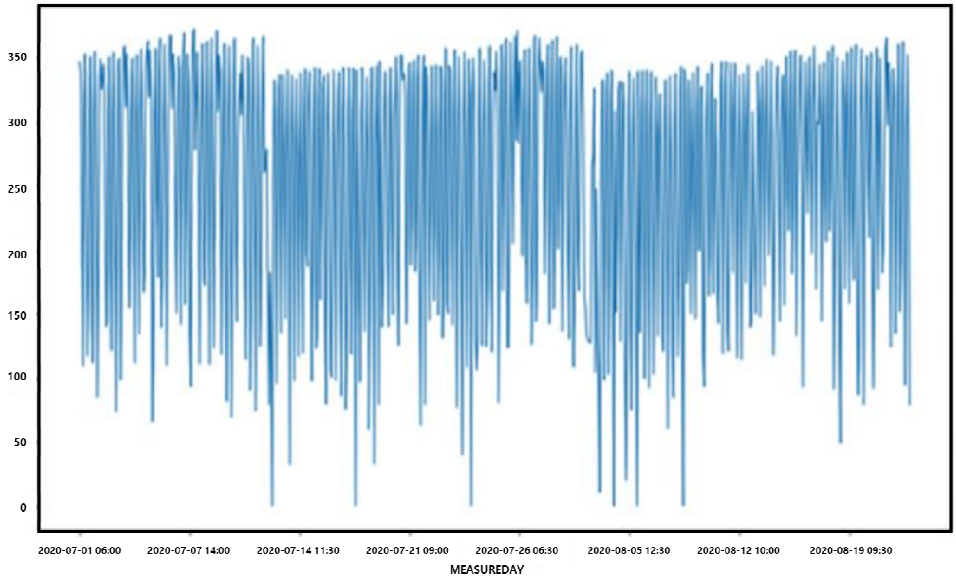
Table 1 This table shows the model used for prediction by each method

Regression methods	
Boosting	Ada boost
	Graident boost
	XGBoost
	LightGBM
Tree	Decision tree
	Random forest
SVR	RBF SVR
	Linear SVR
	Nu SVR
Deep learning	RNN
	LSTM
	GRU

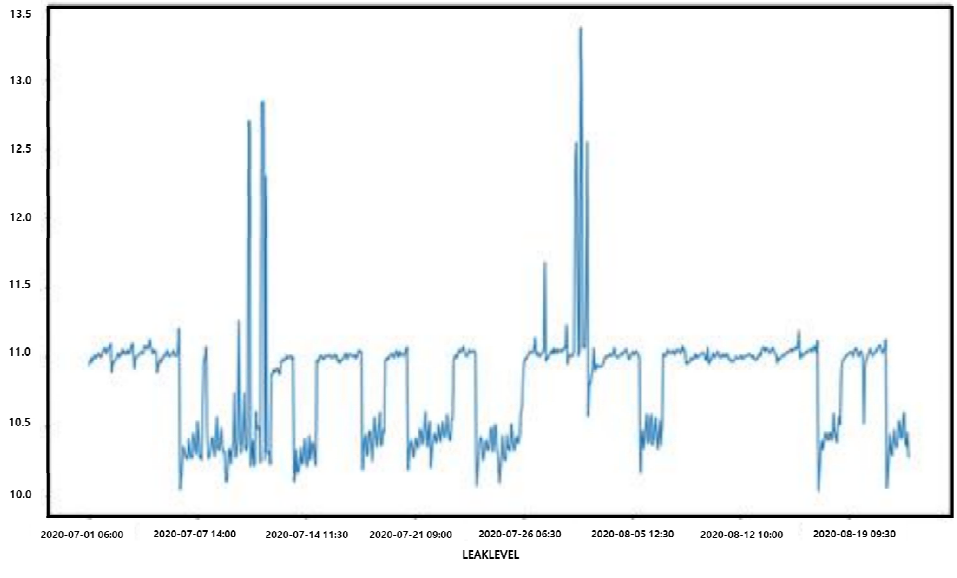
4 Data preprocessing

In the study, frequency and leak level data of inverter were used to construct a time series failure prediction model. The data were collected through actual inverter operation. The collection period was from July to August, 2020, and 40 to 70 data were collected per minute. As a result of data sensing, a total of 8,954,665 time data were obtained. These data are too voluminous to apply general analysis techniques. Therefore, a research method differentiated from the existing time series is required. In this study, machine learning techniques and deep learning techniques were applied to solve these time series problems. In addition, the following preprocessing was performed to model the inverter data.

Figure 1 This figure visualises the data as a graph, (a) graph of frequency data; frequency values usually range from 100 to 350 (b) graph of leak level data; leak level data is represented by the dense and repeatedly form (see online version for colours)



(a)



(b)

The following preprocessing was performed to model the inverter data. First, to match the unity of working time, data for a specific working time period was extracted. In the case of the data used in the study, the data collection time coincides with the total working hours, but in the case of the actual working environment, the working hours may be different depending on various circumstances, so work to unify them was necessary.

Data extraction was carried out by setting the time when the task was most active from 6:00 to 16:00. Secondly, data of time points were created by the unit of 30-minute. Because similar values tend to be repeated continuously in the case of frequency and leak level data, 30 minutes was set as a unit of time point for the efficiency of model configuration. Thirdly, data transformation for model training was performed. The raw data were substituted between 0 and 1 using the min-max scalar normalisation technique. In addition, time shift was performed to predict the next time point through 12 time points. The basic statistics for the data are shown in Table 2, Figure 1 show graphs of frequency and leak level, respectively.

In this study, data from July to August 9, 2020 was used as train data for model learning, and the rest of the data was used as test data. The model is randomly tested 100 times, and the mean error rate and standard deviation are presented together. After comparing the prediction performance of the model through the test data, it predicts the next 6 hours (12 time points) using the actual data.

Table 2 This table shows the basic statistics of frequency data and leak level data

Basic-statistics	*Frequency*	*Leak level*
Count	756	748
Mean	251.76	10.81
Std.	93.12	0.37
Min	0	10.04
25%	170.48	10.44
50%	278.17	11.00
75%	336.11	11.03
Max	372	13.37

5 Results

5.1 Model evaluation

5.1.1 Frequency data

In order to monitor the presence or absence of an inverter failure, a suitable model and appropriate evaluation criteria are required. In the study, root mean square error (RMSE) and mean absolute error (MAE) were used as evaluation indicators to evaluate the predicted values for each model. For the reliability of the predicted values derived from each model, each model was repeatedly measured 100 times to derive the result value.

Table 3 and Table 4 show the error rate for the result of predicting the frequency data through the prediction models. The following table compares the predicted results between various methods. As a result, the error value of the deep learning method was the smallest with RMSE of about 0.2, and the deviation of the result value was also the smallest.

Table 3 This table shows the RMSE of the model

Root mean square error								
Methods	*Model*	*Mean*	*Std.*	*Min*	*25%*	*50%*	*75%*	*Max*
Boost	Ada	0.40	0.01	0.37	0.39	0.40	0.40	0.44
	Gradient	0.54	0.00	0.54	0.54	0.54	0.55	0.55
	XGB	0.55	0.00	0.55	0.55	0.55	0.55	0.55
	LGBM	0.58	0.00	0.58	0.58	0.58	0.58	0.58
Tree	Decision	0.61	0.00	0.54	0.56	0.56	0.57	0.57
	Random	0.56	0.00	0.55	0.56	0.56	0.57	0.57
SVR	RBF	0.54	0.00	0.54	0.54	0.54	0.54	0.54
	Nu	0.56	0.00	0.56	0.56	0.56	0.56	0.56
	Linear	0.52	0.00	0.52	0.52	0.52	0.52	0.52
Deep learning	RNN	0.19	0.00	0.19	0.19	0.19	0.19	0.19
	GRU	0.19	0.00	0.19	0.19	0.19	0.19	0.19
	LSTM	0.19	0.00	0.19	0.19	0.19	0.19	0.19

Notes: The data used for this prediction is frequency data. Measurements were repeated 100 times, and the basic statistics are shown.

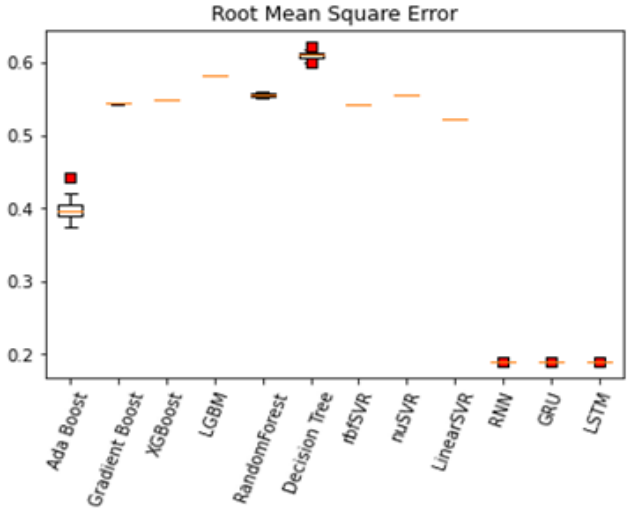
Table 4 This table shows the MAE of the model

Mean absolute error								
Methods	*Model*	*Mean*	*Std.*	*Min*	*25%*	*50%*	*75%*	*Max*
Boost	Ada	0.33	0.01	0.30	0.32	0.33	0.34	0.37
	Gradient	0.45	0.00	0.45	0.45	0.45	0.45	0.45
	XGB	0.45	0.00	0.45	0.45	0.45	0.45	0.45
	LGBM	0.49	0.00	0.49	0.49	0.49	0.49	0.49
Tree	Decision	0.52	0.00	0.50	0.51	0.52	0.52	0.53
	Random	0.46	0.00	0.46	0.46	0.46	0.46	0.47
SVR	RBF	0.45	0.00	0.45	0.45	0.45	0.45	0.45
	Nu	0.46	0.00	0.46	0.46	0.46	0.46	0.46
	Linear	0.42	0.00	0.42	0.42	0.42	0.42	0.42
Deep learning	RNN	0.14	0.00	0.14	0.14	0.14	0.14	0.14
	GRU	0.14	0.00	0.14	0.14	0.14	0.14	0.14
	LSTM	0.14	0.00	0.14	0.14	0.14	0.14	0.14

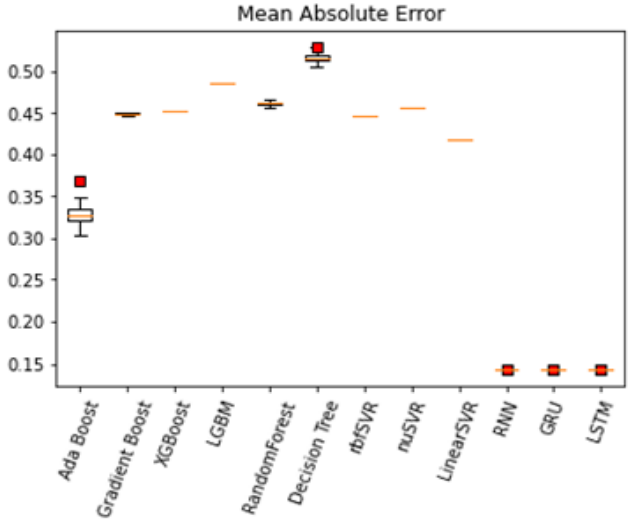
Notes: The data used for this prediction is frequency data. Measurements were repeated 100 times, and the basic statistics are shown.

Except for the deep learning method, almost the same error rate is shown. However, among the boosting techniques, Ada boost shows good results compared to other machine learning methods. In particular, it is seen that there is a significant difference compared to the same boosting techniques such as gradient, XGB, and LGBM. It showed the lowest error rate among similar boosting models. The results of comparison for the error values are shown with the box plots shown in Figure 2.

Figure 2 This figure shows the RMSE and MAE values of the frequency data prediction model as a box-plot (see online version for colours)



(a)



(b)

Note: Decision tree has the highest error rate and deep learning method has the lowest.

5.1.2 Leak level data

The error rate of the result of predicting leak level data through the prediction model is shown in Table 5 and Table 6. As a result of the prediction, it can be seen that the deep learning method has fewer error values than other methods. It has an error rate of about 0.08 based on RMSE. In the case of leak-level data, there is usually a small error rate because there is less variation in the raw data.

It can be seen that other techniques except deep learning have a high level of error. The average error level of the other techniques is about 0.144 based on the RMSE. This is about twice that of deep learning models.

Table 5 This table shows the RMSE of the model

Root mean square error								
Methods	*Model*	*Mean*	*Std.*	*Min*	*25%*	*50%*	*75%*	*Max*
Boost	Ada	0.13	0.00	0.11	0.12	0.13	0.13	0.13
	Gradient	0.18	0.00	0.18	0.18	0.18	0.19	0.19
	XGB	0.15	0.00	0.15	0.15	0.15	0.15	0.15
	LGBM	0.14	0.00	0.14	0.14	0.14	0.14	0.14
Tree	Decision	0.19	0.00	0.19	0.19	0.19	0.19	0.19
	Random	0.14	0.00	0.14	0.14	0.14	0.14	0.14
SVR	RBF	0.11	0.00	0.11	0.11	0.11	0.11	0.11
	Nu	0.14	0.14	0.14	0.14	0.14	0.14	0.14
	Linear	0.12	0.12	0.12	0.12	0.12	0.12	0.12
Deep learning	RNN	0.08	0.00	0.08	0.08	0.08	0.08	0.08
	GRU	0.08	0.00	0.08	0.08	0.08	0.08	0.08
	LSTM	0.08	0.00	0.08	0.08	0.08	0.08	0.08

Notes: The data used for this prediction is leak level data. Measurements were repeated 100 times, and the basic statistics are shown.

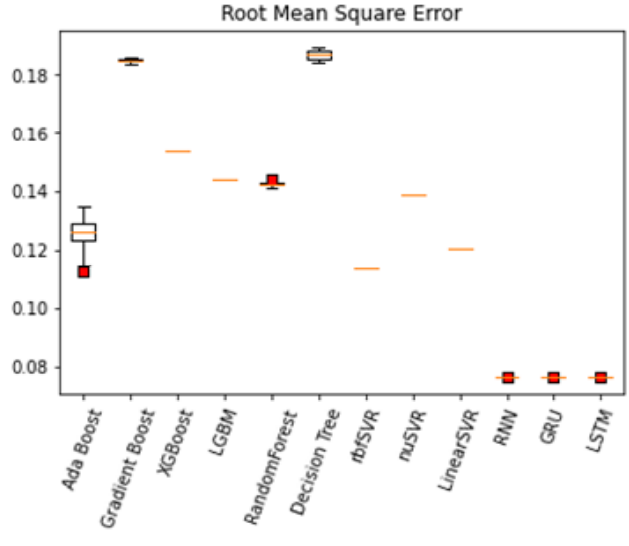
Table 6 This table shows the MAE of the model

Mean absolute error								
Methods	*Model*	*Mean*	*Std.*	*Min*	*25%*	*50%*	*75%*	*Max*
Boost	Ada	0.9	0.01	0.07	0.08	0.09	0.09	0.10
	Gradient	0.12	0.00	0.12	0.12	0.12	0.12	0.12
	XGB	0.11	0.00	0.11	0.11	0.11	0.11	0.11
	LGBM	0.11	0.00	0.11	0.11	0.11	0.11	0.11
Tree	Decision	0.12	0.00	0.12	0.12	0.12	0.12	0.12
	Random	0.10	0.00	0.10	0.10	0.10	0.10	0.10
SVR	RBF	0.08	0.00	0.08	0.08	0.08	0.08	0.08
	Nu	0.10	0.00	0.10	0.10	0.10	0.10	0.10
	Linear	0.09	0.00	0.09	0.09	0.09	0.09	0.09
Deep learning	RNN	0.03	0.00	0.03	0.03	0.03	0.03	0.03
	GRU	0.03	0.00	0.03	0.03	0.03	0.03	0.03
	LSTM	0.03	0.00	0.03	0.03	0.03	0.03	0.03

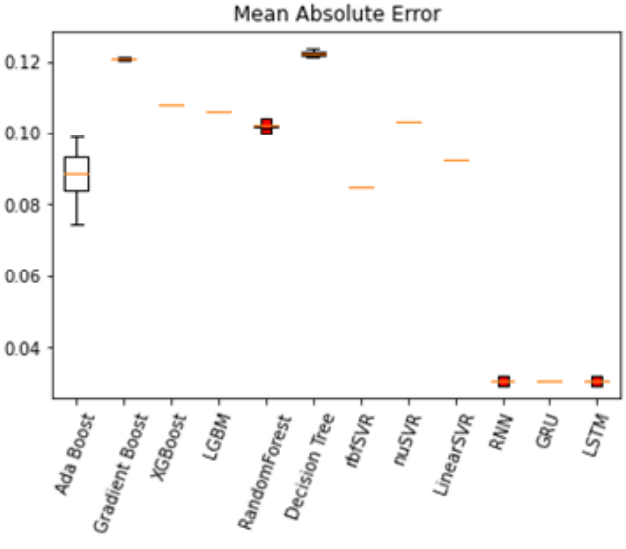
Notes: The data used for this prediction is leak level data. Measurements were repeated 100 times, and the basic statistics are shown.

Figure 3 shows box-plots for various models. For other models, the variance of the results is very small, but Ada boost has a very large variance. Ada boost has a margin of error of 0.2 based on RMSE.

Figure 3 This figure shows the RMSE and MAE values of the leak level data prediction model as a box-plot (see online version for colours)



(a)



(b)

Note: Gradient boost and decision tree have the highest error rate, and deep learning method has the lowest error rate.

5.2 Performance

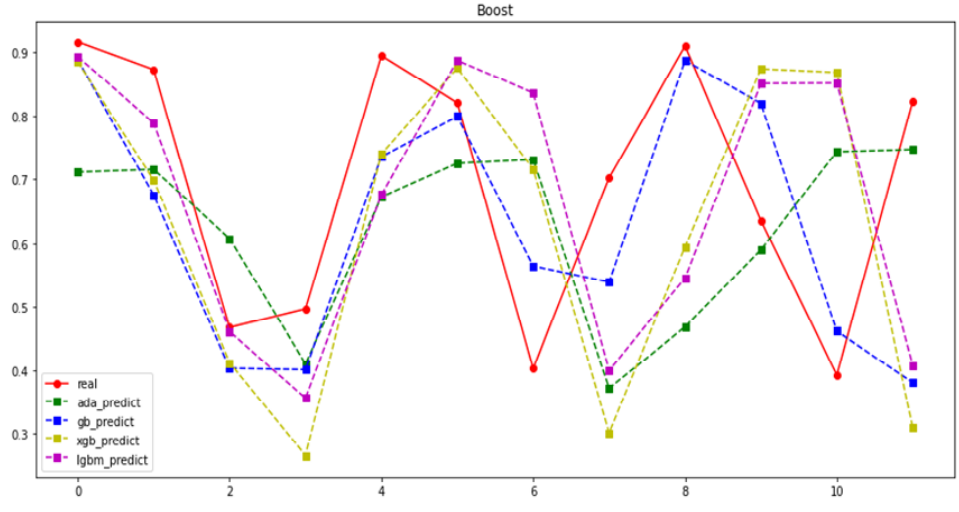
5.2.1 Prediction of frequency data

Figures 4–7 show the result of predicting frequency data through the models. The next time point was predicted using 12 time points. In the case of frequency data, 12 points in

the future were predicted from 06:00 on August 10, 2020. This is the starting point of the test set. That is, the value of 06:00 on August 10, 2020 is predicted using the last 12 time points (August 9, 2020 12:00–August 9, 2020 18:00) of the train set. The next predicted value is also predicted using the previous 12 time points.

The result of the boosting methods is shown in Figure 4. In the case of the AdaBoost model, which is the best model among machine learning models, it shows bad results in the actual prediction graph. Although the flow of the actual value is reflected, it appears that it does not approach the pole of the actual value. The reason why the AdaBoost model showed the lowest error rate is that it maintains the result value that is somewhat close to the average value. In the case of other Boost method models, the pole value is found, but the time point of the predicted value cannot be grasped at the exact location. Therefore, the error rate was higher than that of the AdaBoost method.

Figure 4 This figure is a graph of predicting frequency data using the boost method (see online version for colours)



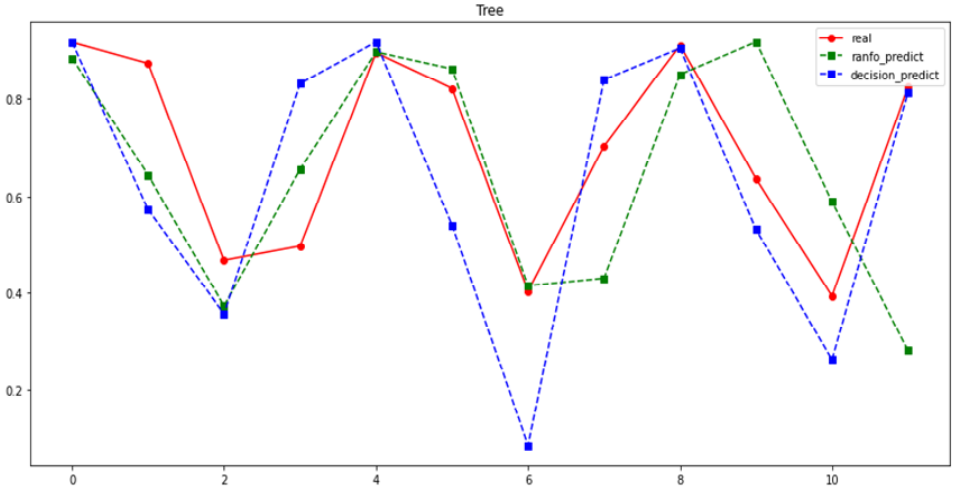
Note: Although the flow with the actual value appears similar, it does not perform an accurate prediction.

Figures 5 and 6 show the results of predicting frequency data using the tree method and the SVR method.

Among the tree techniques, the RF model has superior predictive value compared to other models. Not only is it following the flow of real values, but it is also following the values of the poles. However, in the case of the decision tree model, the prediction rate is lower than that of the RF. Also, there is a large error at the midpoint of the prediction.

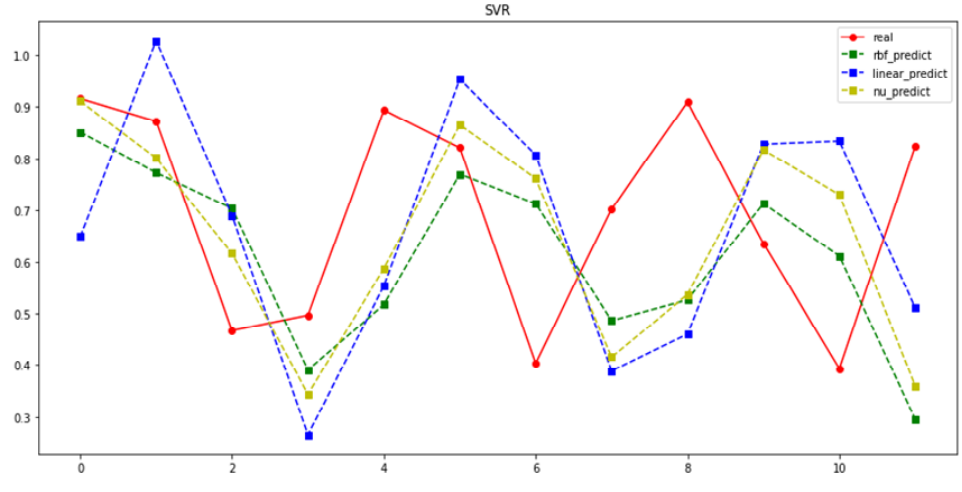
The SVR method is similar to the result of the Boosting method. It is not possible to predict a value at an exact point in time.

Figure 5 This figure is a graph of predicting frequency data using the tree method (see online version for colours)



Note: The prediction of decision tree is inaccurate compared to RF.

Figure 6 This figure is a graph predicting frequency data using the SVR method (see online version for colours)

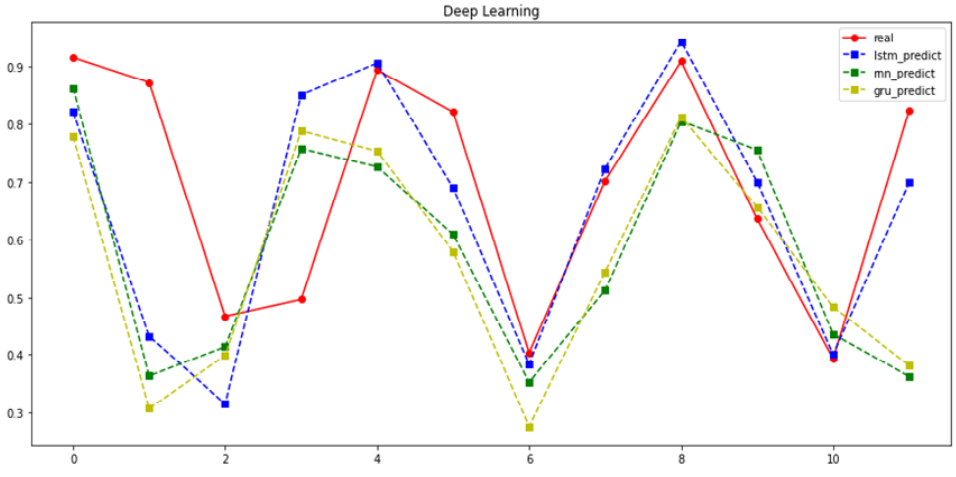


Note: All three models used in this study show inaccurate prediction results.

Figure 7 shows the result of predicting the next time value using the deep learning technique. RNN and GRU show similar patterns, while LSTM shows different predictions from the two models. In the case of RNN and GRU, the predicted value is lower than the actual result value. In the case of the LSTM model, it most accurately represents the overall change trend among deep learning techniques. Although it is somewhat sluggish at the beginning of the forecast, it shows the closest predictive model in all forecasts after that. According to Table 5 and Table 6, although the model performance is similar, the actual prediction graph shows the best results. Therefore,

LSTM has the highest prediction accuracy among all models and is the best model for frequency data prediction.

Figure 7 This figure is a graph predicting frequency data using a deep learning method (see online version for colours)

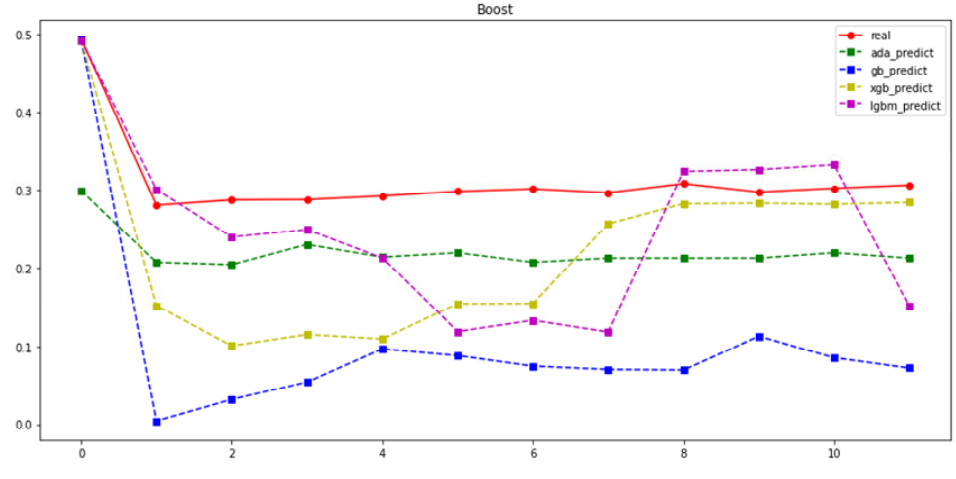


Note: The LSTM model shows the most accurate result graph.

5.2.2 Prediction of leak level data

Figures 8–11 show the predicted value of leak level data as a graph. In the case of leak level data, the next time point was predicted using the previous four time points. Because the leak level values are relatively densely gathered, they sensitively reacted to changes in the model. From 06:00 on July 29, 2020, the next 12 time points were predicted.

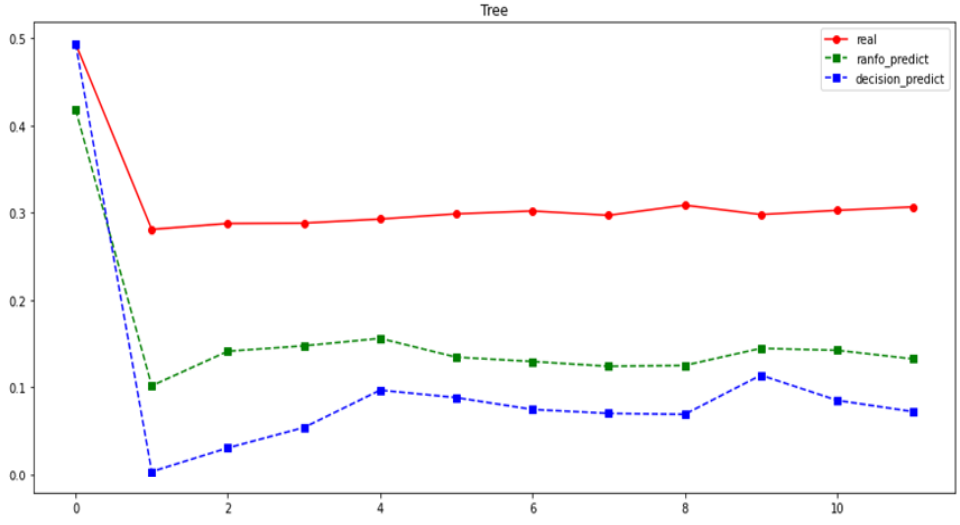
Figure 8 This figure is a graph predicting leak level data using the boost method (see online version for colours)



Note: All models used in this study show inaccurate prediction results.

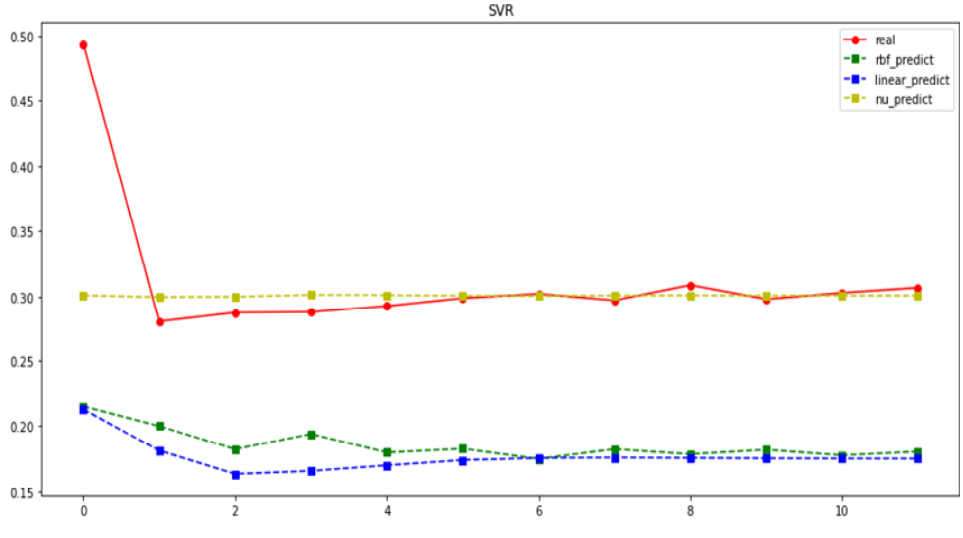
Figure 8 shows the result of predicting leak level data using the boosting method. For leak-level data, the predicted values between the boosting methods vary greatly. In general, the predicted values are not accurate. In the case of the gradient boost model, the flow of the predicted values is similar, but the error rate is high. Also, the Ada boost model shows a flat prediction value compared to LGBM and XGB. For this reason, among the boosting methods, Ada boost the lowest prediction error.

Figure 9 This figure is a graph of predicting leak level data using the tree method (see online version for colours)



Note: Overall, the tree method shows predicted values that are lower than the actual values.

Figure 10 This figure is a graph predicting leak level data using the SVR method (see online version for colours)



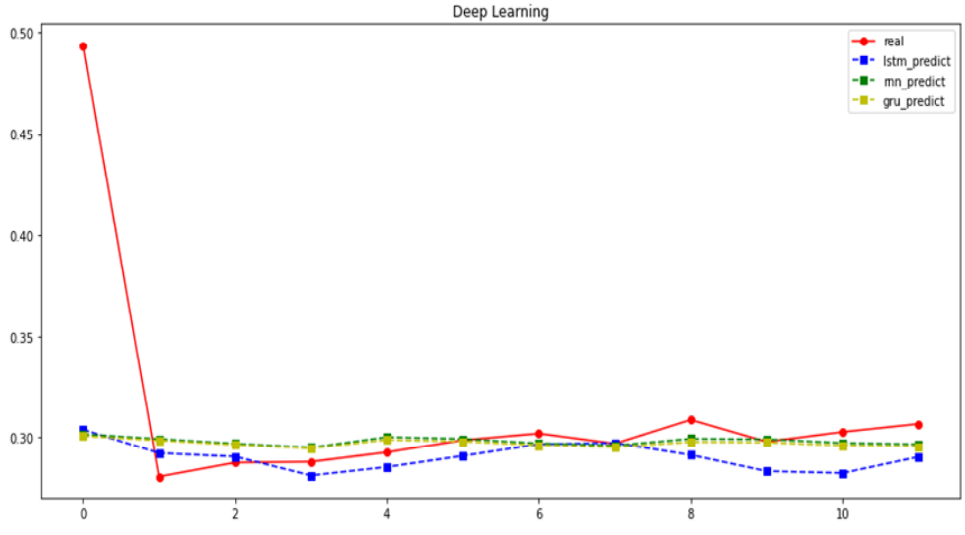
Note: The Nu SVR model shows the predicted value closest to the actual value.

Figure 9 and Figure 10 show the result of predicting leak level data using the tree method and the SVR method. As a result of the tree method shown in Figure 9, the tree method is not good for predicting leak level data. Because, like the gradient boost, it shows an inaccurate prediction value. In the case of RF, it showed somewhat better prediction value than decision tree, but there are many errors compared to the actual value.

The graph of the prediction result using the SVR method is shown in Figure 10. The SVR method also showed low prediction performance. However, in the case of the Nu SVR model, the analysis result was close to the actual value. The value at the first time point showed a large error rate, but from the second time point, the predicted value was similar to the actual value.

Figure 11 shows the result of predicting the leak level value at the next time point using the deep learning technique. In the case of leak level data, LSTM, RNN, and GRU all show similar analysis results. As with the frequency data, it shows a bit sluggish at the beginning of the forecast, but shows the closest predictive model in all forecasts after that. Unlike other machine learning methods, the deep learning method shows a similar type of prediction value. The prediction value of the first time point shows a large error value, which is similar to the prediction value of Nu SVR. Therefore, the predictive power of the deep learning method is the most stable compared to the machine learning method used in this paper.

Figure 11 This figure is a graph predicting leak level data using a deep learning method (see online version for colours)



Notes: Deep learning methods represent relatively accurate prediction graphs. However, the predictive power of LSTM is somewhat lower than that of the other two models.

6 Discussion and conclusions

6.1 Discussion

In this study, frequency and leak level data were used to predict inverter failure. Inverter data was analysed and prediction results were compared using various machine learning models and deep learning models. As a result of comparison, the error rate of the deep learning model was generally low. In addition, the prediction results using each model also showed that the deep learning model was excellent. This is because deep learning models are advantageous when learning data for a long period of time. In particular, in the case of LSTM and GRU, more accurate learning can be performed through the long-term memory gate, and the performance of such learning can be confirmed through the result graph of the predictive model.

In this study, an excellent predictive model was presented through a deep learning model, but there are some limitations. In the actual working environment, it is necessary to reflect various variables because more factors can affect the failure. When analysing with a single variable, the amount of change according to time change cannot be sufficiently explained, so there is a limit to the improvement of the model's performance. In addition, in the case of the data used in this study, since the data was not collected under a certain working time, some data were extracted by arbitrarily set daily work hours and then analysed. In this process, key factors in the data may be omitted. If the main factor is omitted, it is difficult to improve the error rate.

Therefore, in future research, it is necessary to adjust the sensing environment to utilise the entire data set. It is expected that prediction accuracy will be further improved if all data can be utilised. Also, in this study, we tried to predict a long time period, but in this case, a lot of errors appear compared to predicting a short period. Further discussion is needed on whether predicting a certain point in time is appropriate for the actual environment.

6.2 Conclusions

In this study, we tried to construct a time series-based failure prediction model using actual inverter data. The inverter data used in this study consists of frequency and leak level, and when an abnormality occurs, the corresponding value increases significantly. The measured inverter data consists of a total of 8,954,665 time points. Traditional analysis methodologies have limitations in analysing such a large volume of time series data. Therefore, in this study, we tried to analyse the time series process data generated by the inverter using various machine learning methods. The machine learning methods used in this study can be divided into boosting methods, tree methods, SVR methods, and deep learning methods. For each technique, we tried to predict the following 12 time point values through a representative model. The performance of the model was measured by RMSE and MAE, and it was repeated 100 times to ensure the reliability of the results.

As a result of analysing frequency data through various techniques, the deep learning technique showed the highest performance. In the case of the boosting method and the SVR method, the flow of the actual value is well followed, but the value at the time of prediction shows a large difference from the actual value. In particular, in the case of the Adaboost model, the superiority of the model itself was confirmed through the error rate,

but when applied to prediction, it was confirmed that there was a lot of difference from the actual value. In the case of RF, one of the tree techniques, the prediction value was relatively accurate compared to other models, but it was insufficient compared to the deep learning technique. The deep learning method showed the lowest error rate when compared to other methods. In addition, when drawing an actual prediction graph, it showed high accuracy compared to other techniques. In fact, when the prediction graph was drawn using the LSTM model, which is one of the deep learning techniques, it showed a slight difference from the actual value at the beginning of the prediction, but showed almost the same as the actual value from the 5th time point. Therefore, it is effective to use the LSTM model when predicting frequency data.

As for the result of analysing leak level data, the deep learning method showed the highest performance as well. In particular, in prediction, LSTM, RNN, and GRU all showed similar results. Although the initial predicted value deviated significantly, the subsequent predicted value appeared close to the actual value. In the case of other methods, most of the prediction values deviated significantly, and in the case of the Nu SVR model, the results were similar to those of the deep learning method. Therefore, it can be said that it is effective to use a deep learning method when predicting the leak level.

In this study, various preprocessing techniques were applied to efficiently analyse big data time variables. In this process, key attributes of data may be diluted or lost. In addition, it did not reflect various environmental factors that may affect the inverter. Therefore, in future research, we intend to propose a more general and accurate inverter life prediction algorithm through the use of various variables and algorithm optimisation.

Acknowledgements

Author contributions: Sang-Ha Sung – analysis of literature, analysis of experimental data, validation of model, draft and final copy of the manuscript. Chang Sung Seo – analysis of literature, validation of model, draft and final copy of the manuscript. Min Ho Ryu – consulting, literature analysis. Sangjin Kim – research idea, formulation of research goals and objectives, guidance and consulting, examine of calculation results. All authors have read and agreed to the published version of the manuscript.

This research was funded by Dong-A University, South Korea.

References

Anju and Sharma, N. (2017) 'Survey of boosting algorithms for big data applications', *International Journal of Engineering Research & Technology (IJERT)*, Vol. 5, No. 11, ISSN: 2278-0181.

Awad, M. and Khanna, R. (2015) 'Support vector regression', *Efficient Learning Machines*, pp.67–80, DOI: https://doi.org/10.1007/978-1-4302-5990-9_4.

Bentéjac, C., Csörgó, A. and Martínez-Muñoz, G. (2019) *A Comparative Analysis of XGBoost*, ArXiv 2019, DOI: 10.1007/s10462-020-09896-5.

Chang, C-C. and Lin, C-J. (2011) 'LIBSVM: a library for support vector machines', *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 2, No. 3, pp.1–27, DOI: https://doi.org/10.1145/1961189.1961199.

Chen, T. and Giestrin, C. (2016) 'XGBoost: a scalable tree boosting system', *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785–794, DOI: https://doi.org/10.1145/2939672.2939785.

Cheon, K.M. and Yang, J. (2020) 'An ensemble model for machine failure prediction', *Journal of Society of Korea Industrial and Systems Engineering*, Vol. 43, No. 1, pp.123–131, DOI: https://doi.org/10.11627/jkise.2020.43.1.123.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014) *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*, arXiv preprint arXiv:1406.1078.

Connor, J.T., Martin, R.D. and Atlas, L.E. (1994) 'Recurrent neural networks and robust time series prediction', *IEEE Transactions on Neural Networks*, Vol. 5, No. 2, pp.240–254, DOI: 10.1109/72.279188.

Guo, L., Li, N., Jia, F., Lei, Y. and Lin, J. (2017) 'A recurrent neural network based health indicator for remaining useful life prediction of bearings', *Neurocomputing*, Vol. 240, No. 31, pp.98–109, DOI: https://doi.org/10.1016/j.neucom.2017.02.045.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T-Y. (2017) 'LightGBM: a highly efficient gradient boosting decision tree', *Advances in Neural Information Processing Systems (NIPS 2017) 2017*, Vol. 30, pp.3149–3157.

Lee, G-Y., Kim, M., Quan, Y-J., Kim, M-S., Kim, T.J.Y., Yoon, H-S., Min, S., Kim, D-H., Mun, J-W., Oh, J.W., Choi, I.G., Kim, C-S., Chu, W-S., Yang, J., Bhandari, B., Lee, C-M., Ihn, J-B. and Ahn, S-H. (2008) 'Machine health management in smart factory: a review', *Journal of Mechanical Science and Technology*, Vol. 32, pp.987–1009, DOI: 10.1007/S12206 -018-0201-1.

Lee, J.H., Yoo, S-Y., Shin, S-c., Kang, D-H., Lee, S-s. and Lee, J.C. (2019) 'Fault diagnosis of bearings using machine learning algorithm', *Journal of the Korean Society of Marine Engineering*, Vol. 43, No. 6, pp.1876–1886, DOI: https://doi.org/10.1016/j.eswa.2010.07.119.

Lee, S.W. (2019) 'Smart factory overseas trends in major countries', *Proceedings of Symposium of the Korean Institute of communications and Information Sciences*, pp.1039–1039.

Liu, Y.C., Hu, X.F. and Sun, S.X. (2019) 'Remaining useful life prediction of cutting tools based on support vector regression', *IOP Conference Series: Materials Science and Engineering*, Vol. 576, DOI: doi:10.1088/1757-899X/576/1/012021.

Mathew, V., Toby, T., Singh, V., Rao, B.M. and Kumar, M.G. (2018) 'Prediction of remaining useful lifetime (RUL) of turbofan engine using machine learning', *IEEE International Conference on Circuits and Systems*, DOI: 10.1109/ICCS1.2017.8326010.

Moura, M.d.C., Zio, E., Lins, I.D. and Droguett, E. (2011) 'Failure and reliability prediction by support vector machines regression of time series data', *Reliability Engineering & System Safety*, Vol. 96, No. 11, pp.1527–1534, DOI: https://doi.org/10.1016/j.ress.2011.06.006.

Oh, H-W. and Huh, J-D. (2020) 'IoT-based smart factory failure prediction analysis technology to improve productivity and quality', *The Institute of Electronics and Information Engineers*, Vol. 47, No. 11, pp.33–43, ISSN: 1016-9288.

Omar, A. and Belkhayat, K. (2018) *XGBoost and LGBM for Porto Seguro's Kaggle Challenge: A Comparison*, Preprint Semester Project.

Rahman, S., Irfan, M., Raza, M., Ghori, K.M., Yaqoob, S. and Awais, M. (2020) 'Performance analysis of boosting classifiers in recognizing activities of daily living', *International Journal of Environmental Research and Public Health*, Vol. 17, No. 3, DOI: 10.3390/ijerph17031082.

Raj, J.S. and Ananthi, J.V. (2019) 'Recurrent neural networks and nonlinear prediction in support vector machines', *Journal of Soft Computing Paradigm*, Vol. 1, No. 1, pp.33–40, DOI: 10.36548/jscp.2019.1.004.

Safavian, S.R. and Landgrebe, D. (1991) 'A survey of decision tree classifier methodology', *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 21, No. 3, pp.660–674, DOI: 10.1109/ 21.97458.

Sampaio, G.S., Filho, A.R.d.A.V., da Silva, L.S. and Silva, L.A. (2019) 'Prediction of motor failure time using an artificial neural network', *Sensor Technologies for Smart Industry and Smart Infrastructure*, Vol. 19, No. 19, DOI: https://doi.org/10.3390/s19194342.

Sapankevych, N.I. and Sankar, R. (2009) 'Time series prediction using support vector machines: a survey', *IEEE Computational Intelligence Magazine*, Vol. 4, No. 2, pp.24–38, DOI: 10.1109/MCI.2009.932254.

Scholkopf, B., Sung, K-K., Burges, C.J.C., Girosi, F., Niyogi, P., Poggio, T. and Vapnik, V. (1997) 'Comparing support vector machines with Gaussian kernels to radial basis function classifiers', *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp.2758–2765, DOI: 10.1109/78.650102.

Sepp, H. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, Vol. 9, No. 8, pp.1735–1780, DOI: 10.1162/neco.1997.9.8.1735.

Song, Y-Y. and Ying, L.U. (2015) 'Decision tree methods: applications for classification and prediction', *Shanghai Archives of Psychiatry*, Vol. 27, No. 2, pp.130–135, DOI: 10.11919/j.issn.1002-0829.215044.

Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y. and Huang, H. (2020) 'Statistical and machine-learning methods for clearance time prediction of road incidents: a methodology review, *Analytic Methods in Accident Research*, Vol. 27, DOI: https://doi.org/10.1016/j.amar.2020.100123.

Tran, V.T., Yang, B-S., Oh, M-S. and Tan, A.C.C. (2008) 'Machine condition prognosis based on regression trees and one-step-ahead prediction', *Mechanical Systems and Signal Processing*, Vol. 22, No. 5, pp.1179–1193, DOI: https://doi.org/10.1016/j.ymssp.2007.11.012.

Viola, P. and Jones, M.J. (2004) 'Robust real-time face detection', *International Journal of Computer Vision*, Vol. 57, No. 2, pp.137–154, DOI: https://doi.org/10.1023/B:VISI.0000013087.49260.fb.

Wu, D., Jennings, C., Terpenny, J., Gao; R.X. and Kumara, S. (2017) 'A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests', *Journal of Manufacturing Science and Engineering*, Vol. 139, No. 7, DOI: https://doi.org/10.1115/1.4036350.

Wu, J., Roy, J. and Stewart, W.F. (2010) 'Prediction modeling using EHR data challenges, strategies, and a comparison of machine learning approaches', *Medical Care*, Vol. 48, No. 6, pp.S106–S113, DOI: http://www.jstor.org/stable/20720782.

Zhao, R., Yan, R., Wang, J. and Mao, K. (2017) 'Learning to monitor machine health with convolutional bi-directional LSTM networks', *Sensors (Basel)*, Vol. 17, No. 2, DOI: https://doi.org/10.3390/s17020273.