AAS 22-666

AUTONOMOUS ANOMALY DETECTION VIA UNSUPERVISED MACHINE LEARNING

Felipe Giraldo-Grueso[*], Renato Zanetti[†] and Michelle R. Simon[‡]

Manual inspection of satellite telemetry data in search of anomalies is a time-consuming detection technique. Most multi-signal systems send back extensive data that a single person cannot easily monitor. Machine learning techniques that scan the data and flag anomalies autonomously are an attractive alternative. The autonomous anomaly detection problem can be divided into two sub-problems: regression analysis and a classification process. In the regression analysis, a machine learning model is trained to reconstruct a given signal, and the classification process categorizes the reconstruction error as anomalous or nominal. This paper studies the autonomous anomaly detection problem and proposes improvements to the regression and classification sub-problems. Regarding the regression analysis, it was found that including the physics of the target signal in the machine learning model yields a lower reconstruction error than a purely data-driven model. The classification approaches studied showed that cluster-based thresholding techniques accompanied by a pruning procedure can outperform non-parametric dynamic thresholds.

INTRODUCTION

Anomaly detection is a crucial process in multi-signal systems. Complex systems such as spacecraft have multiple channels with individual telemetry signals that operators cannot monitor in real-time to identify anomalies and prevent faults and failures. As most systems have many channels that need constant observation, manual monitoring is non-viable, and fluctuations in the data that could lead to larger anomalies often remain undetected. Alternatively, artificial intelligence approaches can be used as autonomous anomaly detection models, specifically those relying on machine learning. Multiple machine learning models have already been used, monitoring CT scans, prices, stocks, and telemetry data.[1] In recent years, machine learning for anomaly detection has proven to be useful and has been an active research field.

Traditional anomaly detection in satellites relies on manual inspection by a qualified person with sufficient knowledge and alarms to signal anomalous behavior when a value strays out of a predetermined limit.[2] Satellites typically generate large amounts of telemetry data which are commonly

[*]Ph.D. Student, Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, Austin, TX 78712.

[†]Assistant Professor, Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, Austin, TX 78712.

[‡]Technical lead for Autonomous Operations Resilience for Tactile Action (AORTA), Airforce Research Laboratory, Kirtland Air Force Base, Albuquerque, NM.

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

multivariate time-series, making traditional anomaly detection inefficient.[2] Recently, artificial neural networks have been used to replace traditional anomaly detection in satellite telemetry data. More specifically, in 2018, a deep learning model based on Long Short-Term Memory (LSTM) layers was implemented to identify anomalies in labeled telemetry data taken from the *Mars Science Laboratory*, and *Soil Moisture Active Passive* satellites.[3] In 2019, Jin et al.[2] trained a denoising autoencoder to detect anomalies in a satellite power subsystem.

When nominal data is available, the autonomous anomaly detection problem can be solved by following two steps.[1] First, a machine learning model (usually an autoencoder or a multi-layer neural network) is trained to reconstruct or predict the nominal data. This first step can be considered a regression problem. Second, the reconstruction error is used to classify each point as either an anomaly or nominal behavior.[4] For the classification process, nearest neighbors and clustering techniques have been used to set automatic anomaly thresholds.[2] An efficient anomaly detection algorithm has to reconstruct the nominal behavior as closely as possible so that the actual anomalies in the data return an unusually high reconstruction error, which the classification algorithms can easily detect.

Most signal reconstruction algorithms are trained to minimize the reconstruction error without considering physics constraints in the training data. Physics Informed Neural Networks (PINNS) were first introduced in 2019 to allow the reigning physics of the problem to be included.[5] These are specific neural networks trained to solve supervised learning tasks while considering any given physics law described by general partial differential equations. If the physics governing the studied phenomenon is known, it can be introduced into the model to achieve better results potentially. Instead of just minimizing the difference between the real values and the values predicted by the network, PINNS minimize the difference between real and predicted values while conforming to the physics present in the problem as soft constraints.[5]

This paper studies the autonomous anomaly detection problem and proposes improvements to the regression and classification sub-problems. For the regression portion, we study how including physics in machine learning models used for signal reconstruction can benefit the signal reconstruction process. Different methods are studied for the classification portion to determine an optimal anomaly threshold efficiently.

RELATED WORK

The regression sub-problem in autonomous anomaly detection has been accomplished as a purely data-driven approach in previous work when applied to satellite data. The results published by Hundman et al.[3] and Jin et al.[2] carry out the regression portion by training a machine learning model based on actual telemetry data from the satellites being studied. As mentioned before, Hundman et al.[3] trained a model based on Long Short-Term Memory (LSTM) layers to identify anomalies in real labeled telemetry data taken from the *Mars Science Laboratory* and *Soil Moisture Active Passive* satellites. In the results presented by Jin et al.,[2] a denoising autoencoder was trained to detect anomalies in a satellite power subsystem.

The classification portion of this work builds on the results presented by Hundman et al.,[3] as their

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

labeled data is readily available for public use. Their classification work relies on non-parametric dynamic thresholding techniques to identify anomalies. In other words, their anomaly threshold is found that, if all values above are removed, it causes the most significant percent decrease in the mean and standard deviation of the reconstruction errors.[3]

DATA PREPARATION

Two different data sets were used, each specifically tailored to demonstrate the proposed contributions in the regression and classification portions.

Regression

The data used for the signal reconstruction was taken from the Advanced Composition Explorer launched on August 25, 1997,[6] which will be referred to as the ACE dataset. More specifically, four-minute average magnetic field measurements taken by the onboard dual triaxial magnetometer with respect to the spacecraft's geocentric solar ecliptic (GSE) coordinates were used. This dataset also contains position measurements in GSE coordinates. The data was filtered using an exponential moving average with a 500-point window and shifted so that all values are positive. Having the position measurements helps understand some of the physics that drives the magnetic field. Following Maxwell's laws and neglecting any relativity contributions:[7]

$$\nabla \cdot \mathbf{B} = 0 \tag{1}$$

$$\nabla \cdot \mathbf{B} = \frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} = 0 \tag{2}$$

That is, the divergence of the magnetic field must be zero for all time. This is known as Gauss's Law of magnetism, expressed in differential form, which denies the existence of magnetic monopoles.[8] The divergence of the magnetic field can be calculated numerically with the experimental data using numerical differentiation, which can be helpful to check whether the data follows Equation (2). In Figure 1 the filtered magnetic field measurements can be seen, and Figure 2 shows the divergence of the magnetic field measurements calculated by following Equation (2) and using numerical differentiation. Despite noise (which is most likely introduced when performing the numerical differentiation[9]), the divergence of the magnetic field approaches zero for all time steps, thus confirming the physics behind the magnetic field measurements. Therefore, the use of this dataset aims to explore how including physics in the loss function of a regression model can help improve the reconstruction of the signal.

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

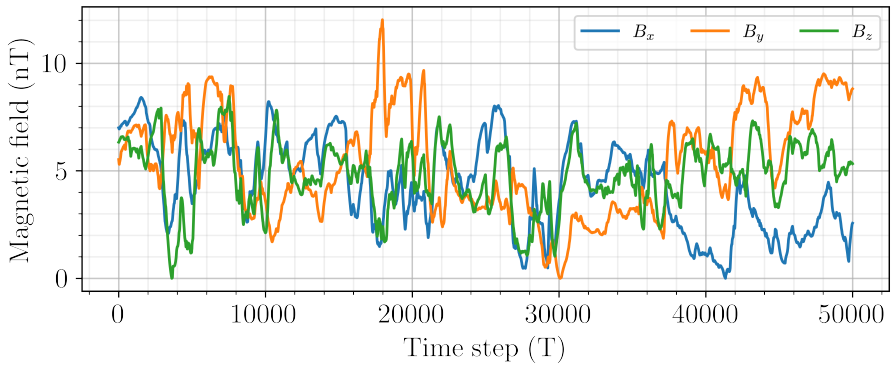

Figure 1. Magnetic field measurements filtered from the ACE dataset.

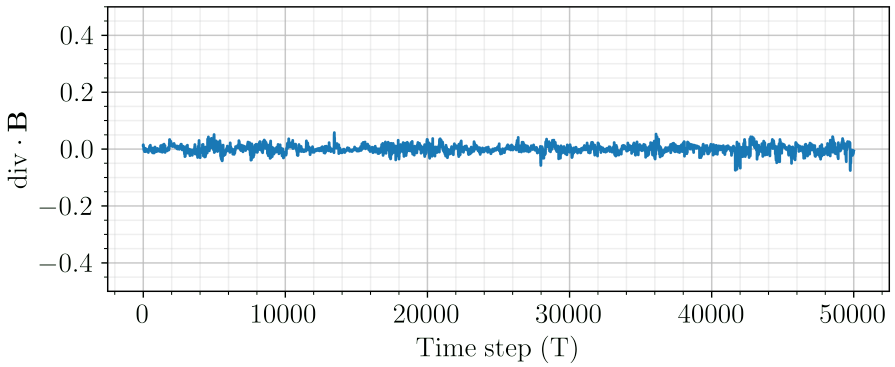

Figure 2. Divergence of magnetic field measurements from the ACE dataset.

Classification

The data used for the classification problem was taken from the *Soil Moisture Active Passive*, and the *Mars Science Laboratory*,[3] hereafter referred to as the SMAP/MSL dataset. For each of these satellites, a different number of telemetry channels are included and contain a time series with one-hot encoded information for commands sent to each satellite module and the actual telemetry value of the channel. Since the data has already been scaled and preprocessed, no data filtering nor scaling was necessary. For this dataset, the training samples include only nominal data, while the test samples have labeled anomalies which are useful for evaluating classification performance. In Figure 3, the training and testing time series for channel P-11 can be seen. It is important to notice that the test dataset is labeled with the exact time step at which an anomaly starts and ends. Therefore, the purpose of this dataset is to explore different thresholding techniques that can improve classification performance.

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

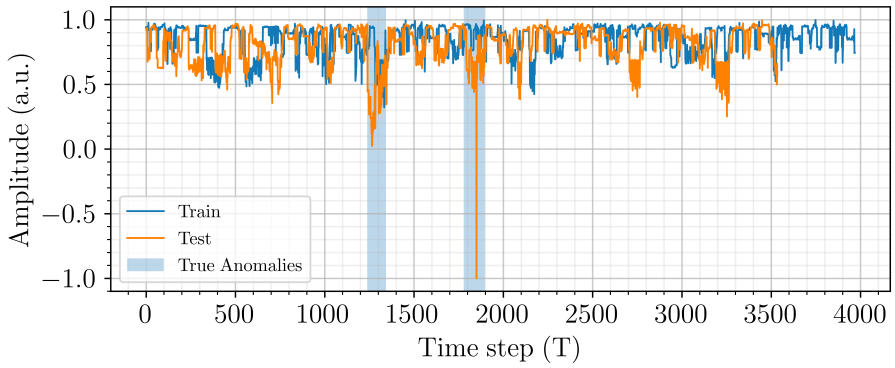

Figure 3. Training and testing time series for channel P-11 of the SMAP/MSL dataset.

LEARNING FRAMEWORK

ACE dataset

In regression problems using machine learning, the model predicts a numerical value given a known input. The model is asked to predict or approximate a function f that maps the input to the desired output.[10] In this sense, signal reconstruction can be considered a regression problem. Several models can be used in the deep learning framework to solve this type of problem. A standard model for signal reconstruction problems is the autoencoder.[1]

Autoencoders An autoencoder is a neural network trained to reconstruct its input and display it as an output (In some cases, the model is trained to copy the input to its output). Its architecture consists of two basic blocks: the encoder and the decoder. The encoder reduces the dimensionality of the input and maps it to the latent space, and the decoder handles the reconstruction of the signal from the latent space.[10] In the specific context of anomaly detection, the autoencoder is trained to be able to reconstruct nominal signal data, such that if there is an anomaly in the signal, the autoencoder will not be able to reconstruct the anomaly.

Loss functions Different training loss functions can be used to train an autoencoder to reconstruct a signal. Most commonly, the loss functions used for signal reconstruction or enhancement are distance metrics between the reconstructed signal and the desired target.[11] With this in mind, a standard metric is the mean squared error:

$$\mathcal{L}_{\text{mse}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

Where N is the number of points, y refers to the desired target, and $\hat{y}$ is the predicted output. The use of these loss functions shows that traditional autoencoders are purely data driven. If the physical properties of the signal that is being reconstructed are known, they can be implemented in the loss function as soft constraints to restrict the output.[5] These soft physical constraints can give the model more information to predict the correct output. With this, a new loss function can be used as follows:

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

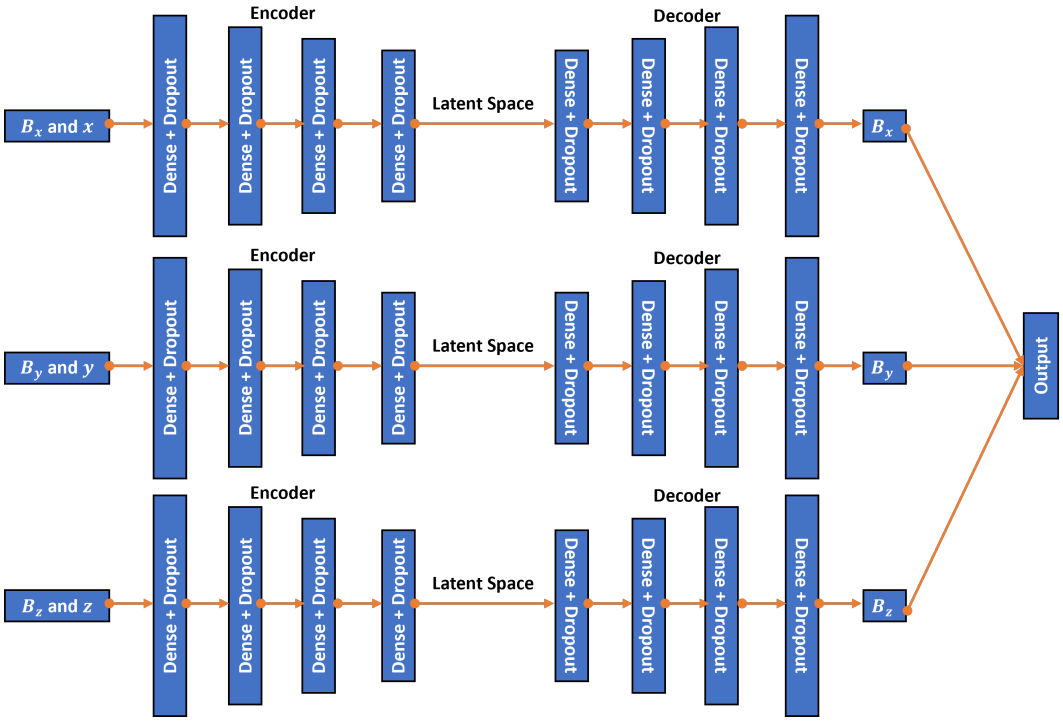

Figure 4. Model used to reconstruct the magnetic field signal in the ACE dataset.

$$\mathcal{L}_{\text{pi}} = \lambda_1 \mathcal{L}_{\text{mse}} + \lambda_2 f(\boldsymbol{x}, \hat{\boldsymbol{y}}) \tag{4}$$

Where f refers to the additional knowledge on the problem which might be a function of both the input and the approximated output of the model, and λ_1, λ_2 are simply scaling factors.

The model used to reconstruct the magnetic field signal in the ACE dataset can be seen in Figure 4. This model consists of three different autoencoders whose task is to reconstruct the signal for each magnetic field component by inputting the position and the magnetic field component itself. The output of each autoencoder is concatenated to use a single loss function that includes all three reconstructed signals. Each dense layer contains a dropout of 0.2 to avoid overfitting and a hyperbolic tangent activation function, except the last one which has a ReLU activation function. The loss function used to train the model can be seen below:

$$\mathcal{L}_{\text{pi}} = \frac{\lambda_1}{N} \sum_{i=1}^{N} \left(\mathbf{B}_i - \hat{\mathbf{B}}_i \right)^2 + \frac{\lambda_2}{N} \sum_{i=1}^{N} \left(\frac{\partial \hat{B}_{x_i}}{\partial x_i} + \frac{\partial \hat{B}_{y_i}}{\partial y_i} + \frac{\partial \hat{B}_{z_i}}{\partial z_i} \right)^2 \tag{5}$$

Where $\boldsymbol{B} = [B_x, B_y, B_z]^T$ refers to the true magnetic field and $\hat{\boldsymbol{B}} = \left[\hat{B}_x, \hat{B}_y, \hat{B}_z\right]^T$ is the reconstructed magnetic field. The first term is the data mean squared error reduced over the number of components and the second term refers to the $\mathcal{L}_2$ physics regularization. Having the position as part of the input to the model allows the calculation of the predicted magnetic field divergence

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

through automatic differentiation. Automatic differentiation permits the model to compute these derivatives more efficiently in terms of accuracy and computational time compared to numerical differentiation.[12]

Bias-Variance Trade-off In statistical learning, the mean squared error of a prediction is defined as:[13]

$$\mathrm{MSE}\left(\hat{y}\right) = \mathrm{E}\left[\left(y - \hat{y}\right)^2\right] = \mathrm{Var}\left(\hat{y}\right) + \mathrm{Bias}^2\left(\hat{y}\right) + \text{Irreducible Error} \tag{6}$$

As a general rule, high complexity models can be prone to overfitting the training data, increasing the prediction's variance while decreasing its bias. As seen in equation 6, the MSE depends on the relative rate of change of these two quantities. Thus, as a model's complexity increases, the prediction bias tends to decrease faster than the variance rises, making the MSE decrease. However, at a specific point, increasing the model's complexity has little impact on the bias but significantly increases the variance, increasing the MSE. To restrain the complexity of a model, regularization techniques are used, which prevent the model from becoming too complex.[13] Suppose a regularization term is introduced in the model's loss function, as seen in equation 5. In that case, the MSE will generally follow the behavior described in figure 5. This is known as the bias-variance trade-off.

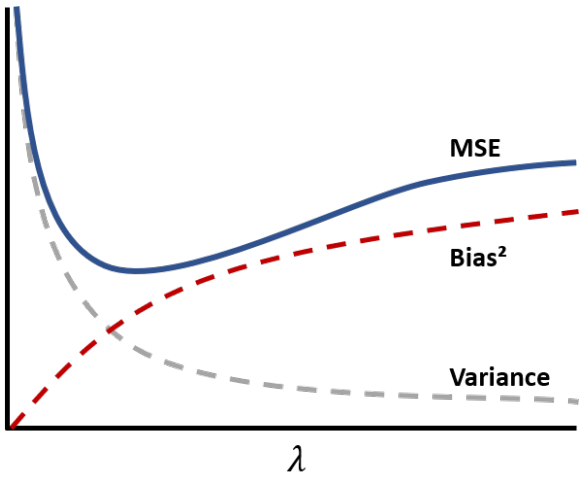

Figure 5. Bias-variance trade-off as a function of the weight of the regularization. As λ increases, the model decreases its complexity. Figure adapted from James et al.[13]

It is important to note that current research has shown that it is possible for the variance to increase as the width of a neural network decreases. Meaning that while the bias-variance trade-off is a good analysis tool, it might not be universal.[14]

Curriculum Regularization The term curriculum regularization was first introduced in 2021 by Krishnapriyan et al.[15] when characterizing the possible failure modes in physics informed neural networks. With this technique, the loss term in the neural network starts from a simple partial differential equation regularization and becomes more complex as the training advances. Curriculum

regularization allows a warm start to the neural network training by finding a proper set of initial weights.[15] Based on this idea, this paper implements a similar technique where the importance of the physics term in the loss function starts to increase gradually as the model is being trained.

SMAP/MSL dataset

Once the signal reconstruction model has been implemented, identifying anomalies becomes a classification problem. The classification problem is responsible for classifying the reconstruction error into nominal behavior and potential anomalies. Therefore, to obtain the reconstruction error for this dataset, a model based on long short-term memory networks was used.

Long Short Term Memory Networks Recurrent networks use Long Short-Term Memory layers that can use feedback connections in order to take into account past representations of recent inputs in the form of activation neurons.[16]

In Figure 6, the sequential network architecture used for the SMAP/MSL dataset can be seen. This is the same architecture presented by Hundman et al.,[3] which has proven to be sufficiently good at reconstructing the signal. In this case, the input is not the entire time series. Instead, the input is a time window of the time series (specifically 250 time steps), and the output is the prediction of the following 10 time steps.[3] The output can then be compared to the actual telemetry data by averaging the prediction, which essentially works as a moving average filter. Since the dataset includes contextual anomalies rather than just point anomalies; the LSTM layers help identify these anomalies.[3] As the physics of this dataset is not known, the mean squared error loss was used to train a different model, following the same architecture as shown in Figure 6, for each channel. It is important to note that the purpose of this dataset is to classify the error between the reconstructed output and the ground truth. Considering that the regression model returns the error between the reconstructed values and the real signal, an error threshold must be set to know when an anomaly has been detected. This threshold establishes the limit between the nominal behavior and the anomalies in the signal. Different techniques can be used to establish the anomaly threshold. In this work, two different alternatives have been studied.

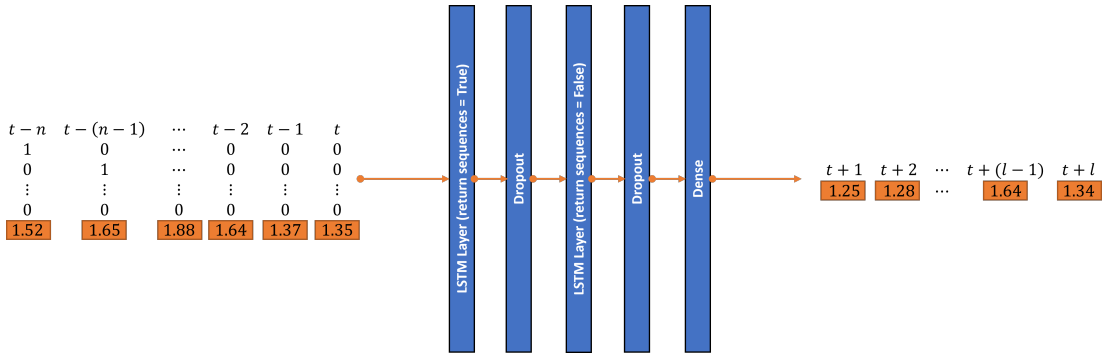

Figure 6. Sequential network architecture used for the SMAP/MSL dataset. The orange boxes represent the telemetry value at each time step. Figure adapted from Hundman et al.[3]

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

Gaussian Assumption The Gaussian assumption method is based on the premise that the error between the predicted values and the ground truth follows a Gaussian distribution. Therefore, the anomaly threshold (ξ) is set as:[17]

$$\xi = \mu(\text{test MAE}) + \alpha\sigma(\text{test MAE}) \tag{7}$$

Where μ is the mean, σ is the standard deviation, α is a scaling parameter, and MAE refers to mean absolute error. If the testing MAE is higher than the threshold at a given point, the model can flag this point as an anomaly. A high α will produce true positives but might return an increased number of false negatives, while a low α will result in an increased number of true positives and a large amount of false positives. Usually, the Gaussian assumption will not hold in practice,[18] so other models can be used to find the necessary threshold.

K-means clustering A different method to find the anomaly threshold is the K-means clustering algorithm. K-means clustering is an algorithm that divides a dataset into K distinct, non-overlapping clusters.[13] In this specific case, the reconstruction error is clustered into two groups representing nominal behavior and anomalies. To understand how K-means clustering works, let C_n and C_a be the clusters representing nominal behavior and anomalies respectively. Now let $\sigma(C_n), \sigma(C_a)$ denote the standard deviation of the reconstruction errors inside each cluster. Therefore, the best possible cluster configuration will be the one that minimizes:

$$W(C_a, C_n) = \sigma(C_a) + \sigma(C_n) \tag{8}$$

The anomaly threshold (ξ) can be set as the lowest reconstruction error in the anomalies cluster.[2]

Pruning procedure Once the anomaly threshold has been set, a pruning procedure in the predicted anomalies must be done. If the anomaly threshold is set too low, in addition to predicting real anomalies, it is possible to set a high rate of false positives. Therefore, the pruning procedure takes care of dealing with false positives. In this work, the pruning procedure identifies all the anomalies predicted, calculates the difference (ϕ) between their reconstruction error and the anomaly threshold, and calculates the ratio between $\phi_{\max}$ and the rest. Then, a pruning threshold (γ) can be set to prune the anomalies if the above mentioned ratio is below the pruning threshold.

In Figure 7, the pruning procedure used can be seen. The predicted anomalies include four false positives since the error signal goes above the threshold five times. The pruning procedure eliminates the false positives and gives only the true positive. It is essential to mention that the pruning procedure will not eliminate all the false positives every time. Regardless, different techniques such as grid search, Bayesian search, coarse to fine random search can be used to find a pruning threshold that maximizes a desired metric. Finding a pruning threshold by maximizing a metric is only valid if the dataset is labeled so quantities such as true positives, false negatives, and false positives can be calculated.

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

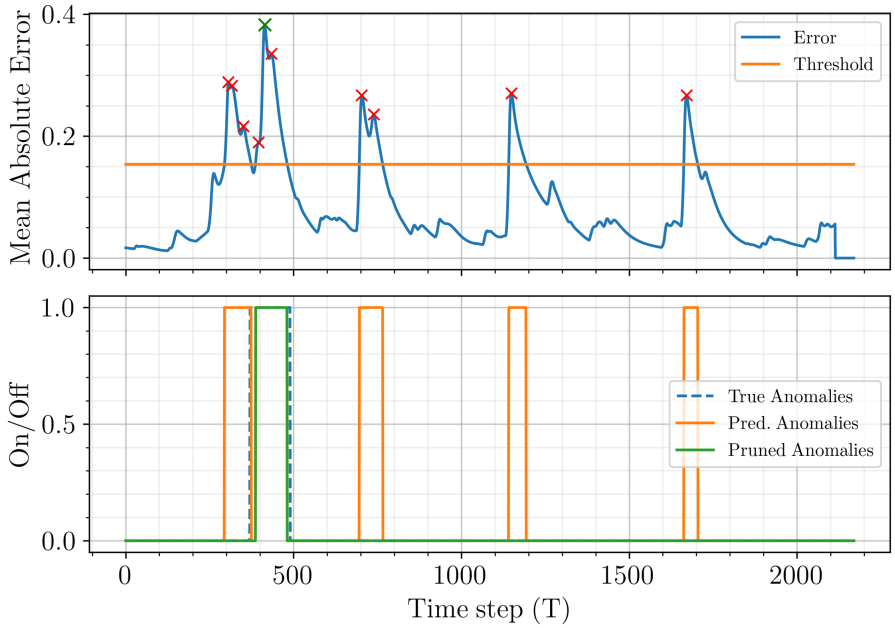

Figure 7. Reconstruction error (Top) between the real and predicted telemetry value for channel T-12. Comparison (Bottom) between true, predicted, and pruned anomalies for T-12 in the SMAP/MSL dataset.

Training and testing

To train the ACE model (Figure 4), the dataset was divided into a training set containing 90,000 data points and a test set with 30,000 data points. The Adamax optimizer[19] with an initial learning rate of 0.01 and a decay rate of 0.0001 was used, and the model was trained for 100 epochs with a batch size of 500, a validation split of 0.2, and an early stop callback monitoring validation loss with a patience of ten epochs.

The SMAP/MSL model (Figure 6) was trained independently for each channel in each dataset. The Adamax optimizer[19] with an initial learning rate of 0.001 was used, and the model was trained for 35 epochs with a batch size of 30, a validation split of 0.1, and an early stop callback monitoring validation loss to avoid overfitting.

Metrics To test the ACE model's ability to reconstruct the signal, the standard MSE metric described in Equation (3) was used. With this, the distance between the reconstructed and real signal on the test set was calculated for different model configurations.

To test different thresholding techniques, the following metrics were used:[13]

$$\text{Precision } (P) = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall } (R) = \frac{TP}{TP + FN} \tag{10}$$

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R} \tag{11}$$

Where TP refers to true positives, FP to false positives, FN to false negatives, and β is a scaling factor. A lower β value gives more weight to precision and less to recall, whereas a larger β, gives less weight to precision and more weight to recall. Then, with these three metrics, the model's performance can be obtained, which can be helpful for fine-tuning and threshold searches. Having introduced these concepts, it is essential to note that a pruning procedure can only improve the precision of the model, and in some cases, it might decrease recall. For the SMAP/MSL dataset, the true positives, false positives, and false negatives were recorded as follows:[3]

- A true positive is recorded if any portion of the predicted anomalies overlaps any true anomaly. Only one true positive is recorded if multiple portions of the predicted anomalies fall within a true anomaly.
- All predicted anomalies that do not overlap with any true anomaly are recorded as false positives.
- If no portion of the predicted anomalies overlaps a true anomaly, a false negative is recorded.

RESULTS

Regression

Physics Regularization For the signal reconstruction of the magnetic field, the ACE model was first trained by varying the loss function scaling factor λ_2 and keeping $\lambda_1 = 1$ constant. A low value of λ_2 refers to a data driven model, while a high value of λ_2 describes a model with a high physics regularization. Figure 8 shows the average mean squared error between the magnetic field from the test set and the reconstructed signal as a function of the parameter $\lambda_2 \in [0, 10]$. The model was trained ten different times, starting from different initial weights. As it can be seen, there is a tendency for the error to decrease as the weight of physics in the loss function increases. This demonstrates that the model benefits from the constraints induced by including a physics-based regularization term in the loss function. The lowest error is found at $\lambda_2 = 2.0$. As with standard regularization methods,[13] the mean squared error reaches a point where it stops decreasing (in this case, at $\lambda_2 = 2.0$) and adopts an increasing trend as λ grows from this point on. Along with reducing the test MSE, including a physics-based regularization term in the training loss function reduces the variance of the results obtained. As the test MSE is closely related to the variance plus the bias squared,[13] it can be seen from Figure 8 that increasing the physics weight lowers the variance, thus lowering the overall test MSE.

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

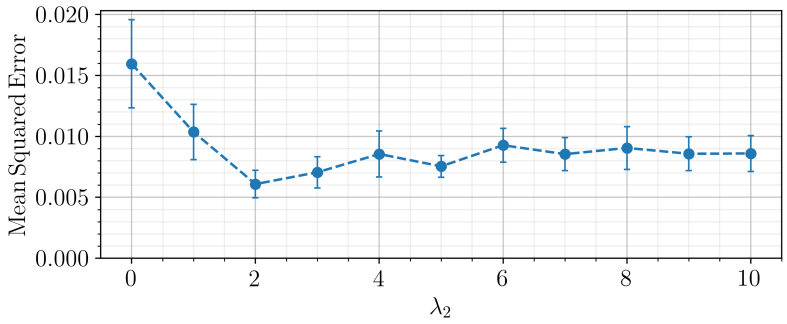

Figure 8. Average MSE between the reconstructed magnetic field and the original signal from the ACE test set as a function of $\boldsymbol{\lambda_2} \in [0, 10]$.

Figure 9 shows the average mean squared error between the magnetic field from the test set and the reconstructed signal as a function of $\lambda_2 \in [10^1, 10^{10}]$. With this figure, it is easier to see the increasing trend that the test MSE adopts as λ_2 grows beyond its optimal point. The increase found for large values of λ shows that a purely physics-based model will not perform as well compared to a mixed model. It is important to note that there is an increase in variance for very high values of λ_2 which could be due to the fact that the physics used in the regularization term are not enough to reconstruct the magnetic field.

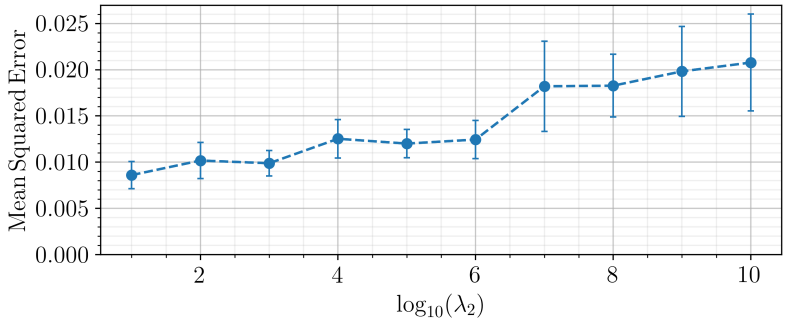

Figure 9. Average MSE between the reconstructed magnetic field and the original signal from the ACE test set as a function of $\boldsymbol{\lambda_2} \in [10^1, 10^{10}]$

Figure 10 shows the average total training epochs as a function of λ_2. Since the training scheme included an early stop callback monitoring validation loss with a patience of ten epochs, a higher number of total training epochs means that the model avoids overfitting the data at the early stages of training. For $\lambda_2 = 0$, which means that the model is purely data-driven, the training is stopped, on average, before it reaches 30 epochs as the model starts overfitting the data at around this point. As the weight of the physics-based regularization term increases, it can be seen from the figure that the average total training epochs adopts an increasing tendency. Thus, including the physics in the loss function helps the model avoid overfitting the training data as expected. It is important to note that training for more epochs will not guarantee a lower test MSE, as a high regularization might lead to an under-fitted model due to the bias-variance trade-off.[13]

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

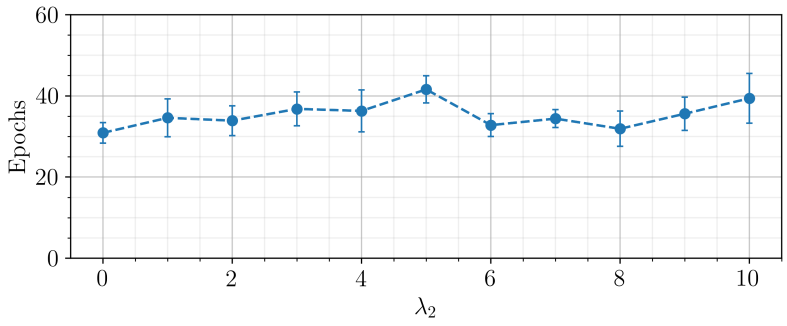

Figure 10. Average number of epochs as a function of $\lambda_2 \in [0, 10]$.

Figure 11 shows the real and averaged reconstructed signals by the lowest error model ($\lambda_2 = 2$ according to Figure 8) for the three components of the magnetic field in the ACE dataset. This reconstruction refers to data that the model has never seen in training. Therefore, it can be established that the model successfully reconstructs the signal as it follows closely the ground truth.

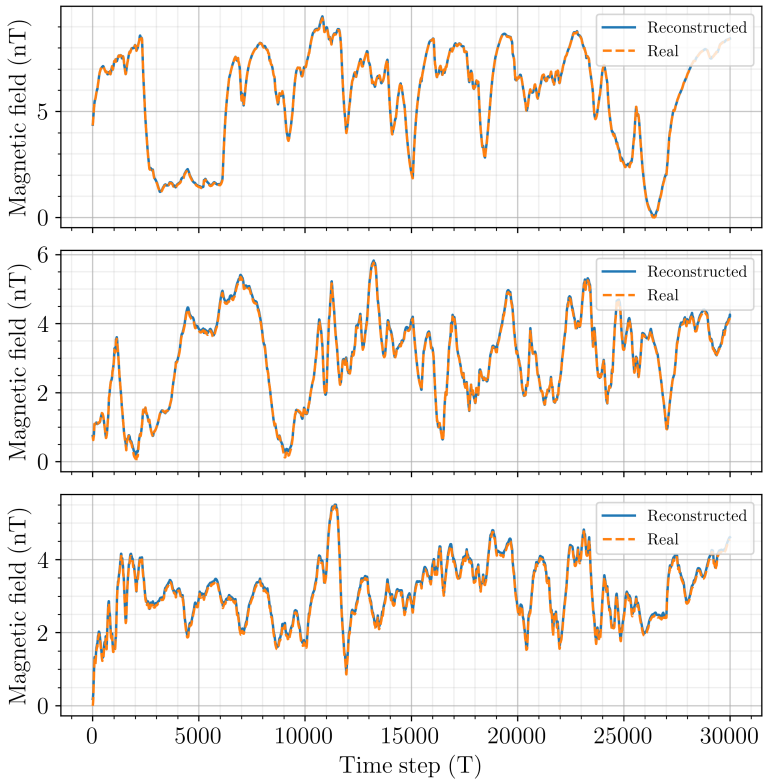

Figure 11. Real and averaged reconstructed signals for the three components of the magnetic field in the ACE dataset for the best physics regularization model. The top figure shows B_x, the middle figure shows B_y and the bottom shows B_z.

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

Curriculum Regularization As mentioned before, curriculum regularization was used to train the model while gradually increasing the weight of the physics term in the loss function (λ_2). Ten models were trained starting from different starting weights and varying the physics weighting term from zero to six. This range of λ_2 yielded better results, although it is clear that the range of λ_2 is case-dependent. The weight of the physics-based regularization term was varied in increments of one every fifteen epochs of training. Figure 12 shows the real and averaged reconstructed signals for the three magnetic components using the curriculum regularization scheme mentioned before. As it can be seen, the reconstructed signals in this figure follow more closely the real values when compared to Figure 11.

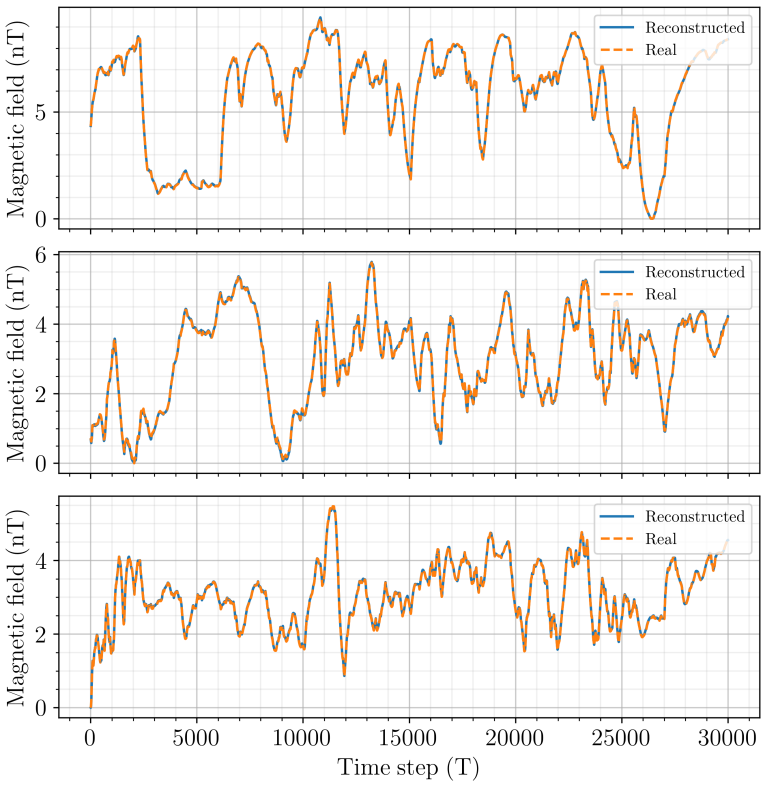

Figure 12. Real and averaged reconstructed signals for the three components of the magnetic field in the ACE dataset using curriculum regularization. The top figure shows B_x, the middle figure shows B_y and the bottom shows B_z.

Table 1 shows the average reconstruction mean squared error between the real and reconstructed signal for the two different training techniques used. From the results obtained, it can be seen that the curriculum regularization training scheme achieves a lower reconstruction error when compared to the physics regularization used in the previous section. The improvement found shows that slowly increasing the weight of the regularization portion in the loss function can help achieve a lower reconstruction error which fits with previous results reported in literature.[15] Still, characteristics

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

such as determining the model's sensitivity to the rate of increase of the regularization term are questions that need to be answered but lay beyond the scope of this work.

Table 1. Comparison between the best averaged physics informed model ($\lambda_2 = 2$) and the averaged model trained by curriculum regularization.

Training Technique	Mean Squared Error
Physics Regularization	0.00609 ± 0.00113
Curriculum Regularization	0.00324 ± 0.00059

Having shown that the inclusion of physics in the model's loss function can help reconstruct the desired signal, the classification procedure was studied and analyzed.

Classification

As mentioned above, two different thresholding techniques were studied for the classification procedure. Once the SMAP/MSL model was trained for signal reconstruction, the Gaussian assumption technique was first used to classify the reconstruction error as anomalous or nominal. Using $\alpha = 1$, the anomaly threshold was found as discussed before. Having determined the anomaly threshold for each channel, a line search was performed to find the optimal pruning threshold to maximize the $F_{0.5}$ score.

In Figure 13 the line search for the pruning threshold in channels T-9 and F-8 is shown. First, it can be seen that the pruning procedure for the F-8 channel improves the classification accuracy of the gaussian assumption classifier since the $F_{0.5}$ score increases as the pruning threshold grows. The optimal pruning threshold for this channel (defined as the lowest threshold that maximizes the $F_{0.5}$ score) lies close to 0.6, meaning that only the highly significant anomalies are being considered, and the less important anomalies (possibly false positives) are being pruned. The pruning procedure for the T-9 channel brings no improvement to the $F_{0.5}$ score. Since the pruning procedure takes care of eliminating false positives, seeing no improvement in the line search simply means that the initial anomaly threshold results in no false positives.

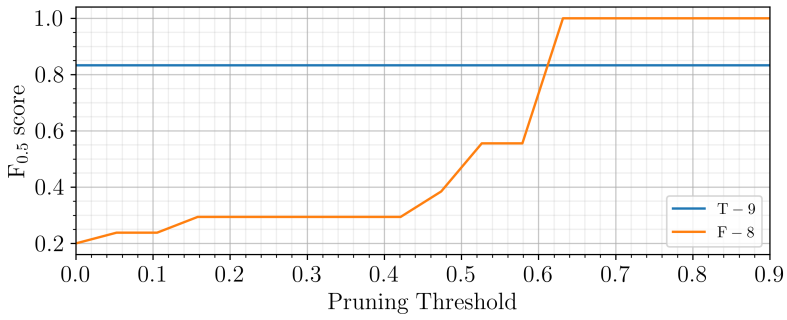

Figure 13. $F_{0.5}$ score as a function of the pruning threshold for the predicted anomalies found using the gaussian assumption.

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

The behavior discussed previously can be seen more clearly in Figure 14. Figure 14 shows the true, predicted, and pruned anomalies for the channels T-9 and F-8. In channel T-9, the initial anomaly threshold sets no false positives, and the pruned anomalies are the same as the initially predicted anomalies. Setting no false positives means that the anomaly threshold is set correctly to maximize precision. Regardless, the anomaly threshold set for channel T-9 fails to predict one true anomaly, suggesting that recall can still be improved by using other thresholding techniques. Channel F-8 shows the importance of pruning after predicting anomalies. As shown, while the predicted anomalies include the true anomaly, multiple false positives are also present. The pruning procedure eliminates these false positives, given that the most significant reconstruction error is found at the true anomaly. If the true anomaly is not associated with the most significant reconstruction error, the pruning procedure may decrease recall, which shows the importance of the regression portion of the autonomous anomaly detection problem. This same process was done for the rest of the channels in the SMAP/MSL dataset, where the global results can be seen in table 2.

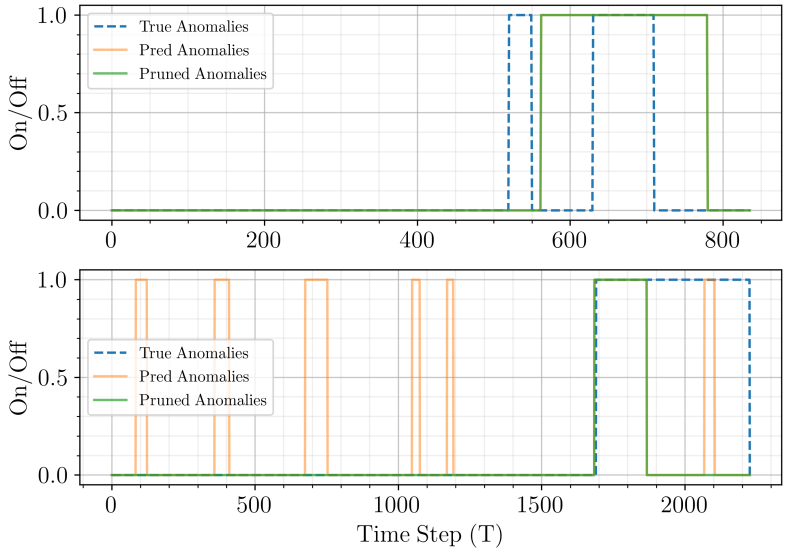

Figure 14. True, predicted, and pruned anomalies for channel T-9 (top) using the Gaussian assumption threshold. True, predicted, and pruned anomalies for channel F-8 (bottom) using the Gaussian assumption threshold.

Following with the second technique, K-means clustering was used to find the anomaly threshold. Figure 15 shows the K-means clustering classification technique. The algorithm divides the reconstruction error into two clusters, one for nominal behavior and one for anomalies. The anomaly threshold can be found by selecting the lowest error in the anomalies cluster. Thus, a base anomaly threshold was obtained by performing this procedure for each channel in the SMAP/MSL dataset. Once again, having determined the base anomaly threshold for the K-means clustering classifier, a line search was performed to find the best pruning threshold for each channel.

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

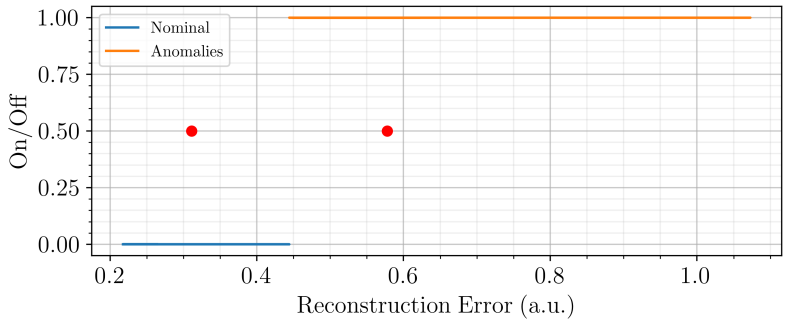

Figure 15. Error classification using K-means clustering for an arbitrary channel (E-11). The red dots represent the centroids of each cluster.

Figure 16 shows the line search for the pruning threshold regarding the K-means clustering classifier in channels T-9 and F-8. Just as with the Gaussian assumption model, the $F_{0.5}$ score for the F-8 channel increases as the pruning threshold grows. In this case, the optimal pruning threshold is higher when compared to the Gaussian assumption model. A higher pruning threshold means that the initial anomaly threshold is set at a lower number. This is also true for channel T-9; as the anomaly threshold is set lower, the classifier can obtain a perfect score by predicting all true anomalies while still having no false positives. Just as with the Gaussian assumption, the K-means classifier predicts no false positives for channel T-9, meaning that the pruning procedure results in no improvement when considering the $F_{0.5}$ score.

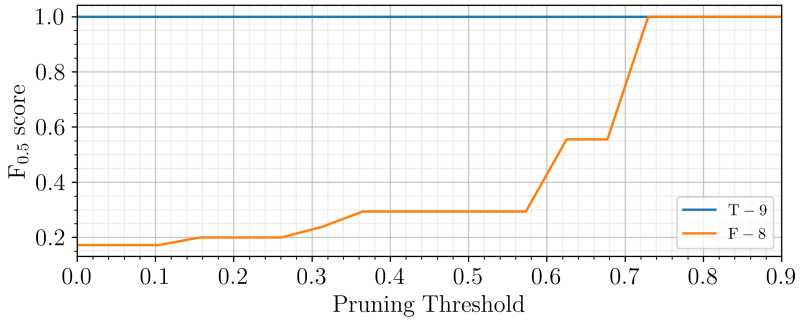

Figure 16. $F_{0.5}$ score as a function of the pruning threshold for the predicted anomalies found using the K-means clustering algorithm.

Figure 17 shows the pruning procedure in the two channels discussed above. In the same way as the Gaussian assumption model, the initial anomaly threshold found with the K-means model for channel T-9 sets no false positives. The pruned anomalies are the same as the predicted anomalies. However, the anomaly threshold set by the K-means model can identify all true anomalies in the signal, meaning that for this specific case, the K-means thresholding technique is more appropriate. For channel F-8, it can be seen that the predicted anomalies found with the K-means thresholding technique result in a higher number of false positives than with the Gaussian assumption technique. Regardless, the pruning procedure removes these false positives and achieves a perfect score. Once

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

again, this is only possible since the highest reconstruction error found by the regression portion is precisely at the true anomaly.

Table 2 summarizes the generalized $F_{0.5}$ scores for the model configurations presented above. To calculate each $F_{0.5}$ score in this table, the total number (including every channel) of true positives, false positives and false negatives for each satellite was found, and Equation 11 was used to calculate the resulting score. This table shows that the K-means clustering technique obtains a better $F_{0.5}$ score in both satellites being studied compared to the Gaussian assumption technique. These results show that, as mentioned before, reconstruction errors will not always follow a Gaussian distribution, leaving other thresholding techniques better suited for anomaly detection. Regardless, straightforward techniques such as line searches have been shown to achieve satisfactory results.

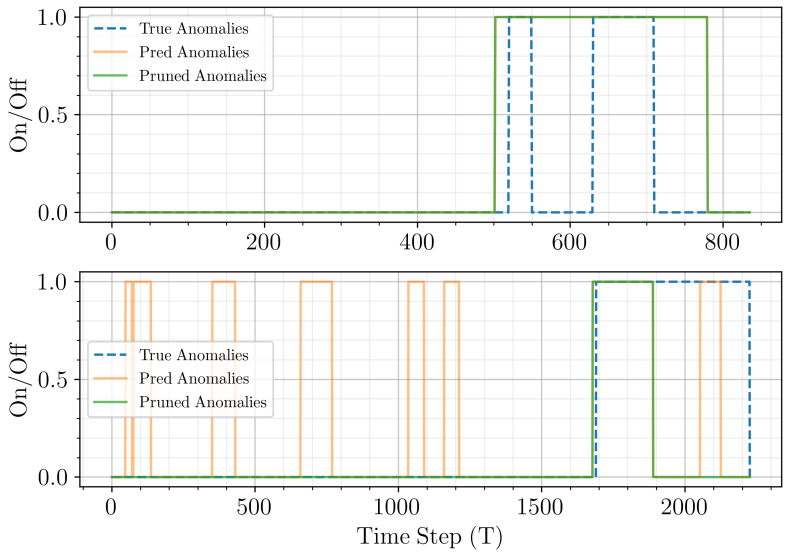

Figure 17. True, predicted, and pruned anomalies for channel T-9 (top) using the K-means clustering threshold. True, predicted, and pruned anomalies for channel F-8 (bottom) using the K-means clustering threshold.

Table 2. Comparison between the different thresholding techniques. The first score refers to SMAP and the second score to MSL.

Thresholding Technique	$\mathbf{F_{0.5}}$ **score** (SMAP)	$\mathbf{F_{0.5}}$ **score** (MSL)
Gaussian assumption	0.843	0.781
K-means clustering	0.869	0.938

Having encountered the best configurations for the classifiers, the results were compared with the initial results obtained by the original paper[3] from which the dataset was published. As mentioned before, the SMAP/MSL model used in this paper for the signal reconstruction is the same as the one presented by Hundman et al.,[3] which allows a direct comparison of the thresholding techniques. In

Figure 18, the $F_{0.5}$ score for the original paper's classification model, which uses a non-parametric dynamic thresholding technique, the Gaussian assumption classification model, and the K-means clustering classification model is shown. This figure shows that the original paper outperforms the Gaussian approach. In contrast, the K-means clustering technique outperforms the original article in both satellites. With this, it can be seen that out of the three thresholding techniques considered, K-means clustering accompanied by the pruning procedure described before has the highest $F_{0.5}$ score.

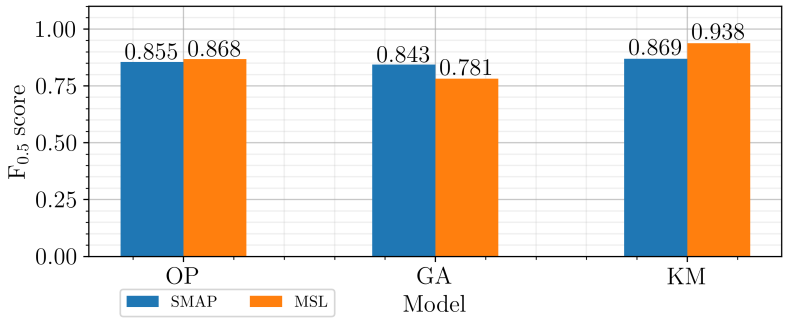

Figure 18. $F_{0.5}$ score for the *Soil Moisture Active Passive* (SMAP) and *Mars Science Laboratory* (MSL) using the original paper's classification model[3] (OP), the Gaussian assumption classification model with pruning (GA) and the K-means clustering classification model with pruning (KM).

CONCLUSION

The results presented above show improvements in the two sub-problems that make up the autonomous threat detection problem.

Regarding the regression sub-problem, it was shown that when using a data-driven machine learning algorithm to reconstruct a signal, including the known physics of the signal into the model's loss function can lower the reconstruction error by restricting the model from becoming too complex. Also, by varying the weight of the physics in the loss function, an optimal balance was established between the data's importance and the physics' importance in the model. Finding an optimal balance showed that a purely physics-based or entirely data-driven model will not perform as well as a model relying on both physics and data. As well as this, by using a curriculum regularization training scheme, it was found that slowly increasing the weight of the regularization term in the loss function can yield a lower reconstruction error.

For the classification sub-problem, the results obtained showed that the thresholding process is case-based, as some thresholding approaches might work better for different applications. In the case of the two satellites being studied, the $F_{0.5}$ score obtained with the K-means clustering algorithm outperformed previous models relying on dynamic thresholding techniques. The two thresholding methods studied showed that a suitable regression will yield a high recall when classifying anomalies, and the pruning procedure can take advantage of the proper regression results to increase the precision of the model.

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

ACKNOWLEDGMENT

This material is based on research sponsored by Air Force Research Laboratory (AFRL) under agreement number FA9453-21-1-0045. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.

REFERENCES

[1] R. Chalapathy and S. Chawl, "Deep Learning for Anomaly Detection: A Survey," *CoRR*, Vol. abs/1901.03407, 2019.

[2] W. Jin, B. Sun, Z. Li, S. Zhang, and Z. Chen, "Detecting Anomalies of Satellite Power Subsystem via Stage-Training Denoising Autoencoders," *Sensors*, Vol. 18, 14, p. 3216.

[3] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Söderström, "Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding," *Proceedings of the 2018 International Conference on Knowledge Discovery Data Mining*, London, United Kingdom, July 2018.

[4] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection," *Proceedings of the 2016 International Conference in Machine Learning Anomaly Detection Workshop*, New York, NY, June 2016.

[5] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, Vol. 378, 2018, pp. 686–707.

[6] E. Stone, A. Frandsen, R. Mewaldt, E. Christian, D. Margolies, J. Ormes, and F. Snow, "The Advanced Composition Explorer," *Space Science Reviews*, Vol. 86, No. 1/4, 1998, pp. 1–22.

[7] D. J. Griffiths, *Introduction to Electrodynamics*. Boston, MA: Pearson, 2013.

[8] J. D. Jackson, *Classical Electrodynamics*. New York, NY: Wiley, 1999.

[9] I. Knowles and R. J. Renka, "Methods for Numerical Differentiation of Noisy Data," *Proceedings of the 2012 Variational and Topological Methods: Theory, Applications, Numerical Simulations, and Open Problems Conference*, Flagstaff, AZ, June 2012.

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[11] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," *Proceedings of the 2021 International Conference on Telecommunications and Signal Processing*, Virtual Conference, July 2021.

[12] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, "Automatic Differentiation in Machine Learning: A Survey," *Journal of Machine Learning Research*, Vol. 18, No. 1, 2017, p. 5595–5637.

[13] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: with applications in R*. New York, NY: Springer, 2017.

[14] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, and I. Mitliagkas, "A Modern Take on the Bias-Variance Tradeoff in Neural Networks," *CoRR*, Vol. abs/1810.08591, 2018.

[15] A. S. Krishnapriyan, A. Gholami, S. Zhe, R. M. Kirby, and M. W. Mahoney, "Characterizing possible failure modes in physics-informed neural networks," *Proceedings of the 2021 Advances in Neural Information Processing Systems Conference*, Virtual Conference, Dec. 2021.

[16] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, Vol. 9, No. 8, 1997, pp. 1735–1780.

[17] J. P. Buzen and A. W. Shum, "MASF - Multivariate Adaptive Statistical Filtering," *Proceedings of the 1995 International Computer Measurement Group Conference*, Nashville, TN, Dec. 1995.

[18] C. Wang, K. Viswanathan, L. Choudur, V. Talwar, W. Satterfield, and K. Schwan, "Statistical techniques for online anomaly detection in data centers," *Proceedings of the 2011 IFIP/IEEE International Symposium on Integrated Network Management and Workshops*, Dublin, Ireland, May 2011.

[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Proceedings of the 2015 International Conference for Learning Representations,*, San Diego, CA, May 2015.

Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-3350. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the US Government.