Multiforecast-based Early Anomaly Detection for Spacecraft Health Monitoring

Prajjwal Yash
prajjwal@ursc.gov.in
U R Rao Satellite Centre
Bengaluru, India

Sharvari Gundawar
sharvari@ursc.gov.in
U R Rao Satellite Centre
Bengaluru, India

Nitish Kumar
nitish@ursc.gov.in
U R Rao Satellite Centre
Bengaluru, India

Uma B R
uma@ursc.gov.in
U R Rao Satellite Centre
Bengaluru, India

Krishna Priya G
priya@ursc.gov.in
U R Rao Satellite Centre
Bengaluru, India

Purushottam Kar
purushot@cse.iitk.ac.in
Indian Institute of Technology Kanpur
Kanpur, India

ABSTRACT

Early detection of impending anomalies is a strong desirable for spacecraft operation as it can allow preemptive action to safeguard the mission objectives. Methods abound for just-in-time anomaly detection but early detection is a much more sought after goal. In this paper, we present MEND, a simple-yet-powerful model for early anomaly detection for spacecraft health monitoring. In experiments, MEND was able to provide strong alerts for impending anomalies as much as 10-15 minutes before the onset of the anomaly which could give a system admin valuable time to perform curative action. It is notable that none of the other models considered, including state-of-the-art zero-shot time series prediction models, were able to achieve this. MEND is based on simple, explainable elements such as self-supervised operation-mode detection and self-disagreement-based anomaly detection which, as a side effect, offer insights that may aid root cause analysis and may be of independent interest. Code for MEND is available at https://github.com/purushottamkar/mend

CCS CONCEPTS

• **Computing methodologies** → **Anomaly detection**.

KEYWORDS

spacecraft health monitoring, time series learning, anomaly detection, self-disagreement, clustering, regression analysis

ACM Reference Format:
Prajjwal Yash, Sharvari Gundawar, Nitish Kumar, Uma B R, Krishna Priya G, and Purushottam Kar. 2024. Multiforecast-based Early Anomaly Detection for Spacecraft Health Monitoring. In *7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD) (CODS-COMAD 2024), January 04–07, 2024, Bangalore, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3632410.3632458

1 INTRODUCTION

The holy grail of space exploration is a system that achieves complete autonomy and allow the spacecraft to become self-sustaining and capable of achieving mission objectives that span months, years or even decades, without requiring much or any human feedback or intervention. However, most current missions operate in a hybrid mode wherein there do exist on-board controllers that can perform minor remediation such as correcting the sun-angle, but vast amounts of telemetry data are still sent to mission control, in the form of multi-channel time series, to be inspected manually (often continuously) so that anomalies can be identified, and remedial actions be signalled back to the spacecraft.

Early and automated detection of anomalies thus becomes critical in the present setup as it can enable ground staff to quickly respond and take preemptive or curative action to safeguard the craft as well as the mission. There has been a lot of work on time series modelling and anomaly detection but a lot of it focuses on *just-in-time* anomaly detection which does not leave the system admin with much headroom to take curative action. For example, automated spacecraft health monitoring based on out-of-limit (OOL) detection on telemetry has been a popular paradigm but offers limited utility as it requires a fault in the subsystem to manifest at significant levels before it can be detected.

Our Contributions. We develop MEND for early detection of anomalies and present a case study on the power subsystems of low-Earth orbit (LEO) satellites. MEND incorporates novel elements:

(1) A novel R-DTW metric to reduce instances of false positives.
(2) Unsupervised detection of modes of operation of the spacecraft subsystem being studied that reduces reliance on expert input and manual intervention. This also offers key insights into subsystem performance for designers and monitors.
(3) A novel *self disagreement*-based anomaly detection paradigm that reduces MEND's reliance on real-time telemetry which may be unavailable or delayed in the event of an anomaly.
(4) MEND is light enough to be deployed on a spacecraft yet detects anomalies as much as 10-15 minutes ahead of time.
(5) MEND outperforms several state-of-the-art method [3, 6, 8] including those that utilize deep models.

Takeaways from the Case Study. Some specific lessons that can be taken from this study are summarized here:

(1) Although univariate modelling using powerful non-linear models is popular in literature, especially given its convenience, multivariate modelling can offer superior performance even with simpler, explainable models.

CODS-COMAD 2024, January 04–07, 2024, Bangalore, India

ACM ISBN 979-8-4007-1634-8/24/01…$15.00
https://doi.org/10.1145/3632410.3632458

(2) Infusing domain knowledge, such as identifying the modes of operation of a system, into the anomaly detection system can build powerful models. MEND demonstrates how these modes can be identified without much human supervision.
(3) Self-disagreement based anomaly detection can effectively detect anomalies in advance while reducing the reliance on real-time telemetry.

On multiple experiments on the electrical battery subsystem of a low-Earth orbit (LEO) satellite, MEND outperformed several competitor techniques including state-of-the-art zero-shot time series models. MEND outperformed its competitors in terms of modelling nominal behavior as well as in terms of early prediction of anomalies. MEND consistently offered alerts as much as 10-15 minutes before the onset of the anomaly which competitor algorithms were not able to approach.

1.1 Related Works

Time series analysis has a rich history, featuring traditional models like Vector Autoregression (VAR) and Vector Autoregressive Integrated Moving Average (VARIMA) [7]. Modern non-linear models, such as Long Short-Term Memory networks (LSTMs) [6] and the Neural Basis Expansion Analysis for Time Series (N-BEATS) method [9], including its recent extension NBEATSx [8], have gained prominence for their versatility in handling time series data, including those with exogenous variables.

Anomaly detection in time series data has been extensively studied. Common approaches involve using thresholds to identify anomalies, with various methods for determining these thresholds, such as non-parametric thresholding [6] and oblique thresholds [2]. Recent works have also explored the use of sparse representation techniques along with anomaly detection techniques such as one-class SVM[4] and transfer learning principles [3]. Dissimilarity metrics, including the well-known Dynamic Time Warp (DTW) distance and its efficient variant, FastDTW [11], are crucial for comparing time series sequences for discrepancies.

The availability of user-friendly packages has accelerated research in time series analysis. Widely used tools like scikit-learn [10], statsmodels [13], tslearn [15], and Darts [5] facilitated working with time series data by providing a variety of algorithms and analysis tools.

The proposed method MEND capitalizes on these accessible tools and introduces an early anomaly detection algorithm that can outperform leading models like NBEATS and traditional models such as AutoRegressive (AR) when exogenous variables are involved.

2 PROBLEM SETTING

The battery subsystem of a spacecraft is a critical component that powers different instruments and other subsystems within the spacecraft. Spacecrafts require self-contained power systems as they operate far away from traditional power sources such as electrical grids.

The battery voltage and current of a low-earth orbit (LEO) spacecraft are tracked as endogenous variables whose temporal evolution are known to interact with each other. Additionally, one or more exogenous variables (abbreviated as exog hereon) are also available. These may consist of planned payload operations that are known to impact battery voltage and current. Two points are noteworthy about these exog variables. Firstly, they are independent variables in that their value is unaffected by the battery parameters but they themselves do affect the battery parameters. Secondly, the value of these exog variables is known well-ahead of time since these can be derived from telecommands sent from ground stations that, for example, plan payload operations well in advance. The time-series is subsampled at intervals of 30s after eliminating telemetry breaks from the dataset. A whole orbit for an LEO satellite lasts around 100 minutes, and thus 200 timestamps in the time series constitute roughly one complete Earth orbit.

Table 1: Dataset Details

Number of Parameters	3 (2 endogenous and 1 exogenous)
Training + Validation Data	3 days (80:20 split)
Resolution	30 seconds
Nominal Day 1	1 day of nominal performance
Nominal Day 2	1 day of nominal performance
Anomalous Day 1	Onset of anomaly at timestamp 1850.
Anomalous Day 2	Onset of anomaly at timestamp 90.

The case studies presented in this paper are related to the charging and discharging of the battery systems in the spacecraft S/C A. The actual date-time stamps in the data-set were replaced with sequential integers in line with the data-sensitivity policy of the Indian Space Research Organization. Four data-sets were compiled as a result along with the training and validation data-set:

The anomalies in the dataset are associated with poor charging of the battery system that was close to its end-of-life (EOL). It is notable that an anomaly detection system based on classical Out-Of-Limit (OOL) analysis of the telemetry could not have detected the anomalous behavior in the system since the time series values were always within bounds in these datasets. Instead, what was anomalous was the charging and discharging curves and their characteristics.

The goal in this paper is to develop an accurate, explainable model that is able to generate alerts for impending anomalies well-ahead of their onset.

3 MEND: MULTIFORECAST-BASED EARLY ANOMALY DETECTION

Data Description. As a part of the dataset, we are offered time series for battery voltage $\{v_t\}$, battery current $\{c_t\}$ and the exogenous variable $\{\mathbf{e}_t\}$. The battery current and voltage are univariate time series but the exogenous variable may be multivariate. The MEND algorithm first extracts augmented features $\{\mathbf{a}_t\}$ from the endogenous features (which are themselves multivariate), treats these as endogeneous variables and then performs multivariate modelling of the entire set of endogenous variables i.e., $(v_t, c_t, \mathbf{a}_t)$ while using instantaneous values of the exogenous variable $\{\mathbf{e}_t\}$.

3.1 Domain-specific Augmented Features for Anomaly Detection

Most spacecraft subsystems exhibit multiple modes of operation that can be characterized by observable parameters such as relative

orientation and attitude, operation of payload or other specifications defined by the spacecraft mission. Being able to recognize these modes automatically can be beneficial to modelling, anomaly detection and subsystem health monitoring.

Modes of Operation in the Case Study. Modular subsystems are often designed to persistently transit among several modes of operation that are defined and baked into the system design. For example, the three known modes of operation for the LEO spacecraft power subsystems under consideration are fast charge, slow charge and discharge. The exact mode of operation is determined by various factors such as depth of discharge of the battery, whether the spacecraft is experiencing an eclipse or not or whether any onboard payload operations are underway or not. MEND aims to learn the evolution of these modes of operation as a result of the interplay of the endogenous parameters.

Unsupervised Mode Identification. MEND employs unsupervised learning techniques such as time-series clustering to automatically identify the various modes of operations of the subsystem. This allows the algorithm to track transitions between modes which in turn helps in the early detection of anomalies. The endogeneous parameters as well as their derivatives i.e., $\{\mathbf{f}_t\} \stackrel{\text{def}}{=} \{(v_t, c_t, v'_t, c'_t)\}$ were used to identify the modes of operation in an unsupervised manner by using a time-series clustering algorithm [15]. Note that the derivatives of v_t and c_t signify the rate of change of the respective endogenous variables. Such derivatives are useful in capturing the dynamics of several systems. However for power subsystems in particular, these hold special significance since these can capture rates of battery charge and discharge, and allow easy identification of anomalies that either lead to slow charge or rapid discharge.

The algorithm slices the time series (using a rolling window with a stride) and creates K clusters (where K is a hyperparameter) each with its cluster center $\{\mathbf{c}_k\}_{k\in[K]}$. Having received these cluster centers, each slice $\mathbf{f}_t$ of the time series is endowed with a *soft* assignment to these clusters by calculating the vector $\mathbf{a}_t \in \mathbb{R}_+^K$ where $\mathbf{a}_t = [a_{tk}]$ defined as $a_{tk} \stackrel{\text{def}}{=} (\text{DTW}(\mathbf{f}_t, \mathbf{c}_k))^- 1$ where DTW is the dynamic time warp distance [1] between the time series $\mathbf{c}_k$ and $\mathbf{f}_t$. Note that this augmented feature tells us what was the dominant mode of operation in the time slice t.

For experiments, the number of clusters was chosen to be $K = 3$ using an elbow plot (see Figure 1) and an implementation of the Kneedle algorithm [12] from the scikit-learn library [10]. Upon inspection, it was confirmed that the identified clusters were genuinely distinct and corresponded to the three well-known modes of operation of the power system battery of this spacecraft, namely fast charge, slow charge and discharge. We shall see that integrating this domain-knowledge leads to the generation of essential insights. When faced with test data, the augmented features are generated similarly by evaluating the inverse DTW distance of the telemetry slices from these cluster centers.

3.2 MEND: Multivariate Modelling

MEND employs the vector autoregression (VAR) model [14] implemented using a [13] time-series wrapper in order to learn the nominal operation of the subsystem in a self-supervised manner. The VAR model has been shown to be effective in modelling physical systems and continue to be popular in the forecasting community

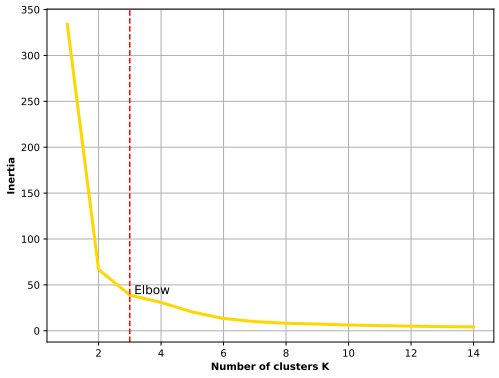


Figure 1: Elbow plot for time-series k-means clustering: system has three distinct modes of operation

[9]. They also have small memory and compute footprint which makes them ideal for resource-constrained environments like onboard deployment on spacecraft where computational resources, power consumption, and real-time processing constraints are significant factors. The model learns the trajectory of the multivariate time-series $\{v_t, c_t, \mathbf{a}_t\}$ as a function of the past values and exogenous parameters $\{\mathbf{e}_t\}$. The Akaike Information Criteria (AIC) was employed to select the optimal lag (order of the model L) as 100. As mentioned earlier, the exog parameter ($\mathbf{e}_t$) values were known well in advance and their instantaneous values were used to train the model. A linear VAR model of order L with $M+K$ model parameters was obtained that generated a forecast at timestamp t as

$$\mathbf{y}_t = A \cdot \mathbf{x}_t + \mathbf{b},$$

where $\mathbf{b} \in \mathbb{R}^{M+K}$ is the model bias, $\mathbf{y}_t \in \mathbb{R}^{M+K}$ are the model predictions, $\mathbf{x}_t \in \mathbb{R}^{L(M+K+E)\times 1}$ is the vector of lagged telemetry, augmented features and exogenous variables reshaped as a vector, $E = \dim(\mathbf{e}_t)$ is the dimensionality of the exogenous variable, and $A \in \mathbb{R}^{(M+K)\times L(M+K+E)}$ is the linear model. For MEND, $L = 100$ as chosen above, and $M = 2$ since the 2 endogenous parameters and $K = 3$ as those many augmented parameters are being modelled. We note that the instantaneous exogenous variables are assumed to be always available, as discussed in Section 2. To reiterate, the endogenous variables of this system are battery current, voltage, and the augmented features.

Three days of nominal telemetry was used to train the VAR model. A test case is presented in subsequent sections in which the power system's health was monitored for a spacecraft. The creation of augmented features as the modes of operation turns out to be an integral part of MEND's architecture. Note that including the augmented features as endogenous variables implies that during training, the model is being trained to learn how do the modes of operation evolve alongside the telemetry.

Forecasting at Test Time. The forecasting horizon for a time-series model (i.e. how far ahead in time is it asked to predict the modelled parameters) plays a critical role in early detection of anomalous behavior. A small forecasting horizon would make the

model short-sighted and unable to give anomaly alerts early enough. However, a forecasting horizon that is too large could also be suboptimal since if the model yields noisy predictions at large forecasting horizons, then it would be unable to distinguish its own prediction errors from an incipient anomaly. MEND works with a forecasting horizon of 100 steps by regressing on its own predictions; it generates forecast for a single timestamp using true telemetry data and the learnt model and uses this forecast to extend the prediction horizon by one timestamp and so on.

Specifically, suppose $\mathbf{x}_t = [\mathbf{n}_{t-1}, \mathbf{n}_{t-2}, \ldots, \mathbf{n}_{t-L}]$ where $\mathbf{n}_\tau \in \mathbb{R}^{M+K+E}$ is the set of endogenous and exogenous variables at timestamp t, then first we obtain the prediction $\mathbf{y}_t \in \mathbb{R}^{M+K}$ for the next time step. Then, a new pseudo vector is created as $\tilde{\mathbf{x}}_{t+1} = [[\mathbf{y}_t, \mathbf{e}_t], \mathbf{n}_{t-1}, \ldots, \mathbf{n}_{t-L+1}]$, where $\mathbf{e}_t$ is the instantaneous value of the exogenous variable. $\tilde{\mathbf{x}}_{t+1}$ is now used alongside the VAR model to generate $\tilde{\mathbf{y}}_{t+1}$ and the process is repeated till the prediction horizon is exhausted. This process allows MEND to use the previous 100 timestamps (roughly 50 minutes) of telemetry and exog variables, alongwith the next 100 timestamps of exog, to forecast the next 100 timestamps of telemetry and augmented features.

Multiforecasts for Early Anomaly Detection. MEND reduces its reliance on real-time telemetry by adopting a strategy that uses self-disagreement to detect anomalies. Specifically, the model is used to create multiple forecasts at the same timestamp with a stride. Put simply, the timestamps $[t-L, t]$ are used to obtain predictions for the timestamps $[t, t+L]$, then the timestamps $[t-L+s, t+s]$ are used to obtain predictions for the timestamps $[t+s, t+L+s]$, then $[t-L+2s, t+2s]$ are used to predict $[t+2s, t+L+2s]$ and so on, where s is the stride length. Notice that this gives us multiple predictions for any future timestamp. These predictions are analysed and disagreements are used to sound alarm for an impending anomaly.

Advantages of Multi-forecasts. Multi-forecasts give MEND several advantages such as recognizing early signatures of deviation from nominal behavior and eliminating dependence on real-time telemetry to check for discrepancies. The second point is a bit subtle and merits more discussion: anomaly detection systems that rely on their predictions not matching actual telemetry before sounding an alarm are heavily dependent on the availability of real-time telemetry. If for some reason telemetry is delayed, perhaps due to the anomaly itself, then such systems would fail to sound any alarm. MEND's self-disagreement-based technique avoids this pitfall.

Alert Generation via the Relaxed DTW metric. MEND uses a modified form of dynamic time warping (Relaxed-DTW) to assess the level of self-discrepancy to generate alerts (see Figure 2). As the name suggests, the R-DTW algorithm is aimed at reducing false positives by not doing a very strict comparison of two time series being compared for disagreement. This flexibility offered by R-DTW is useful when comparing time series data which may exhibit variations in speed, phase or duration. The choice of R-DTW as a metric is further motivated in Section 4.1.

If the self-disagreement is beyond a certain threshold, an alarm is sounded. This threshold is trained on nominal data.The threshold is set to twice the maximum R-DTW score on a validation dataset that covers a nominal day. To further reduce false-positives, MEND uses a 2D threshold similar to [2]. An alarm is sounded only if the R-DTW-based score exceeds the threshold continuously for 10 timestamps.

input : Two sequences, $s1$ and $s2$
output: Relaxed DTW score, rel_dtw

$dtw\_distance, warp\_path \leftarrow fastdtw(s1, s2)$;
$dtw\_score \leftarrow dtw\_distance$;
$path\_x \leftarrow$ list of x-coordinates in $warp\_path$;
$path\_y \leftarrow$ list of y-coordinates in $warp\_path$;
$matrix \leftarrow$ compute euclidean distance matrix(s1, s2);
$df\_dtw \leftarrow$ create a DataFrame with column 'path_x' and values from $path\_x$;
$df\_dtw['path\_y'] \leftarrow$ values from $path\_y$;
$df\_dtw['distance'] \leftarrow$ flatten the matrix values at indices $[path\_x][path\_y]$;
$rel\_dtw \leftarrow \frac{1}{\text{length of } s1} \times \left(\sum_{\text{group by path\_x}} \min(\text{distance}) + \sum_{\text{group by path\_y}} \min(\text{distance})\right)$;
return rel_dtw;

Figure 2: The Relaxed DTW distance (R-DTW).

We note that MEND's technique of alert generation differs from the alternative of performing multiple forecasts using different horizons. Instead, MEND makes prediction at the same horizon but with a stride. This has the benefit that the quality of all of MEND's multiple forecasts is expected to be the same whereas performing forecasts with differing horizons may affect the quality of the longer term forecasts which may cause false positives.

4 EXPERIMENTAL RESULTS

Extensive experiments were done comparing MEND with several competitor algorithms as well as variants of MEND itself to assess the impact of its design choices. All experiments were carried out on a 64-bit machine with Intel® Core™ i7-8700 CPU @ 3.20GHz, 6 cores, 16 GB RAM and Windows 10 Pro OS. The list of competitors and MEND variants is briefly described below.

(1) MEND: this is the base algorithm with all design choices turned on including augmented features (AF) and multi-forecast (MF).
(2) NBEATSx: this is a state-of-the-art deep neural zero-shot time series forecasting model [8] implemented via the Darts package for univariate time series modelling in the presence of exogenous variables [5].
(3) LSTM: this is an LSTM-based method [6] for which code was available and which surpasses other state-of-the-art techniques such as [3] for anomaly detection experiments (see [3, Table 2]).
(4) Anomaly Prediction: this is a method designed specifically for anomaly detection in spacecraft telemetry data. Although code was not available for this method, the method was implemented afresh using standard components from the sklearn library [10] such as K-means clustering, K-SVD and one-class Support Vector Machine (OCSVM).
(5) Bayesian RR: this is a simple linear regression model implemented in [10] that was adapted to univariate time series prediction with exogenous variables via a wrapper provided by the Darts package [5].

(6) MEND-MF: this was a variant of MEND that received augmented features but was forced to generate anomaly alerts using a single forecast.
(7) MEND-AF: this was a variant of MEND that was allowed to use multiple forecasts to generate anomaly alerts but did not receive augmented features.
(8) MEND-MF-AF: this variant of MEND neither receive augmented features nor was allowed multiple forecasts.

4.1 Nominal Behavior Modeling

To demonstrate the performance of MEND for nominal behavior modeling, it was compared to two competitor models and one MEND variant. Specifically, MEND-AF was implemented by withholding augmented features from MEND and using a forecasting horizon of 100 timestamps (recall that 30 seconds elapse in each timestamp). Bayesian RR was implemented with a forecasting horizon of 100 timestamps to assess the performance of univariate modelling vs multivariate modelling as is done by MEND.

Finally, a state-of-the-art deep time series model that allows exogenous variables, namely NBEATSx was implemented with a forecasting horizon of 20 timestamps. Note that NBEATSx was given a shorter forecasting horizon since it was found to offer poor performance with larger horizons. This is a known limitation of complex non-linear models which inherently limits their ability to make early detection of anomalies. Note that the variants MEND-MF and MEND-MF-AF were not implemented in this study since multiforecasting is most useful for anomaly alert generation and these datasets had no anomalies.

The same training data-set was used for all algorithms. In Table 2, MEND demonstrates superior performance than competitors across various metrics. Notably, R-DTW imposes a more substantial penalty on the worst-performing model (NBEATSx) compared to the average model (MEND-AF), a distinction that is not observed with other metrics. This is understandable for normalized metrics such as MAPE but even unnormalized metrics such as RMSE and MAE do not seem to distinguish between the best and worst performing model as well. This indicates the suitability of R-DTW to scenarios where the temporal sequences being compared may exhibit variations in speed, phase, or duration. R-DTW effectively highlights performance disparities between models.

Table 2 shows the errors obtained by various models on the nominal behavior modelling exercise while Figure 3 depicts the same pictorially. MEND offers by far the most superior modelling behavior. It is curious that removing augmented features leads to a sharp drop in performance. Bayesian RR which is a univariate model (models only battery voltage) has worse performance than both MEND variants which do joint multivariate modelling of battery current, voltage and augmented features. This shows the benefits of multivariate modelling. Finally, NBEATSx struggles to give accurate predictions demonstrating the sufficiency of simple linear models with careful augmentation for this task.

Comparison to LSTM-based model. The model in [6] was set up to accept multivariate input (battery voltage and battery current) with an input sequence length of 100 steps (same as other models) and a forecasting horizon of 20 steps (a fifth of the forecasting horizon of MEND which was 100 steps). Figure 4 shows that MEND offers an RMSE of 0.31 which is significantly smaller than [6] which offered an RMSE of 0.54. It is noticeable that this was despite the LSTM-based method being offered a handicap of a much smaller forecasting horizon whish supports the earlier observation of deep models struggling to maintain accuracy with increasing horizons.

Table 2: A comparison of MEND, its variant MEND-AF and competitors based on various metrics (lower is better for all metrics). MEND offers the best performance across all metrics followed by MEND-AF. The numbers in parentheses indicate the ratio of metric achieved by a particular method to that achieved by MEND i.e., how worse a metric value does that method offer as compared to MEND (these ratios are not reported for MEND itself since they would all be unity). It is notable that MEND-AF offers 20-73% worse performance than MEND whereas other competitors frequently offer performance that is more than 100% worse than MEND.

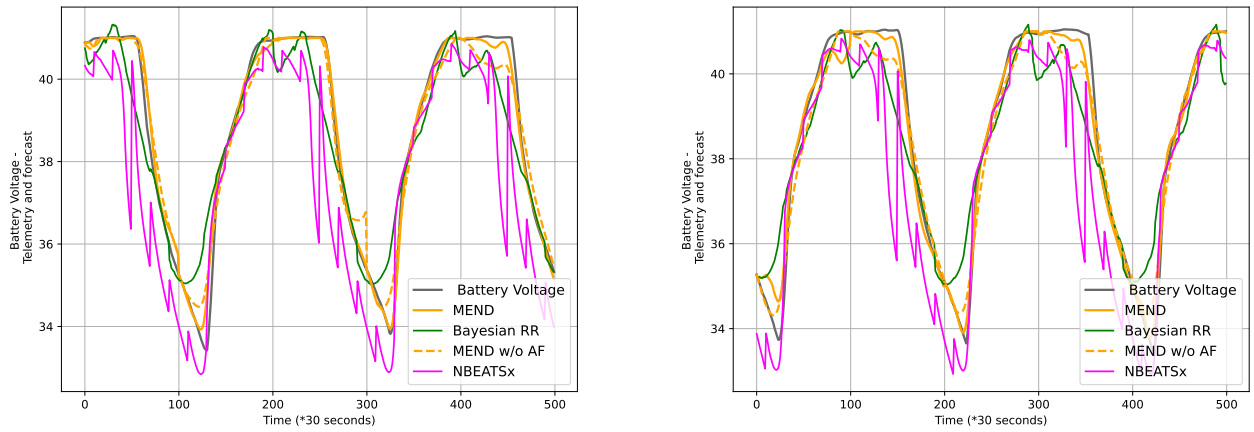
Dataset	Metric	MEND	MEND-AF	Bay_RR	NBEATSx
Nom 1	R-DTW	0.533	0.927 (1.739 ×)	2.247 (4.216 ×)	7.188 (13.486 ×)
	MAE	0.19	0.3 (1.579 ×)	0.6 (3.158 ×)	1.11 (5.842 ×)
	RMSE	0.31	0.46 (1.484 ×)	0.87 (2.806 ×)	1.53 (4.935 ×)
	MAPE	0.52	0.8 (1.538 ×)	1.58 (3.038 ×)	2.89 (5.557 ×)
	sMAPE	0.71	0.89 (1.253 ×)	1.26 (1.775 ×)	1.72 (2.422 ×)
Nom 2	R-DTW	0.67	1.116 (1.666 ×)	2.361 (3.524 ×)	7.452 (11.122 ×)
	MAE	0.17	0.26 (1.529 ×)	0.61 (3.588 ×)	1.09 (6.412 ×)
	RMSE	0.26	0.37 (1.423 ×)	0.86 (3.308 ×)	1.54 (5.923 ×)
	MAPE	0.46	0.68 (1.478 ×)	1.59 (3.456 ×)	2.84 (6.174 ×)
	sMAPE	0.68	0.82 (1.206 ×)	1.26 (1.853 ×)	1.71 (2.515 ×)

4.2 Early Detection of Anomalies

Multi-forecasts help MEND detect early signatures of anomalous behavior. To demonstrate the same, for the two anomalous datasets MEND was compared with two variants, namely MEND-MF that was forced to generate anomaly alerts without using multiple forecasts, MEND-AF that was denied augmented features, and MEND-MF-AF that was denied both. We note that Bayesian RR and NBEATSx were not included for this experiment since they did not give good modelling performance on even nominal data and are expected to perform worse at detecting early signatures of anomaly onset. This is because the threshold for alarm generation would need to be higher for these algorithms given that they incur so much error even on nominal data. With this in mind, only MEND variants were compared for analyzing the impact of multi-forecast in early detection of anomalies.

We also offer comparisons to [4] which is a method designed specifically to detect anomalies in spacecraft telemetry data. Figure 6 presents predictions of this method on the validation dataset. It is notable that the methodology described in [4] gave a very large number of false positives that limits its utility as an anomaly detection technique.

Figure 5 shows plots demonstrating an off-nominal day and how different algorithms are able to generate alerts on the same. MEND is always able to sound an alarm 20-30 timestamps (equivalent to 10-15 minutes) in advance (see Table 3 for exact timestamp data), thus giving the system admins a decent headroom to perform preemptive maintenance or other action to avoid negative impact to



(a) Nominal Day 1 **(b) Nominal Day 2**

Figure 3: Nominal behavior modeling results for MEND vs competitors. MEND offers the most superior nominal modelling behavior given its use of augmented features. Note that MEND-AF struggles to model rapid changes in battery voltages resulting in poor prediction performance.

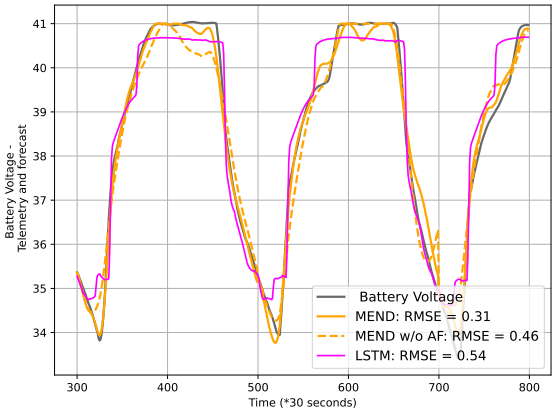


Figure 4: Predictions for battery voltage on nominal day 1. MEND (RMSE = 0.31) and MEND-AF (RMSE = 0.46) perform better than the LSTM-based model by [6] (RMSE = 0.54). Notice that the LSTM-based model is able to predict neither the peaks nor the troughs of the battery voltage correctly.

mission goals. However, all other algorithms fail to generate an advance alarm and are only able to detect the off-nominal behavior once the anomaly has already set in. A point worth noting is that models without multi-forecast must wait for the arrival of off-nominal spacecraft telemetry before they can detect an alarm which inherently limits their ability to make predictions far in advance. This problem is exacerbated due to the fact that spacecraft telemetry can itself experience delays in getting received and processed at ground stations further delaying the anomaly detection process.

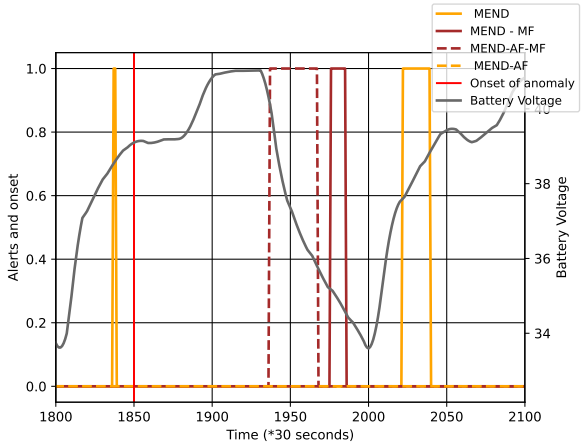
Data	Onset	MEND	MEND-MF	MEND-AF	MEND-AF-MF
Off-nominal Day 1	1850	1820	1970	—	1935
Off-nominal Day 2	90	70	165	—	165

Table 3: Results from early anomaly detection experiments. The second column lists the timestamp at which the onset of anomaly was determined while the subsequent columns report the earliest timestamp at which a method could raise an alarm or alert. It is notable that MEND was able to sound an alert 20-30 timestamps in advance which corresponds to 10-15 minutes of advance notice given that 30 seconds are elapsed in each timestamp. In contrast, all other methods generate their alert well-into the anomalous behavior period.

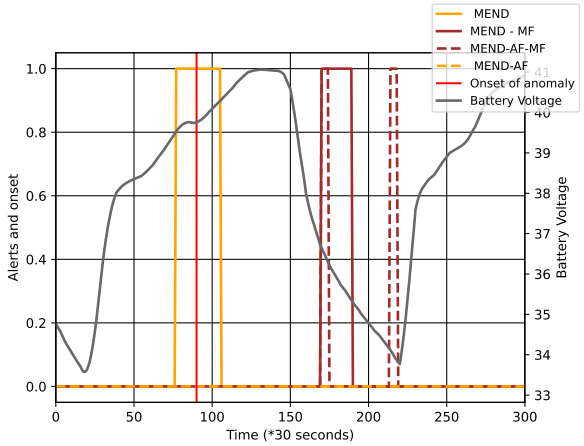
4.3 Insights into anomalous behavior

MEND also generates insights in addition to alarms for anomalous behaviour detection. This is made possible by the model's incorporation of domain knowledge using augmented features. The subsystem health is reflected in the form of interpretable meta-data. The meta-data for the power system health monitoring case study that is described here is linked to the periodicity of various operating modes in the subsystem.

Figure 7 demonstrates this using pair plots. Given a time slice, the augmented feature for that slice reveals the most likely mode of operation during that slice by simply choosing the mode/cluster to which the DTW distance is the closest (see Section 3). Uninterrupted runs of time slices all sharing the same mode of operation are identified and for each such run, meta-parameters viz. average battery current, average battery voltage and the duration of this run are stored as meta-data. Each of the pair-plots in figure 7 is a visual representation of these meta-parameters.



(a) Off-nominal Day 1



(b) Off-nominal Day 2

Figure 5: Comparing MEND variants for early forecasts. It is notable that MEND is always able to generate an alert much ahead of the onset of the anomaly whereas the competitors are either not able to generate an alert in a reasonable amount of time at all (e.g. MEND-AF) or else are able to generate an alert but well-into the anomalous period, giving no advance notice to the system admin to take preemptive action.

Each pair plot presents a 3×3 grid of 9 pair-plots. The diagonal plots in the grid (i.e. top-left, middle-middle and bottom-right) represent the distribution of a specific meta-parameter. The off-diagonal plots show scatter-plots demonstrating correlation any two meta-parameters. Blue color represents slow-charge mode, orange represents fast-charge mode and green represents discharge mode. For instance, the bottom right plot shows the distribution of the length of runs of the 3 modes. Note that on nominal days Figure 7(a,c), all 3 modes last roughly for the same duration which is as expected. However, for off-nominal days Figure 7(b,d), observe the bottom right plots and notice that the slow-charge and fast charge runs are much shorter and the discharge runs are much longer, indicative of a progressing anomaly. Similarly, a look at

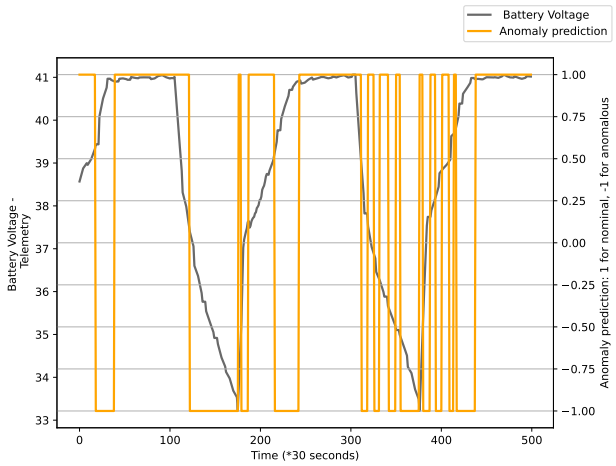


Figure 6: Anomaly predictions made by [4] on the validation dataset. +1 indicates a nominal prediction while -1 indicates an anomaly alert by the method. The abundance of false positives by this method limits its utility.

the scatter plots (the "off-diagonal" plots) reveals that on nominal days, the 3 modes of operation have distinct separation and tight concentration but on off-nominal days, there is much more spread and the scatter plots of different modes seem to be merging into each other. These observations indicate the the augmented features prepared by MEND are very helpful in identifying anomalies.

Comparison between the pair-plots of nominal and anomalous days leads to the following insights since the onset of anomaly:

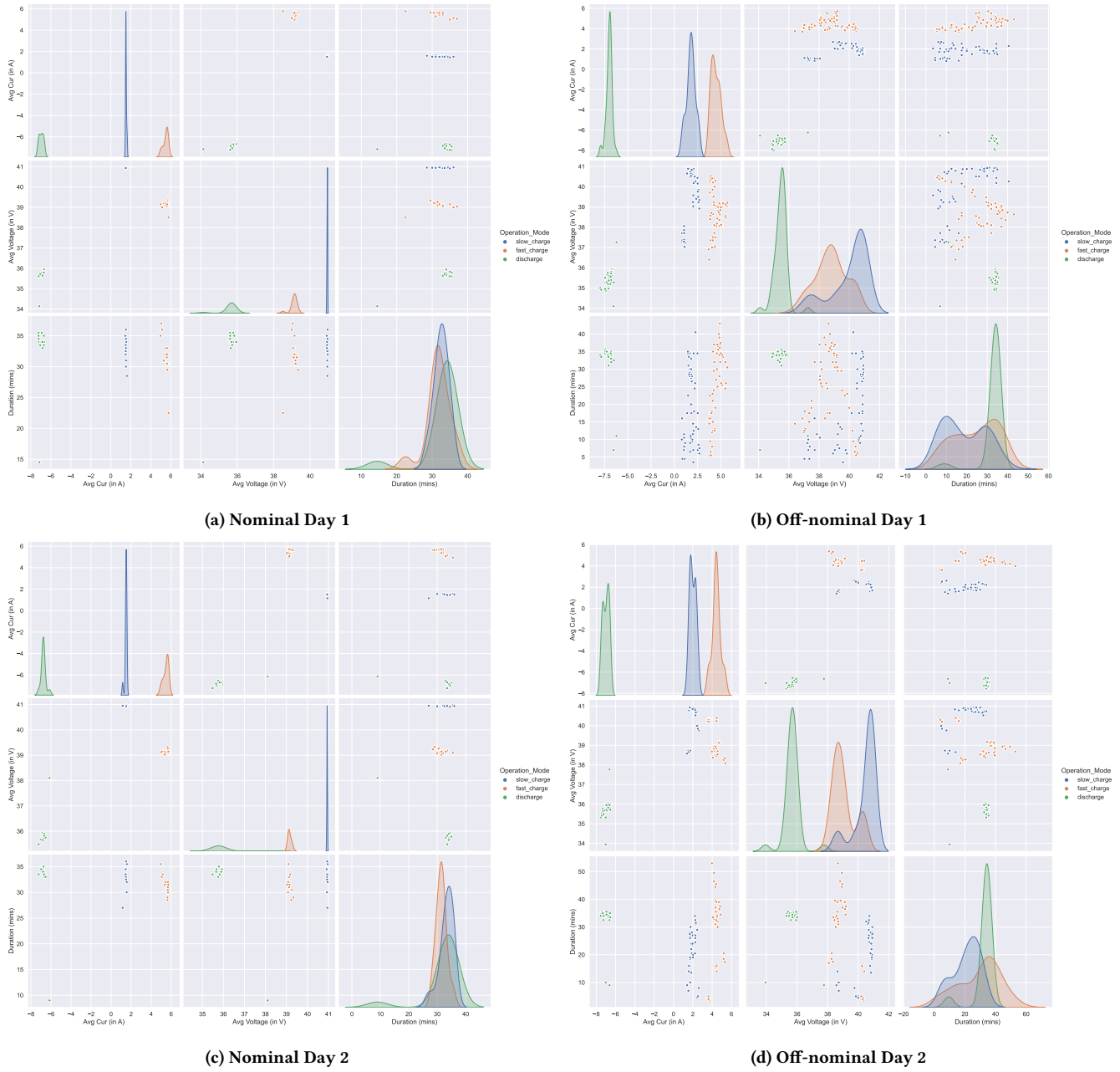
(1) The duration of slow charge mode of operation is reduced.
(2) Fast charge mode of operation has a much higher variance in the battery voltage.
(3) Discharge mode remains unaffected during off-nominal behavior in the subsystem.

Such insights provide valuable information that can help in isolation of the fault that occurred in the subsystem and allow for speedy root-cause analysis.

5 CONCLUSION

This paper developed the MEND method for early detection of anomalies by making careful use of simple models by augmenting them using additional features, multivariate modelling and choice of a novel relaxed time-series metric that helps avoid false positives. MEND made use of multi-forecasting to generate alerts far ahead of the onset of the anomaly. The process of generating augmented features revealed critical aspects of the functioning of the system and explainable insights into faults (see Figure 7).

The study presents certain key takeaways. Multivariate modelling, although more challenging, can offer superior performance than univariate modelling. Overtly complex non-linear models can be suboptimal for tasks with scant training data and supervision. Offering carefully curated features, such as the augmented features in this work, can allow even simple models to perform excellently. Multi-forecasts seem essential to be able to make reliable and early



(a) Nominal Day 1

(b) Off-nominal Day 1

(c) Nominal Day 2

(d) Off-nominal Day 2

Figure 7: Examining the impact of augmented features on anomaly modelling. The pair-plots were obtained from the subsystem on a nominal day (a, c) and off-nominal day (b, d). The discharge mode of operation is unaffected whereas the slow charge mode has lower time-length indicating that it has been affected by the onset of off-nominal behavior.

detection of anomalies. It is important to avoid excessive reliance on real-time telemetry which may experience delays and hinder the anomaly discovery process. Several avenues of future work exist including experimenting with other spacecraft subsystems, further automating the process of augmented feature discovery.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for helpful comments. P.K. thanks Microsoft Research India and Tower Research for research grants.

REFERENCES

[1] 2007. Dynamic Time Warping. In *Information Retrieval for Music and Motion*. Springer Berlin Heidelberg, 69–84. https://doi.org/10.1007/978-3-540-74048-3_4

[2] Wissam Aoudi and Magnus Almgren. 2021. A Framework for Determining Robust Context-Aware Attack-Detection Thresholds for Cyber-Physical Systems. In *Proceedings of the 2021 Australasian Computer Science Week Multiconference* (Dunedin, New Zealand) *(ACSW '21)*. Association for Computing Machinery, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.1145/3437378.3437393

[3] Sriram Baireddy, Sundip R. Desai, James L. Mathieson, Richard H. Foster, Moses W. Chan, Mary L. Comer, and Edward J. Delp. 2021. Spacecraft Time-Series Anomaly Detection Using Transfer Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1951–1960. https://doi.org/10.1109/CVPRW53098.2021.00223

[4] Jiahui He, Zhijun Cheng, and Bo Guo. 2022. Anomaly Detection in Satellite Telemetry Data Using a Sparse Feature-Based Method. *Sensors* 22, 17 (2022). https://doi.org/10.3390/s22176358

[5] Julien Herzen, Francesco LÃ¤ssig, Samuele Giuliano Piazzetta, Thomas Neuer, LÃ©o Tafti, Guillaume Raille, Tomas Van Pottelbergh, Marek Pasieka, Andrzej Skrodzki, Nicolas Huguenin, Maxime Dumonal, Jan KoÅ›cisz, Dennis Bader, FrÃ©dÃ©rick Gusset, Mounir Benheddi, Camila Williamson, Michal Kosinski, Matej Petrik, and GaÃ«l Grosch. 2022. Darts: User-Friendly Modern Machine Learning for Time Series. *Journal of Machine Learning Research* 23, 124 (2022), 1–6. http://jmlr.org/papers/v23/21-1177.html

[6] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 387–395. https://doi.org/10.1145/3219819.3219845

[7] Andrew V. Metcalfe and Paul S.P. Cowpertwait. 2009. *Introductory Time Series with R*. Springer, New York, NY.

[8] Kin G. Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. 2023. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. *International Journal of Forecasting* 39, 2 (2023), 884–900. https://doi.org/10.1016/j.ijforecast.2022.03.001

[9] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2020. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=r1ecqn4YwB

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[11] Stan Salvador and Philip Chan. 2007. Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intell. Data Anal.* 11, 5 (oct 2007), 561–580.

[12] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*. 166–171. https://doi.org/10.1109/ICDCSW.2011.20

[13] Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

[14] Jr Smith, A A. 1993. Estimating Nonlinear Time-Series Models Using Simulated Vector Autoregressions. *Journal of Applied Econometrics* 8, S (Suppl. De 1993), 63–84. https://ideas.repec.org/a/jae/japmet/v8y1993isps63-84.html

[15] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. 2020. Tslearn, A Machine Learning Toolkit for Time Series Data. *Journal of Machine Learning Research* 21, 118 (2020), 1–6. http://jmlr.org/papers/v21/20-091.html