Please select category below:
Normal Paper ☒
Student Paper □
Young Engineer Paper □

Is Deep Learning Superior in Machine Health Monitoring Applications

Wenyi Wang [1], Kilian Vos [3], John Taylor [2], Chris Jenkins [3], Biswajit Bala [2], Leonard Whitehead [1] and Zhongxiao Peng [3]

[1] Defence Science and Technology Group, Aerospace Division, Fishermans Bend, VIC, 3207, Australia
[2] Defence Science and Technology Group, Research Services Division, Fairbairn, ACT, 2600, Australia
[3] University of New South Wales, Sydney, NSW, 2052, Australia

Abstract

In the last few years, applying deep learning (DL) to machine health monitoring (MHM) has gained enormous momentum with an overwhelming claim that DL methodologies are superior to more traditional techniques of MHM. In this paper, we will address this claim using a real-world sensor fault dataset provided by Airbus – a problem of sensor health unsupervised classification. In the 2019 worldwide competition with this dataset hosted by Airbus, Fujitsu Systems Europe (FSE) won the first prize with an F1-score of 93 percent using a DL model based on generative adversarial network (GAN). Another comprehensive study compared the performance of various modified and the existing image encoding methods for the convolutional auto-encoder (CAE) model, where the best classification result was a F1-score of 91 percent using the scalogram as the image encoding method. We will use these two studies as the benchmark for us to compare with some basic statistical analysis methods and the one-class supporting vector machine (SVM). The comparative study demonstrates that the DL-based techniques have great potential but they are not necessarily always superior to traditional methods. We recommend that where possible all future published studies of applying DL methods to MHM include appropriately selected traditional reference methods.

Keywords: faulty-sensor detection, time sequence classification, machine learning, deep learning, statistical signal analysis, machine health monitoring.

Introduction

With the breakthrough success in applying deep learning (DL) to image recognition and natural language processing in the last decade, many researchers are keen to apply DL to machine health monitoring. Recently, we have observed an exponential growth in the number of studies of applying DL to machine health monitoring and fault diagnostics. There have been over one thousand research papers and many review papers in the literature since 2019 [1]. However, critiques could argue whether DL methodologies are always superior to more traditional approaches to solving MHM problems. In this paper, we will discuss this question using a real-world helicopter sensor fault dataset provided by Airbus – a challenging machine learning problem of unsupervised classification. In the 2019 worldwide AI challenge with this dataset hosted by Airbus, the first prize was won by Fujitsu Systems Europe (FSE) with an F1-score of 93 percent using a DL model based on generative adversarial network (GAN), refer to the original paper of the method by Li et al [2]. Another comprehensive study using the Airbus dataset by Garcia et al [3] compared the performance of various modified and the existing image encoding methods for the convolutional auto-encoder (CAE) model. They achieved the best F1-

score of 91 percent using the scalogram as the image encoding method. We will use these two studies as the benchmark for us to compare with some basic statistical analysis and machine learning methods. We conclude in this comparative study that the DL-based techniques have great potential but they are not necessarily always superior to traditional methods. The rest of the paper is structured as follows. The information about the Airbus dataset is summarised in the next section, followed by the review of the results generated using the deep learning approaches. The data were analysed using conventional methods with some commonly used features. A simple statistical method and one-class supporting vector machine (SVM) were applied and their results are reported and compared to those of the deep learning methods. Some concluding remarks are presented in the last section.

The Airbus Helicopters Sensor Fault Dataset

One of the main challenges in aerospace industry is to test the validity of flight test data from heavily instrumented aircraft due to possible faulty sensors. Because of the sheer volume of measured signals that need to be validated, manual validation is no longer possible. It is crucial to automate the validation process. For this purpose, Airbus collected and released a set of helicopter vibration measurement data from different flight tests – publicly available on https://doi.org/10.3929/ethz-b-000415151.

In all operating conditions of the helicopter collected from different flights, accelerometers were placed at different positions of the helicopter, in different directions (longitudinal, vertical, lateral) to measure the vibration signals with a constant sampling rate of 1024 Hz and sampling length of 1 minute. The training data is composed of 1677 accelerometer data sequences from healthy sensors. The testing (or validation) data has 594 sequences consisting of streams from healthy or faulty sensors. Measurement locations and directions in the testing data may or may not be identical to those of the training data. All signals in the dataset were normalised so that absolute values do not have physical meaning.

With this dataset, Airbus hosted a worldwide AI challenge in 2019 to classify the testing sequences into healthy and faulty sequences. Firstly, only the training data were released for model training to ensure that all the models were trained without a priori knowledge about the testing data. After the submission of trained models, Airbus released the testing data.

Review of Results by Deep Learning Methods

Among the 140 competing teams, Fujitsu Systems Europe (FSE) won the first prize (https://www.fujitsu.com/emeia/about/resources/news/press-releases/2019/emeai-20191211-fujitsu-wins-first-prize-for-predictive.html) with an F1-score of 0.93 using a DL model based on multivariate anomaly detection with generative adversarial network (MAD-GAN). There is no publication about the details of the winning method. In the original paper of MAD-GAN [2], an unsupervised multivariate anomaly detection method was proposed based on generative adversarial networks (GANs). Li et al used the long short-term memory based recurrent neural networks (LSTM-RNN) as the base models for both the generator and discriminator in the GAN framework. Their MAD-GAN framework treated the entire variable set concurrently and each data stream independently to capture the latent interactions amongst the variables. They used a novel anomaly score (DR-score) to detect anomalies through the discrimination and reconstruction phases. The test results showed that the proposed MAD-GAN is an effective method in detecting anomalies caused by cyber-attacks on some complex real-world digital systems. Apparently, FSE's winning result proved the efficacy of MAD-GAN in detecting anomalies caused by faulty sensor measurements.

Garcia et al [3] stressed the fact that it is uncommon to find application cases of unsupervised DL (e.g. AE & CNN) based anomaly detection. They compared six image encoding strategies such as Gramian angular field, Markov transition field, recurrence plot, grey scale encoding, spectrogram and scalogram to transform the raw time series data into images for a convolutional auto-encoder (CAE). They defined a more robust encoding method by modifying each of these six existing algorithms. Training the DL model only on healthy condition data, they extracted the 99th percentile in the distribution of the residuals of all sub-series to define the detection threshold τ. They then monitored the maximum residual over the sub-series (for the detection of local anomalies), and measured it against the threshold τ beyond which an anomaly is considered detected. Using the Airbus dataset, they conducted a comprehensive study comparing the modified and the existing encoding methods and showed an improved performance by using the encoded images against using the raw time series. All the modified versions were observed to perform better than their un-modified counterparts, in which the scalogram indicated the best performance with an F1-score of 0.91 and AUC (area under the curve) score of 0.92.

Results by Non-Deep Learning Methods

We will use the results from above two studies where F1-scores were 0.93 and 0.91 by FSE and Garcia et al respectively as the benchmark for us to compare with some basic statistical analysis methods and the one-class supporting vector machine (SVM).

Simple Statistical Analysis Method

Using some basic statistical analyses, we firstly calculate the two most commonly used statistics, i.e. the mean and the standard deviation (STD), for every sequence (1677 in total) in the training data. We then visualize the distributions of the mean and STD values by a 2D histogram (*histogram2* in Matlab) as shown in Fig. 1. As we can see that the distributions are widely spread on both side of the mean values and mostly on the right-hand side of the STD values. With a given rate of outliers, e.g. 5 percent (or 2.5 percent on either side) as opposed to the 2-sigma principle for a Gaussian distribution, we can draw the boundary lines as indicated by the four red dashed lines in Fig. 1. This rate can be prescribed based on the practical requirements of false positive (or false alarm) rate or false negative (or missed detection) rate. For example, in aerospace industry false negative can be fatal thus it should be minimized by using a relatively large rate of outliers, such as the 5 percent chosen here.

In classification with the testing (or validation) data, we obtain the mean and STD values for each of the 594 sequences and compare them with the boundary values obtained from the training data. If either the mean or the STD values from a sequence in the testing data go beyond the respective boundary values, we classify this as positive, i.e. the sequence was from a faulty sensor. The classification result (confusion matrix) is shown in Fig. 2. As the ground truth, there are 297 negative sequences (measured by healthy sensors) and 297 positive sequences (measured by faulty sensors) in the testing data. We can see that the true positive (TP) rate, or the recall score, is very high at 99.66 percent, and a good F1-score, i.e. F1 = 2×TP/(2×TP + FP+ FN) = 2×296/(2×296+23+1) = 0.961, which is in fact better than the results delivered by the deep learning based methods at considerably reduced computational cost and complexity.

When other rates of outliers are chosen, the results are summarized in Table 1. We can see the results at the outlier rate of 3 percent have improved with an F1-score of 0.9732. If we would know a priori the ground truth in the testing data, we could further optimize the F1-score by searching for the outlier rate. We found that the best F1-score of 0.98 can be achieved by setting the outlier rate to 3.6 while using the mean and STD values as the features. We can replace the STD with the skewness (a 3rd order statistic), where the corresponding results can be seen in Fig. 3 and Table 2. Obviously, replacing STD with skewness produces lower F1-scores. The mean plus skewness combination may not be able to detect the possible faulty-sensor sequence of all-zero values, which has a STD of zero and a skewness of infinity (or NAN – not a number).

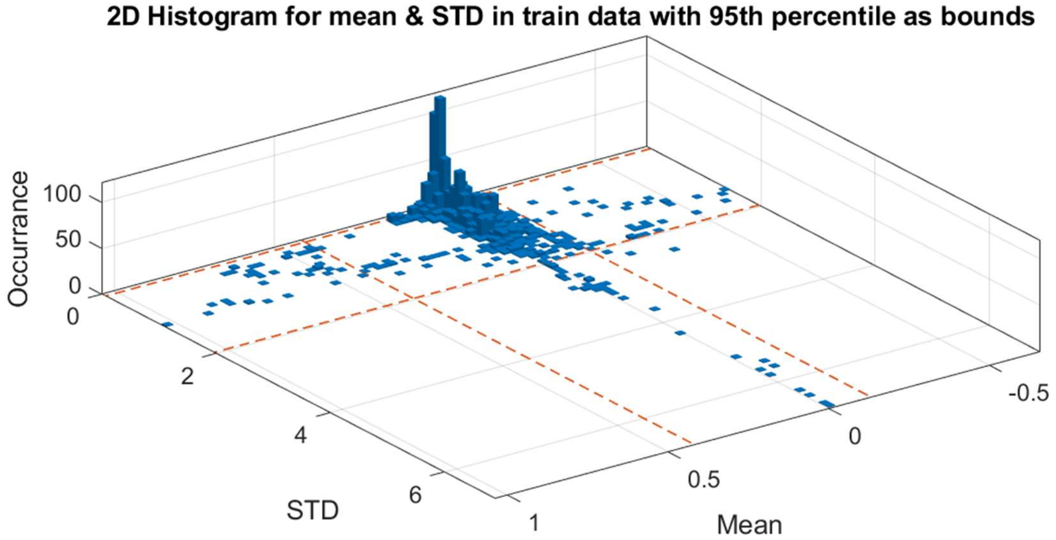

Fig. 1: 2D histogram of mean and STD values in the training dataset

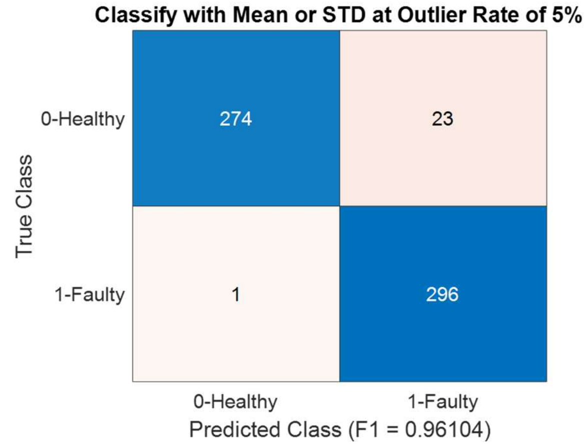

Fig. 2: Confusion matrix using a simple statistical analysis method with a recall score of 99.66% (296/297) and a F1-score of 96.1%.

Table 1: Classification with mean and STD as features

Rate of outliers (297+297=594)	True negative (TN)	False negative (FN)	False positive (FP)	True positive (TP)	F1-Score
1%	293 (98.65%)	32	4	265 (89.23%)	93.640%
3%	288 (96.97%)	7	9	290 (97.64%)	97.315%
5%	274 (92.26%)	23	1	296 (99.66%)	96.104%
3.6%					**98%**

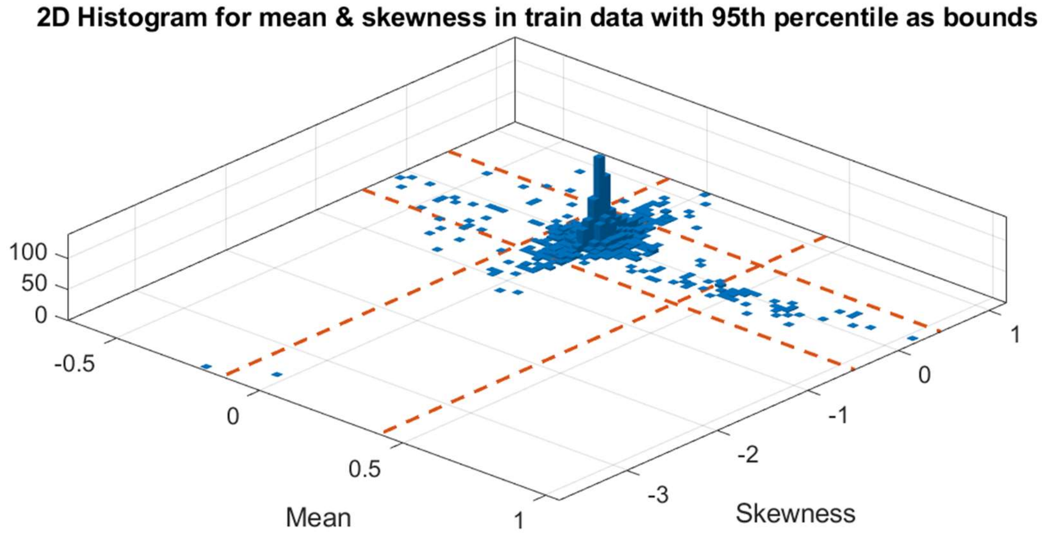

Fig. 3: 2D histogram of mean and skewness values in the training dataset

Table 2: Classification with mean and skewness as features

Rate of outliers (297+297=594)	True negative (TN)	False negative (FN)	False positive (FP)	True positive (TP)	F1-Score
1%	297 (100%)	43	0	254 (85.52%)	92.196%
3%	295 (99.33%)	34	2	263 (88.55%)	93.594%
5%	281 (94.61%)	31	16	266 (89.56%)	91.883%

Table 3: Classification with mean, STD and skewness as features

Rate of outliers (297+297=594)	True negative (TN)	False negative (FN)	False positive (FP)	True positive (TP), Recall	F1-Score
1%	293 (98.65%)	25	4	272 (91.58%)	94.939%
3%	286 (96.30%)	4	11	293 (98.65%)	97.504%
5%	268 (90.24%)	0	29	297 (100.0%)	95.345%
3.85%					**98%**
3.57%				**100%**	

The corresponding results with the 3-feature combination of mean, STD and skewness are listed in Table 3. We can see that an extra feature of skewness only improves the performance marginally from the mean and STD combination at the outlier rate of 3 percent, and produces weaker performance at the 1 and 5 percent outlier rates. Further with the 3-feature combination, we have found the highest recall score (100%) at outlier rate of 3.85 percent, and the highest F1-score (98.01%) at 3.57 percent outlier rate, as shown in Fig. 4. It is worth noting that the detection criterion is through an 'OR' logical operator, where as long as one of the features goes out of bounds, we will have a positive (or faulty sequence) detection. Perhaps the 'OR' operation among the 3 features made the difference for the superb performance by the simple statistical analysis method. In addition, we added the fourth feature of kurtosis (the 4th order statistic) and found that the added feature does not help improve the performance. Despite the fact that it would be difficult to choose the right outlier rate before knowing the ground truth *a* *priori*, we can demonstrate that the simple statistical analysis method with any reasonable outlier rate is at least comparable to the two deep learning based methods discussed previously.

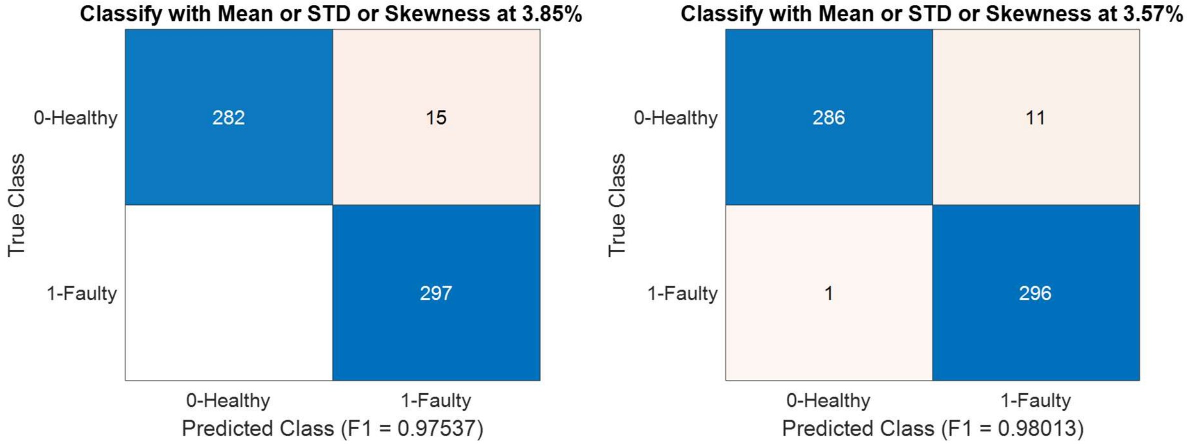

Fig. 4: Confusion matrices using a simple statistical analysis method with the highest recall score of 100% (left) and the highest F1-score of 98.013% (right).

One-class SVM Classification with Simple Statistics

We then tested a one-class support vector machines (SVM), with radial-basis-functions kernel and the boundary factor of υ=0.1 (usually the default value), using the same pre-computed statistical features (mean, STD, skewness and kurtosis), instead of using the entire sequences as input. We obtained similar results when using mean and STD with F1-score of 97%, see Fig. 5 as compared to the F1-score of 96% in Fig. 2. We also conducted a leave-one-out analysis to see which features were the most important out of the four tested, results are presented in Fig. 6. It seems that the most important feature is the mean, as when it is left out the accuracy drops significantly. Also, when leaving out the kurtosis, all 4 accuracy-metrics improve to almost 95% — in agreement with the previous results with simple statistical analysis method.

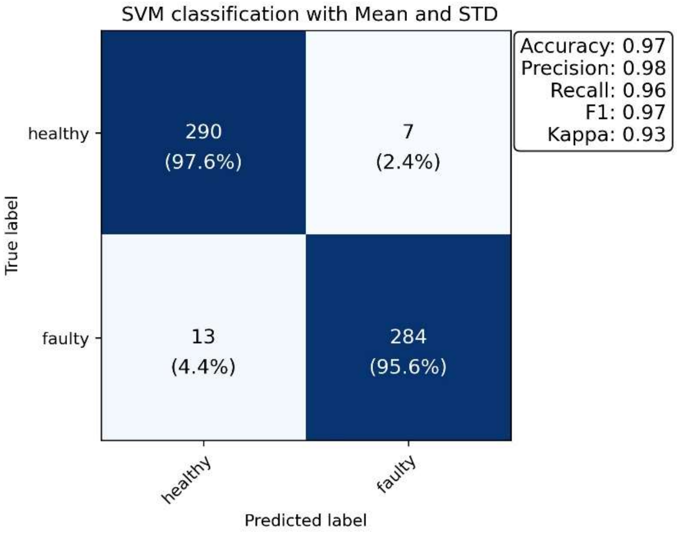

Fig. 5: Classification result using a one-class SVM with mean and skewness as features

We further show in Fig. 7 the performance of different combinations of statistical features, including mean-STD-skewness as well as the same combination including peak-to-peak (maximum minus minimum of each sequence). We can see that the mean-STD-skewness combination performs slightly worse than mean-STD only, while including peak-to-peak produces the best result with a recall score of 0.98 and F1-score of 0.98.

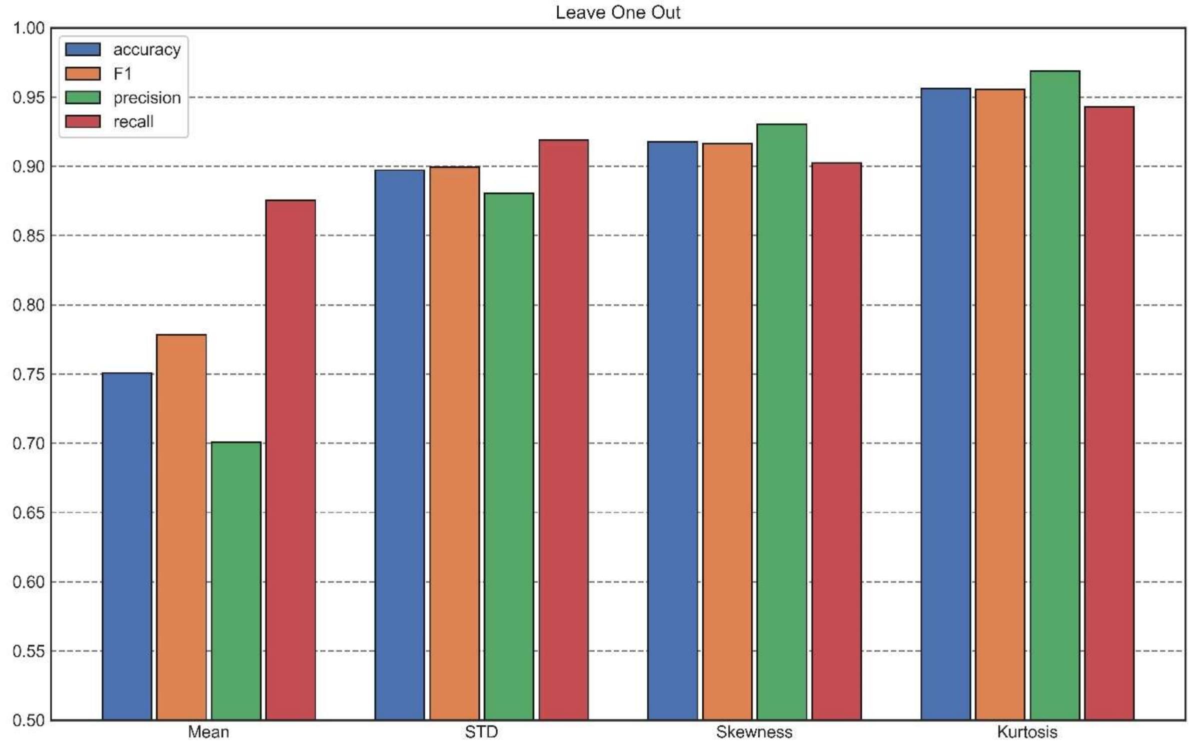

Fig. 6: Leave-one-out analysis showing the effect of excluding one of four different statistical features from the one-class SVM classification.

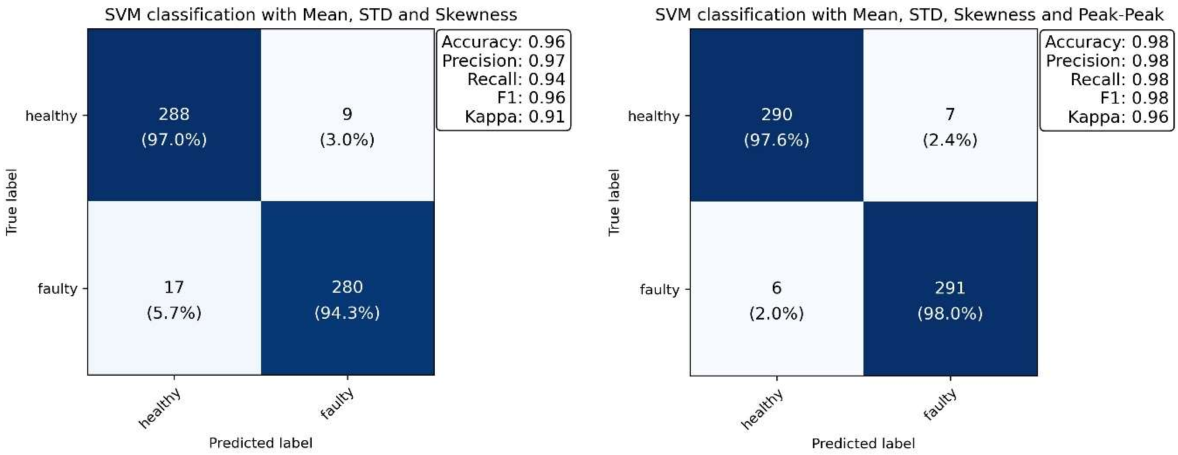

Fig. 7: Classification results of one-class SVM with different combinations of statistical features.

Concluding Remarks

In traditional machine learning, unsupervised classification can be a challenging problem. Deep learning based methods have demonstrated significant potential to address this challenge. With the Airbus dataset, deep learning methods, such as the GAN and CAE, can deliver good performance in classifying sensor data into ‘good’ or ‘bad’ when the deep learning models are trained on ‘good’ data only. In this paper, we are not disputing the effectiveness of deep learning methods, rather we want to remind people that traditional statistical analysis and machine learning methods can often perform as well as, and sometimes better than, the newer and more sophisticated deep learning methods. In the example of the Airbus dataset, most of our simple statistical analysis methods and the one-class SVM with simple statistical features can outperform the deep learning counterparts. However, one might argue that we could have seen the testing data prior to forming our framework, which would make the comparison unfair. Our counter argument would be to begin by using proven simpler methods as benchmarks before starting the journey of applying more complex and computationally expensive deep learning methods. For example, we have demonstrated that we could use mean and STD and 5% outlier in the training data, refer back to Fig. 1, to set up the boundaries/thresholds for anomaly detection without the need of any fine tuning by the testing data.

In conclusion, for machine health monitoring (MHM) problem, it is not necessarily true that deep learning methods are always superior to traditional methods, and it is a good practice to start solving a problem with simpler methods. Similar views and arguments can be found in a case study by Wang et al [4] in another DL application to MHM. Based on the results present here, we therefore recommend that where possible all future published studies of applying DL methods to MHM include appropriately selected traditional reference methods.

References

1. Wang, W., Taylor, J. and Rees, R, "Recent Advancement of Deep Learning Applications to Machine Condition Monitoring Part 1: A Critical Review." *Acoustics Australia* 49, 207–219 (2021), https://doi.org/10.1007/s40857-021-00222-9.
2. Li, D., Chen, D., Jin, B., Shi, L., Goh, J., Ng, S.K., "MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks." *Proceedings of International Conference on Artificial Neural Networks – ICANN 2019: Artificial Neural Networks and Machine Learning: Text and Time Series.* Lecture Notes in Computer Science, vol 11730. Springer, Cham. https://doi.org/10.1007/978-3-030-30490-4_56.
3. Garcia, G.R., Michau, G., Ducoffe, M., Gupta, J.S. & Fink, O., "Time series to images: monitoring the condition of industrial assets with deep learning image processing algorithms." *arXiv preprint* (2020). https://arxiv.org/abs/2005.07031v2.
4. Wang, W., Taylor, J. and Rees, R, "Recent Advancement of Deep Learning Applications to Machine Condition Monitoring Part 2: Supplement Views and a Case Study." *Acoustics Australia* 49, 221-228 (2021), https://doi.org/10.1007/s40857-021-00235-4.