A Report on: Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding

AVISHA BHIRYANI, Technical University Dortmund, Germany

ACM Reference Format:
Avisha Bhiryani. 2022. A Report on: Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. 1, 1 (February 2022), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

During the launch of Columbia orbiter STS-107 a piece of foam shed from the Shuttle external fuel tank struck the leading edge of the orbiter's left wing, compromising the thermal protection system.This turned out to be the very reason for the loss of the Orbiter during entry into the earth's atmosphere. One of the functions of the Data Processing System part of the Modular Auxiliary Data System(MADS) on board the Columbia Orbiter was to support annunciation of vehicle system failures and out-of-tolerance system conditions. [1]

Figure 1 depicts the Pressure taps on the lower left wing of the Orbiter during the launch clearly show that pressure tap V07P8073 among others has been hit at 84 seconds. This data was the first clue of something wrong with the system but was not reported anomalous by the system, so the crew on board as well as the ground crew were unaware about it. Spacecraft are exceptionally complex and expensive machines with thousands of telemetry channels detailing aspects such as temperature, radiation, power, instrumentation, and computational activities. Anomaly detection system are a critical tool here to monitor these channels and avoid potential hazards by reducing the amount of manual intervention needed.[1][3]

2 PAPER SUMMARY

Anomalies are the patterns that do not conform to the past behavior of the data. Anomaly detection methods for multivariate time series data can also be used for spacecraft telemetry data. Thus, challenges central to anomaly detection in multivariate time series data also hold for spacecraft telemetry. Data being monitored are often noisy, high-dimensional, highly non-stationary and dependent on current context. A lack of labelled anomalies necessitates the use of unsupervised or semi-supervised approaches. Specific to spacecraft, finding the source channel or system of anomalies and minimal number of false positives is also important. The authors of the paper have provided a 3-step procedure for anomaly detection in spacecraft telemetry data which can also be applied to any multivariate time series data. Firstly, LSTMs are used to predict high volume telemetry data per channel using inputs as previous telemetry values and encoded command information for the channel. Secondly, an unsupervised thresholding method is then used to automatically assess hundreds to thousands of diverse streams of telemetry data and determine whether

Author's address: Avisha Bhiryani, Technical University Dortmund, Dortmund, Germany, avisha.bhiryani@tu-dortmund.de.

Manuscript submitted to ACM

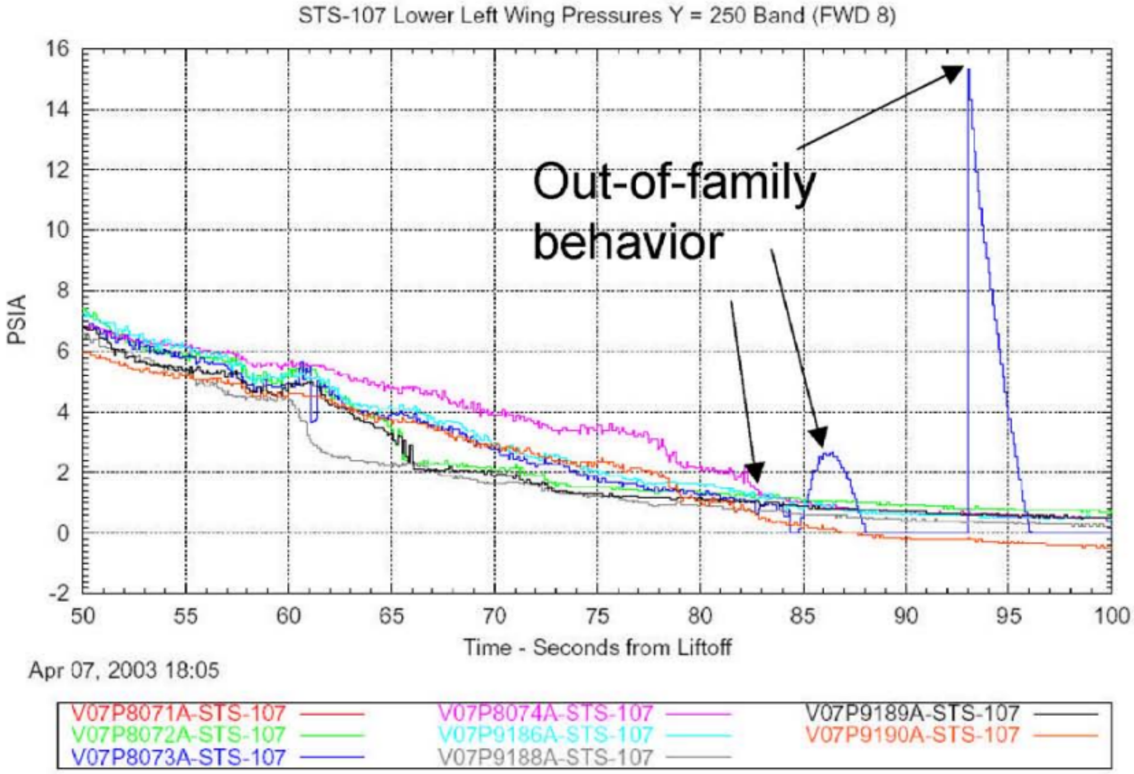

Fig. 1. Anomalous behaviour is clearly noticeable but was not communicated by the system. (https://history.nasa.gov/columbia/Troxell/Columbia).

resulting prediction errors represent spacecraft anomalies. Lastly, strategies for mitigating false positive anomalies are applied.[3][4]

2.1 LSTM - type of Recurrent Neural Network

At the heart of a LSTM is the presence of cell state and the different gates. Cell state is the memory of the network. It can carry relevant information throughout the processing of the sequence. So even information from the earlier time steps can make it's way to later time steps, reducing the effects of short-term memory. The Forget gate decides what is relevant to keep from prior steps, the input gate decides what information is relevant to add from the current step and the output gate determines what the next hidden state should be. The accuracy of an anomaly detection method largely depends on the time window or the amount of previous values considered. As LSTMs overcome Vanishing Gradient problem, they can bridge up to 1000 timestamps. Because of their ability to learn long term correlations in a sequence, they obviate the need for a pre-specified time window and are capable of accurately modelling complex multivariate sequences.[6] Due to the above mentioned properties of LSTM, a single model is created for each telemetry channel using LSTM and each model is used to predict values for that channel.

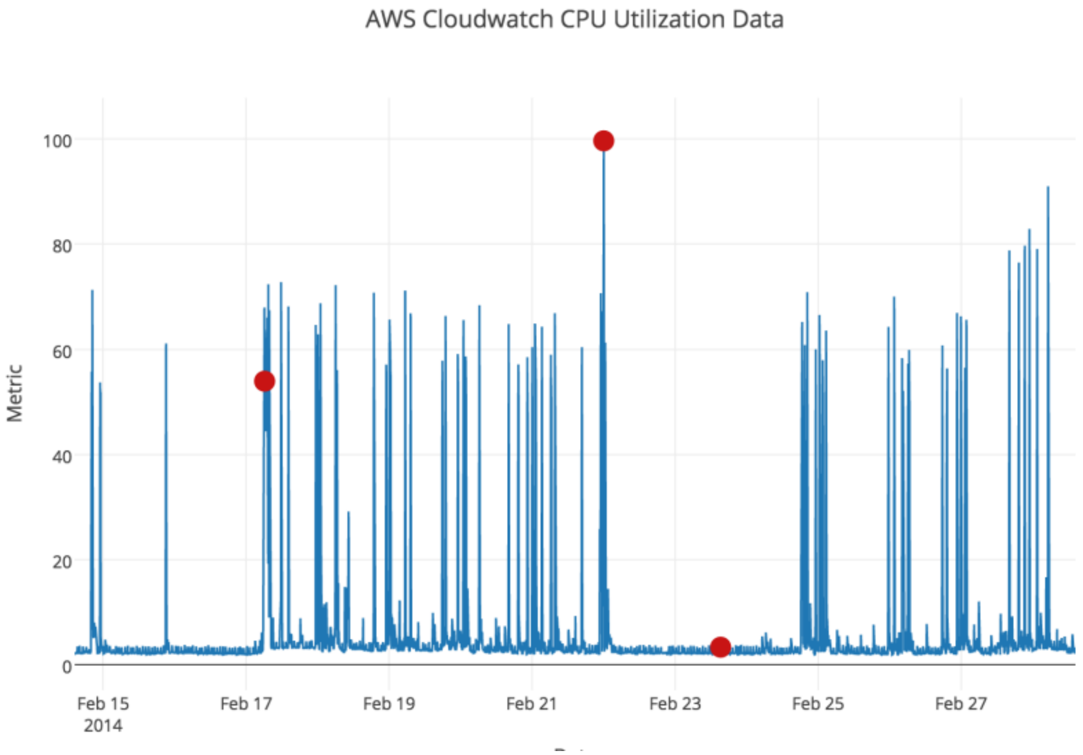

Fig. 2. Anomalies are labeled with red circles. The first anomaly in the is subtle and challenging. The spiking behavior does not return to the baseline as expected, and this is soon the new normal pattern. The second anomaly is a simple spike anomaly after which the system returns to previous patterns. The third anomaly identifies a long period inconsistent with the normal spiking pattern. ([4]).

2.2 Telemetry Value Prediction using LSTM

Every channel being monitored is modeled separately because LSTMs struggle to accurately predict high dimensional data. Modeling each channel independently also allows traceability down to the channel level. If the system were to be trained to detect anomalies at the subsystem level without this traceability, for example, operations engineers would still need to review a multitude of channels and alarms across the entire subsystem to find the source of the issue. Early stopping can be used to limit training to models and channels that show decreases in validation error.[3]

$$\hat{y}^{(t)} = \left\{ \begin{bmatrix} x_1^{(t-l_s)} \\ x_2^{(t-l_s)} \\ \vdots \\ x_m^{(t-l_s)} \end{bmatrix}, \ldots, \begin{bmatrix} x_1^{(t-1)} \\ x_2^{(t-1)} \\ \vdots \\ x_m^{(t-1)} \end{bmatrix}, \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \\ \vdots \\ x_m^{(t)} \end{bmatrix} \right\} \tag{1}$$

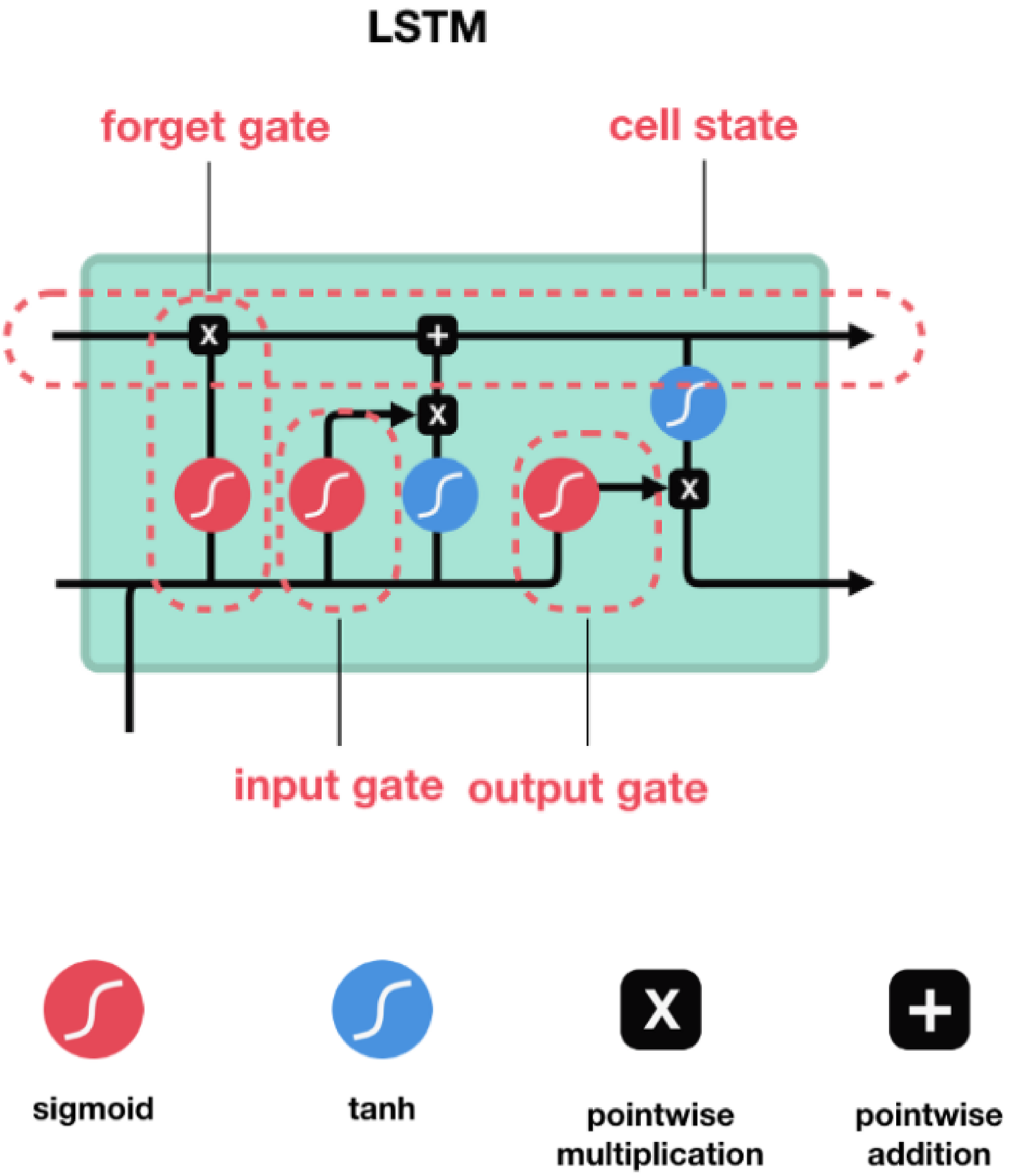

Fig. 3. Long Short Term Memory ([6]).

$$y^{(t)} = \begin{bmatrix} x_1^{(t+1)} \\ x_2^{(t+1)} \\ \vdots \\ x_m^{(t+1)} \end{bmatrix} \tag{2}$$

$$e^{(t)} = [\hat{y}^{(t)} - y^{(t)}] \tag{3}$$

$$e = [e^{(t-h)}, \ldots, e^{(t-l_s)}, \ldots, e^{(t)}] \tag{4}$$

Consider a time series $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$ where each step $\mathbf{x}^{(t)} \in R^m$ in the time series is an m-dimensional vector $\{x_1^{(1)}, x_2^{(2)}, \ldots, x_m^{(t)}\}$, whose elements correspond to input variables. For each point $\mathbf{x}^{(t)}$, a sequence length l_s determines the number of points to input into the model for prediction and a prediction length l_p determines the number of steps ahead to predict, where the number of dimensions d being predicted is $1 \leq d \leq m$. Since telemetry values for a single channel are predicted at a time $d = 1$ is considered. Also $l_p = 1$ is used to limit the number of predictions for each step t and decrease processing time. As a result, a single scalar prediction $\hat{y}^{(t)}$ is generated for the actual telemetry value at each step t. Once a predicted value $\hat{y}^{(t)}$ is generated for each step t, the prediction error is calculated as Equation 3. Each $e^{(t)}$ is appended to a one-dimensional vector of errors as in Equation 4 where h is the number of historical error values used to evaluate current errors. [3]

2.3 Dynamic Error Thresholding

The goal is to detect spacecraft anomalies that are only abnormally high or low smoothed prediction errors. Abrupt changes in values are often not perfectly predicted by LSTMs and result in sharp spikes in error values even when this behaviour is normal. So the set of errors **e** are smoothed to dampen spikes in errors that frequently occur with such predictions.[3]

$$e_s = [e_s^{(t-h)}, \ldots, e_s^{(t-l_s)}, \ldots, e_s^{(t-1)}, e_s^{(t)}] \tag{5}$$

Exponentially-weighted average (EWMA) is used to generate the smoothed errors in Equation 5. A threshold is selected and moothed errors above the threshold are classified as anomalies.[3]

$$\epsilon = \mu(e_s) + z\sigma(e_s) \tag{6}$$

$$\epsilon = argmax(\epsilon) = \frac{(\Delta\mu(e_s)/\mu(e_s)) + (\Delta\sigma(e_s)/\sigma(e_s))}{|e_a| + |E_{seq}^2|} \tag{7}$$

$$s^{(i)} = \frac{max(e_{seq}^{(i)}) - argmax(\epsilon)}{\mu(e_s) + \sigma(e_s)} \tag{8}$$

To determine the threshold:

- From Equation 6, values evaluated for ϵ are determined using z which is an ordered set of positive values representing the number of standard deviations above $\mu(e_s)$ and experimental results show values from two to ten to work.
- From Equation 7, a threshold is found such that, if all values above are removed, that would cause the greatest percent decrease in the mean and standard deviation of the smoothed errors e_s . The function also penalizes for having larger numbers of anomalous values $|e_a|$ and sequences $|E_{seq}|$ to prevent overly greedy behavior.
- From Equation 8, indicating the severity of the anomaly, the highest smoothed error in each sequence of anomalous errors is given a normalized score based on its distance from the chosen threshold.[3]

2.4 Mitigating False Positives

The precision of prediction-based anomaly detection approaches heavily depends on the amount of historical data that is considered. If large amount of historical data is taken into account it becomes expensive to query and process historical data in real-time scenarios and a lack of history can lead to false positives that are only deemed anomalous because of the narrow context in which they are evaluated.

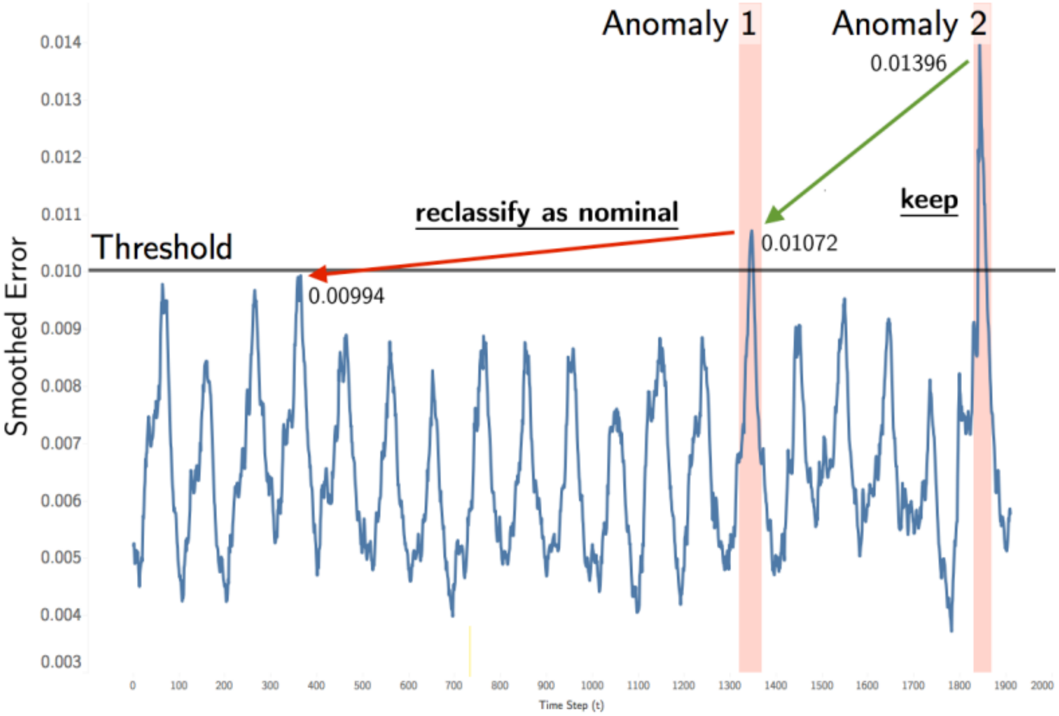

Fig. 4. Pruning Method [3]

$$e_{max} = [0.01396, 0.01072, 0.00994]$$
$$p = 0.1$$
$$d^{(1)} = 0.23 > p$$
$$d^{(2)} = 0.07 < p \tag{9}$$

To mitigate false positives and limit memory usage and computation cost, a pruning procedure in which a new set, e_{max} , is created containing $max(e_{seq})$ for all error sequences sorted in descending order. Also, a maximum smoothed error that is not anomalous is added to the end of e_{max} . The sequence is then stepped through incrementally and the percent decrease between the consecutive values of e_{max} at each step is calculated. If at some step a minimum percentage decrease p is exceeded by $d^{(i)}$, all errors and their corresponding anomaly sequences remain anomalies. If the minimum decrease p is not met, then all those smoothed error sequences are reclassified as nominal.[3] A second strategy for limiting false positives can be employed once a small amount of anomaly labeled data has been gathered. Based on the assumption that anomalies of similar magnitude s generally are not frequently recurring within the same channel, a minimum score s_{min} is set such that future anomalies are re-classified as nominal if $s < s_{min}$.[3]

2.5 Experiment Results

For testing purposes, data from spacecrafts such as SMAP and MSL are used. Both these spacecrafts are very different missions representing varying degrees of difficulty when it comes to anomaly detection. Compared to MSL, operations for the SMAP spacecraft are routine and resulting telemetry can be more easily predicted with less training and less data. MSL performs a much wider variety of behaviors with varying regularity. For both of them, Incident Surprise Anomaly Report were used to obtain labeled anomalies.If multiple anomalous sequences and channels closely resembled each other, only one was kept for the experiment in order to create a diverse and balanced set.[3] A 5 day span is considered around the anomalies, primary anomaly occuring time t_a and the span is $t_s = t_a - 3d$ to $t_f = t_a + 2d$. Each labeled anomalous sequence of telemetry values is compared against the final set of predicted anomalous sequences according to the following:

- True Positive: Only one true positive is recorded even if portions of multiple predicted sequences fall within a labeled sequence.
- False Positive: For all predicted sequences that do not overlap a labeled anomalous region.

Table 1. Model Parameters [3]

Hidden Layers	2
Units in Hidden Layers	80
Sequence Length	250
Training Iterations	35
Drop out	0.3
Batch Size	64
Optimizer	Adam

2.5.1 Model Parameters. Telemetry values are aggregated into 1-minute windows and evaluated in batches of 70 minutes and each 70 minute batch of values is evaluated using 2100 prior values to calculate an error. Each model is shallow with only two hidden layers and 80 units in each layer. The p parameter is an important parameter to control precision and recall, and an appropriate value can be inferred when labels are available. Here, reasonable results were achieved with $0.05 < p < 0.20$.[3]

2.5.2 Final Experiment Result. From Table 2, for the non parametric approach pruning only decreases overall recall by 4.8% (84.8% to 80.0%) while increasing overall precision by 38.6 %(48.9% to 87.5%). If LSTM predictions are poor and resulting smoothed errors do not contain a signal then thresholding methods will be ineffective. The Gaussian tail approach results in lower levels of precision and recall using various parameter settings. Pruning greatly improves precision but at a high recall cost, resulting in an F(0.5) score that is still well below the score achieved by the non-parametric approach with pruning.[3]

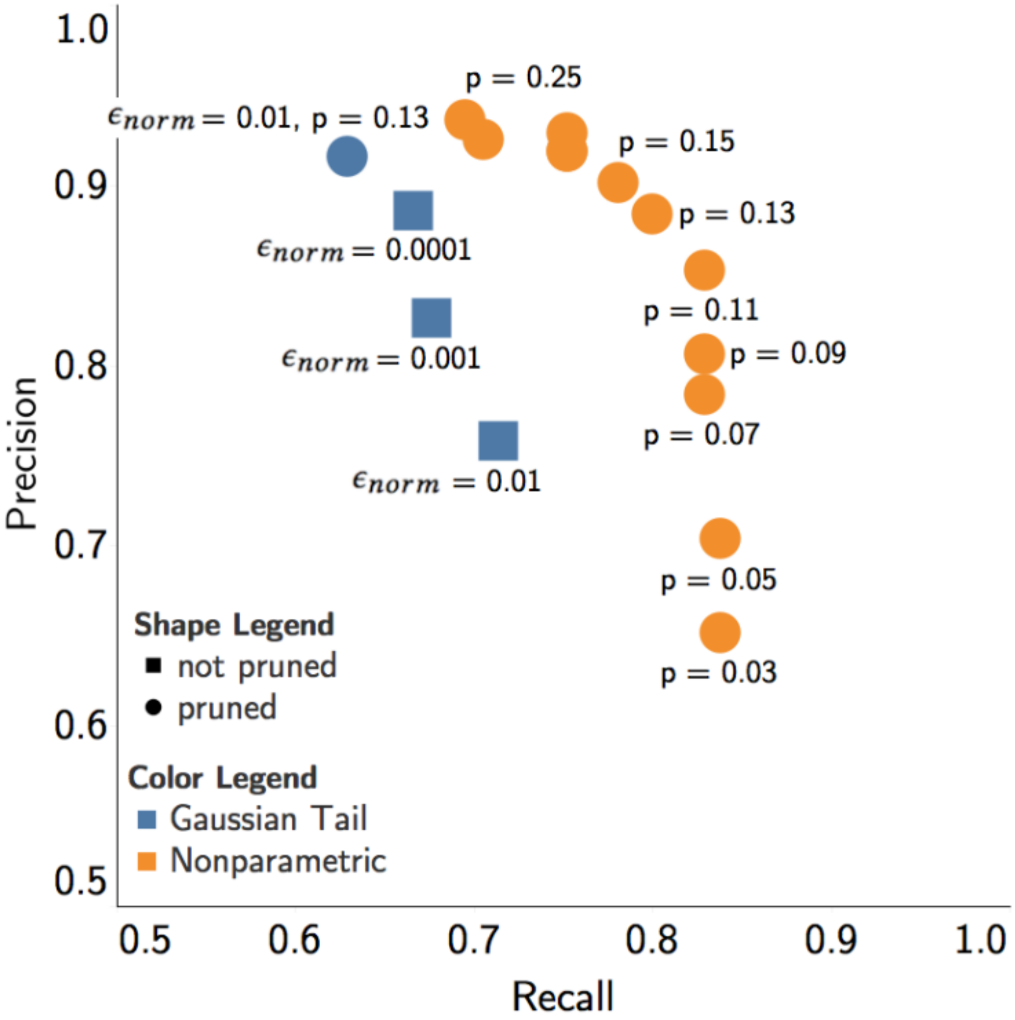

Fig. 5. Overall precision and recall for parametric and non-parameteric approach [3]

3 OTHER ANOMALY DETECTION METHODS FOR SPACECRAFT DATA

3.1 Out of Limits(OOL)

The Out of Limits approach consists of defining an upper and lower threshold so that when a measurement goes above the upper limit or below the lower one, an alarm is triggered. Then engineers will inspect the parameter that is out of limits and determine whether it is an anomaly or not and decide which action to take. It is still the most popular approach because low computational expense, broad and straight-forward applicability and ease of understanding.

Issues with this approach are that as seen in Figure 6,some behaviors are anomalous even if they are within the defined limits. A lot of manual intervention is still needed with this approach.[5]

Table 2. Result for each spacecraft [3]

Thresholding Approach	Precision	Recall	F(0.5) score
Non parametric with Pruning(p = 0.13)			
MSL SMAP Total	92.6% 85.5% 87.5%	69.4% 85.5% 80%	0.69 0.71 0.71
Non parametric without Pruning(p = 0)			
MSL SMAP Total	75.8% 43.0% 48.9%	69.4% 92.8% 84.8%	0.61 0.44 0.47
Gaussian tail(ϵ_{norm} = 0.0001)			
MSL SMAP Total	84.2% 88.5% 87.5%	44.4% 78.3% 66.7%	0.54 0.71 0.66
Gaussian tail(ϵ_{norm} = 0.01)			
MSL SMAP Total	61.3% 82.4% 75.8%	52.8% 81.2% 71.4%	0.48 0.68 0.62
Gaussian tail with Pruning($\epsilon_{norm} = 0.01, p = 0.13$)			
MSL SMAP Total	88.2% 92.7% 91.7%	41.7% 73.9% 62.9%	0.54 0.71 0.66

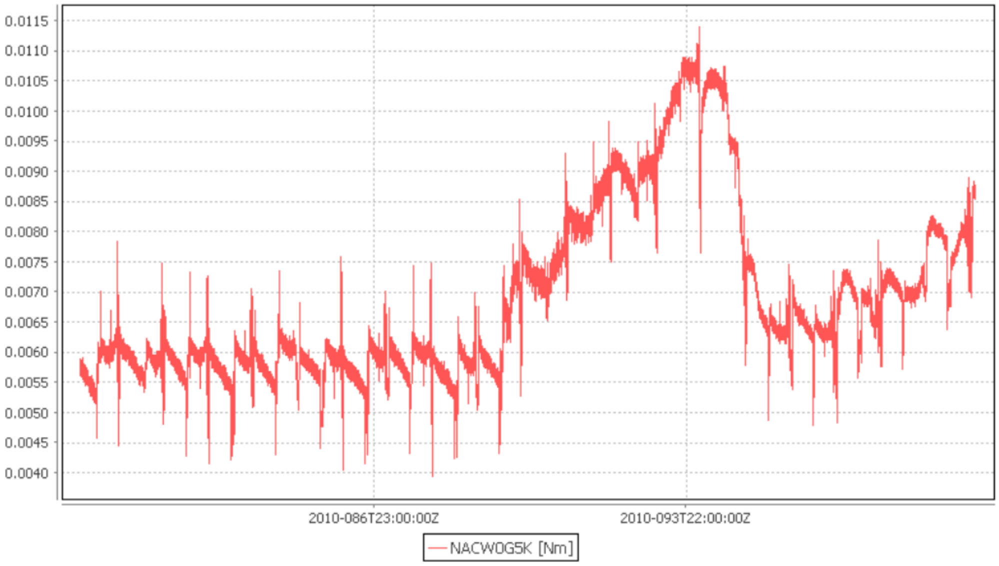

Fig. 6. OOL Approach Limitation [5]

3.2 Nearest Neighbor Based Approach

Orca(currently being used on ISS) uses a nearest-neighbour approach to search for unusual data points in multivariate data sets by calculating the distance of each data point from neighbouring points. Distance between points is measured with the Euclidean distance measure for continuous parameters and the Hamming distance for discrete parameters. The program outputs a score for each point representing the average distance to the nearest k neighbors in the data set. The value of k is specified by the user. Points that have a larger average distance to their nearest neighbors than most other points in the data set are considered anomalies.[2] Issues with this approach are:

- The number of past values to be considered needs to be decided.
- High computational complexity and memory consumption as distances of each point with all other points is to be calculated.
- Can only find global anomalies, does not consider local neighborhood.

3.3 Clustering Based Approach

The IMS(Inductive Monitoring System)(currently being used on ISS) tool uses clustering to analyze archived spacecraft data and characterize nominal interactions between selected parameters. This characterization, or model, is compared with real time or archived system data to detect off nominal behavior. [2]

4 CRITICISM

Table 3. LSTM-NDT(Independent Channel) and MTAD-TF(Dependent Channel) [7]

Dataset	Precision	Recall	F1 score
LSTM-NDT			
MSL	0.5944	0.5374	0.5640
SMAP	0.8965	0.8846	0.8905
SMD	0.5684	0.6438	0.6037
MTAD-TF			
MSL	0.9043	0.8988	0.9015
SMAP	0.9779	0.8192	0.8916
SMD	0.9045	0.9048	0.8940

Limitations with the approach:

- From Table 3, LSTM-NDT has a high score on SMAP, but it performs poorly on MSL and SMD, reflecting that the model is very sensitive to different scenarios[7]. As one of the goals of the authors of the paper was to build an anomaly detection system that can be very well applied to not only different spacecrafts data but also for any multivariate time series data. This goal is not being met.
- Along with the telemetry data, channel specific command information is also an input to the model that the model is very much dependent on. Refining it would lead to more accurate anomaly detection.
- The dependencies between different channels is not taken into account.

5 SUMMARY

An anomaly detection system that caters to the particular need of a spacecraft system considering the time and precision critical environment of a spacecraft has been proposed in the paper I have reported on. Using state of the art deep learning algorithm Long Short Term Memory, gives this approach an advantage of not being heavily dependent on the amount of historical data that can be used to predict telemetry data. A dynamic thresholding approach has been proposed that does not make any assumptions about the prediction error distribution. This makes this approach superior to others that assume data to be normally distributed. Considering the key areas of improvement as those mentioned in the **??** section are addressed in the future refinements of the approach, it can be well applied to a large number of varied anomaly detection tasks in multivariate time series data.

REFERENCES

[1] 2003. *Columbia Accident Investigation Board Report, Volume 3.* Technical Report. NASA.

[2] David Iverson. 2008. Data Mining Applications for Space Mission Operations System Health Monitoring. *SpaceOps 2008 Conference* (2008). https://doi.org/10.2514/6.2008-3212

[3] Christopher Laporte Ian Colwell Tom Soderstrom Kyle Hundman, Valentino Constantinou. 2018. Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (2018), 9. https://doi.org/10.1145/3219819.3219845

[4] A. Lavin and S. Ahmad. 2015. Evaluating Real-time Anomaly Detection Algorithms – the Numenta Anomaly Benchmark. *14th International Conference on Machine Learning and Applications (IEEE ICMLA'15)* (2015). https://github.com/numenta/NAB

[5] Jose MartÃŋnez-Heras and Alessandro Donati. 2014. Enhanced Telemetry Monitoring with Novelty Detection. 35 (2014).

[6] Michael Phi. 2018. *Illustrated Guide to LSTM's and GRU's: A step by step explanation.* https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

[7] C.L. Zhang and H. Y. Wang Q. He, Y. J. Zheng. 2020. MTAD-TF: Multivariate Time Series Anomaly Detection Using the Combination of Temporal Pattern and Feature Pattern. *Hindawi Complexity Volume 2020* 2020, Article 8846608 (2020), 9 pages. https://doi.org/10.1155/2020/8846608