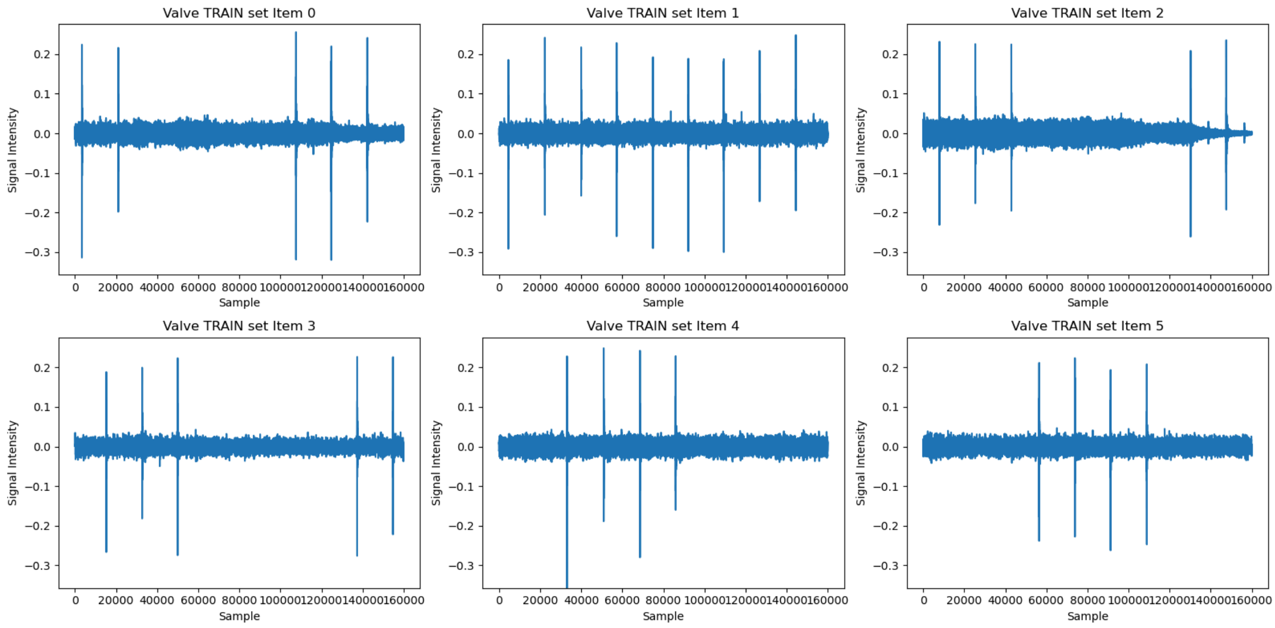
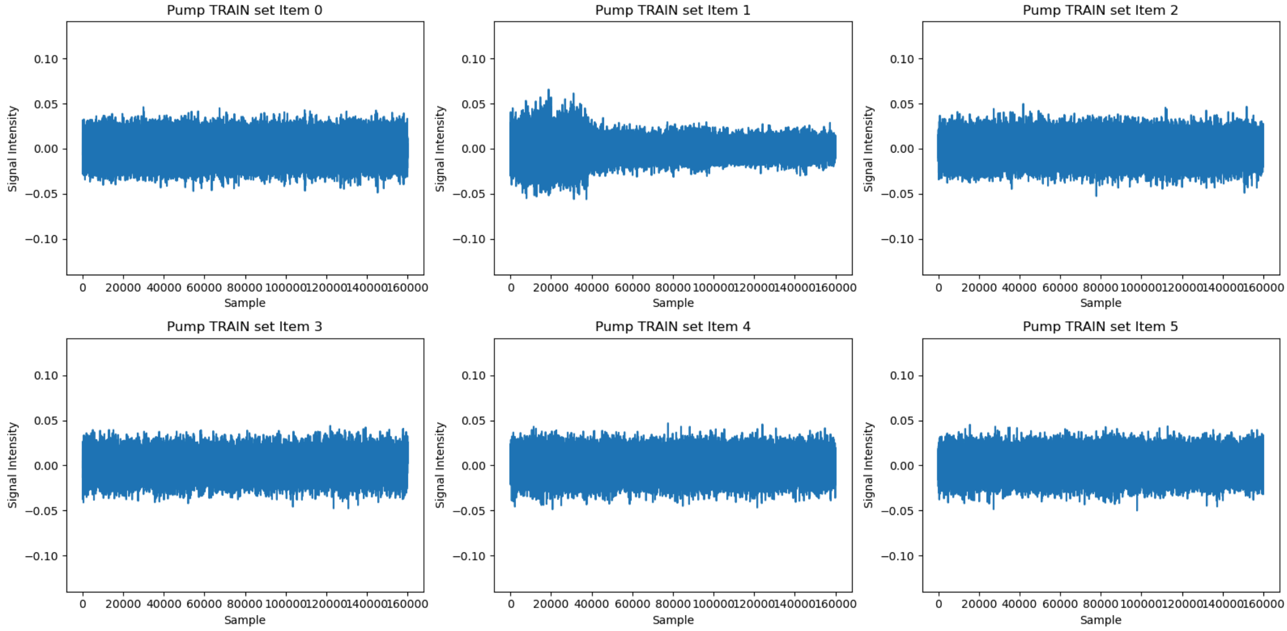
Sophea Bonne, 352901
Joshua Cohen-Dumani, 311105

ML for Predictive Maintenance Applications: Assignment 1

Q1 - Generate plots of raw signals, FFT spectrums or spectrograms from the healthy data of both the pump and valve acoustic signals. Discuss the distinctions between the signals emitted by these two machines.

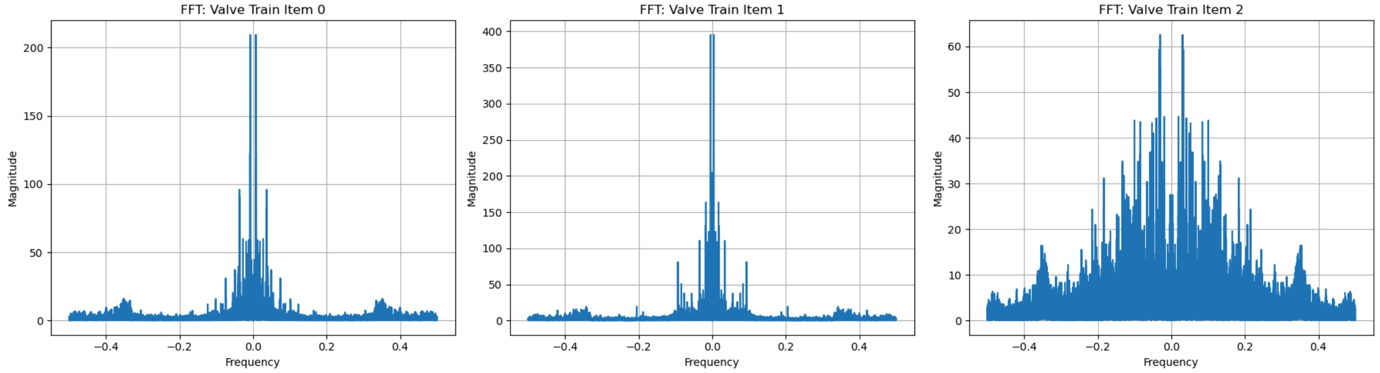
First, let's visualize the raw signals of the valve and pump. We are looking at the first 6 samples of the training set (ie healthy data).

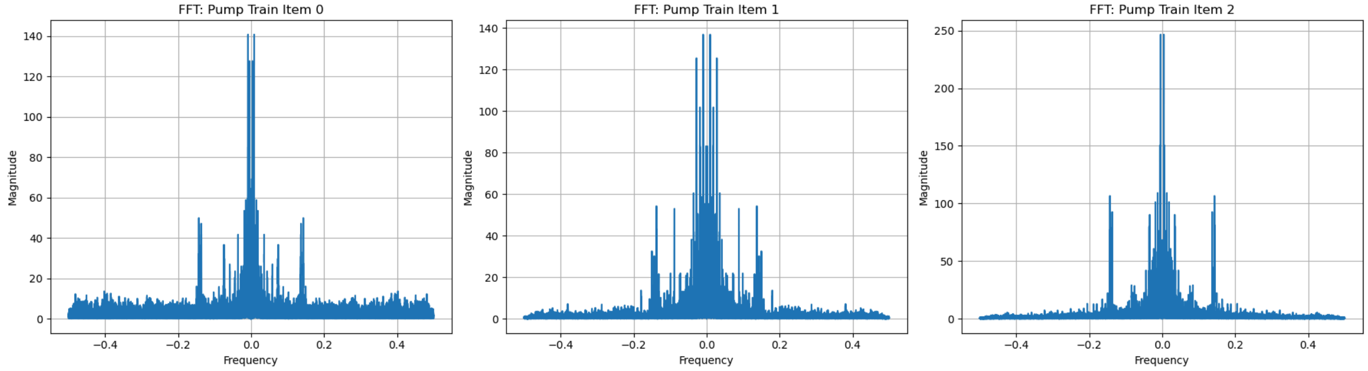




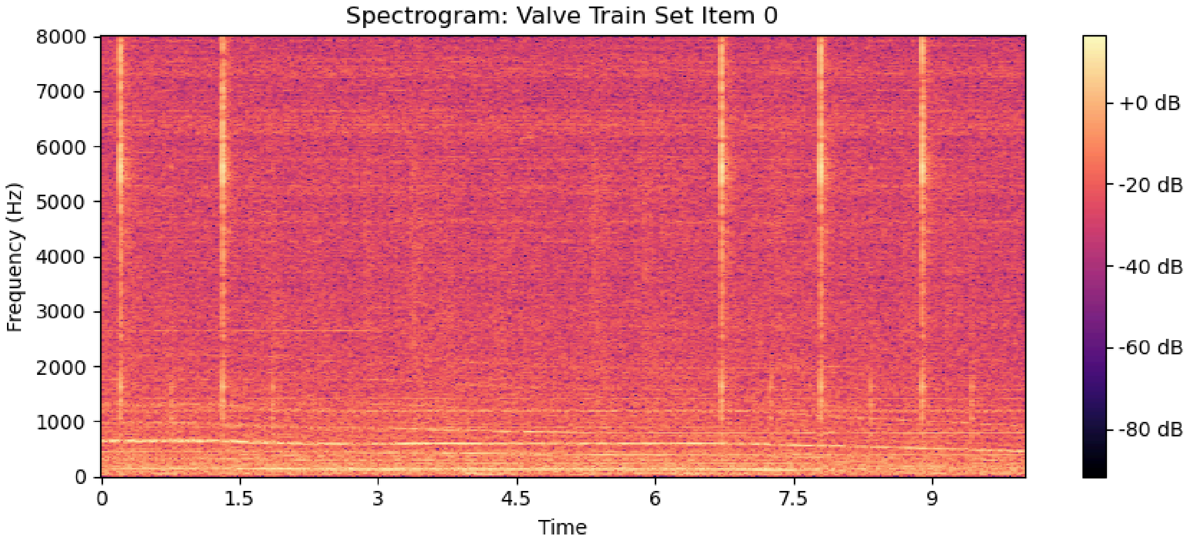
The valve signals have smaller amplitudes with a few samples of higher amplitude. In contrast, the pump signals have greater amplitude amplitude and less variance in amplitude.

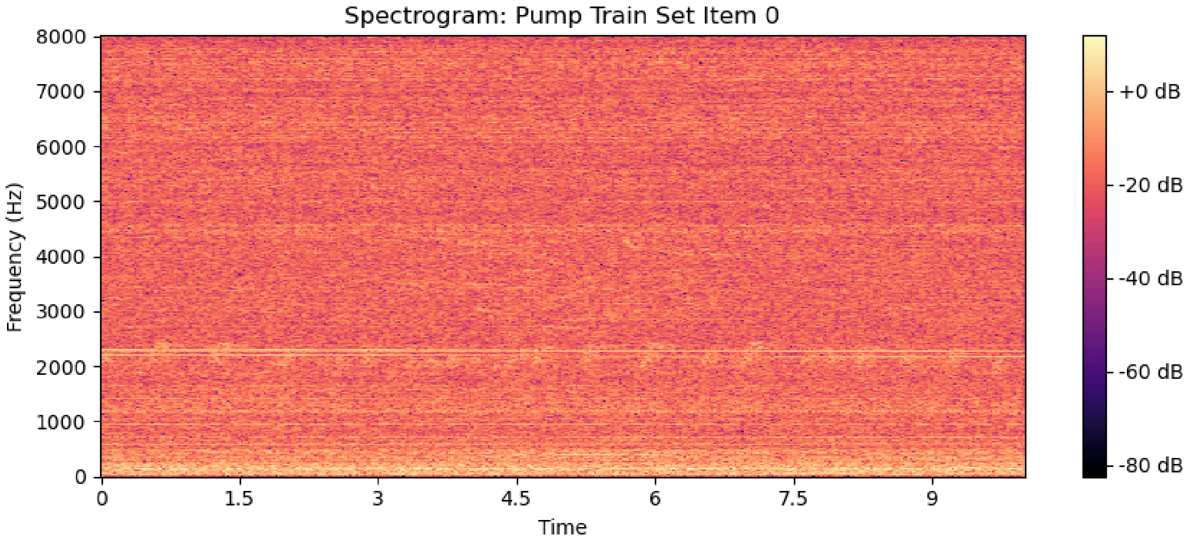
Then we plot the fast fourier transform of these signals, to get a better understanding of the frequency profile of the valve and pump. It looks like they have the biggest peak around 0, with some smaller symmetrical peaks at greater absolute frequency.



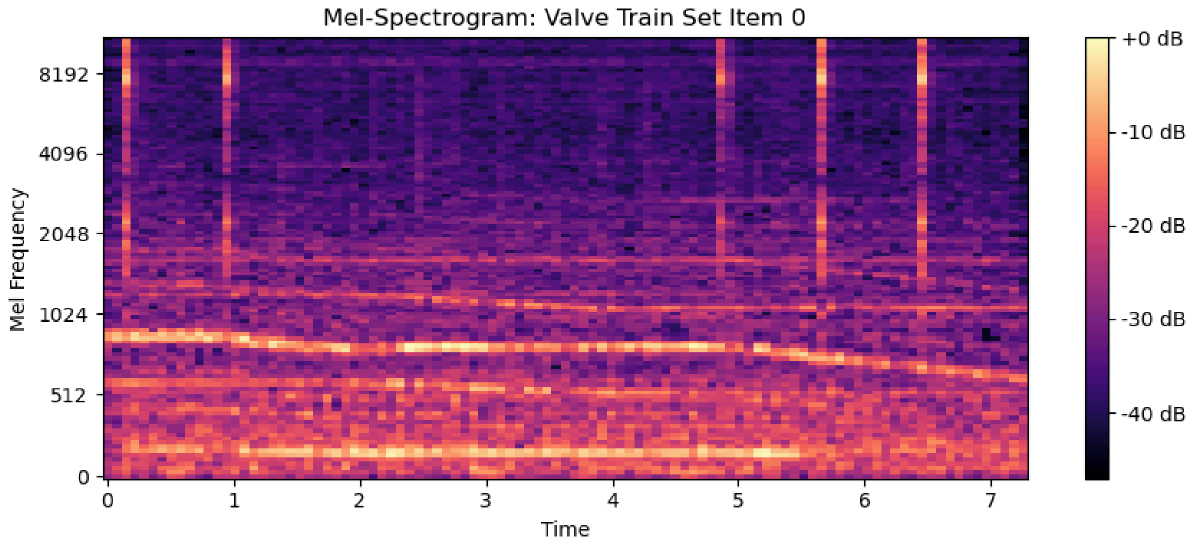


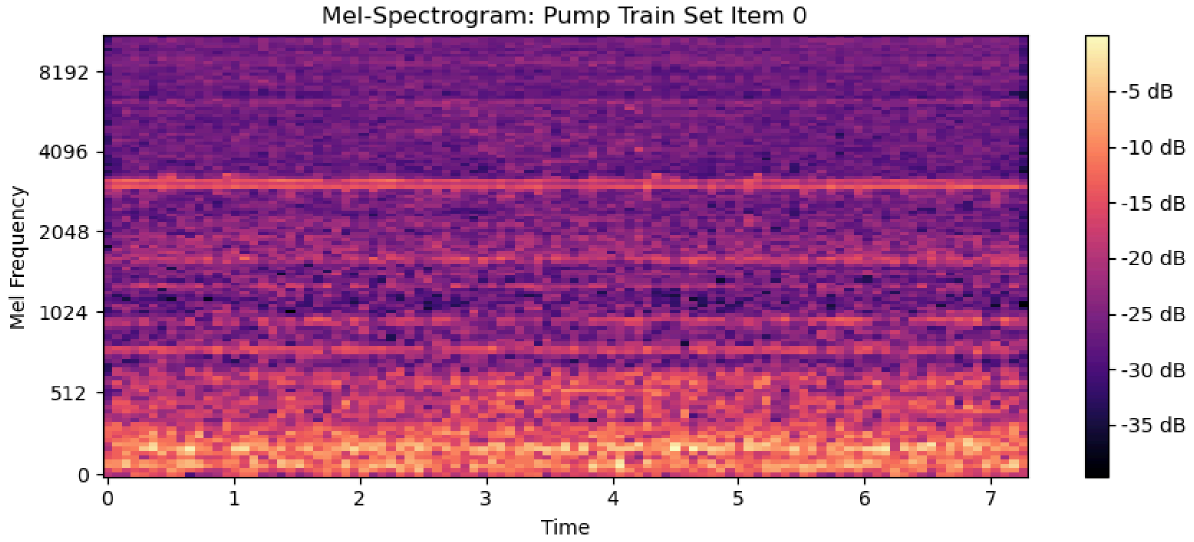
To get a better understanding of the time evolution in the frequency domain, we plot the spectrogram of a sample from both the pump and valve datasets. For the valve, we see that there are a few instances in time where higher frequencies dominate. For example, around 0.1 and 1.4 seconds. For the pump, there is a frequency with constant amplitude across time.





To get a better understanding of what the signal sounds like, let's plot the mel spectrograms for a sample of the valve and pump datasets.





The main difference between the valve and pump seems to be that most samples from the valve have spikes to 0.1-0.2 somewhat regularly (which signals when the valve opens and closes), while the pump has a much more constant signal amplitude, confined between, roughly, ±0.05.

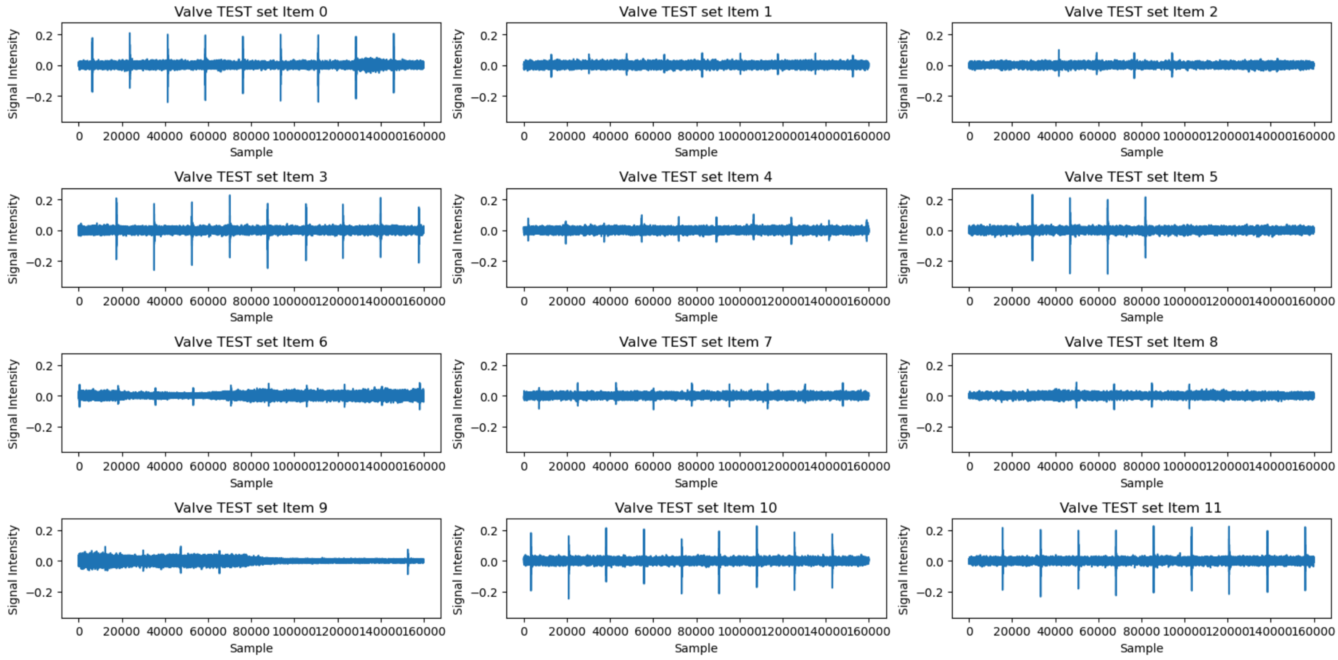
This is also visible in the spectrograms, which clearly show these spikes for the valve while it is much more consistent for the pump.

Questions 2, 3 and 4 have to be answered separately for both the pump and valve datasets. We will answer all the questions for the valve, then all the questions for the pump.

VALVE

Q2 - Visualize the raw signals, spectrum, and spectrograms of the test dataset for the pump/valve dataset. Are there any signals that appear abnormal to you?

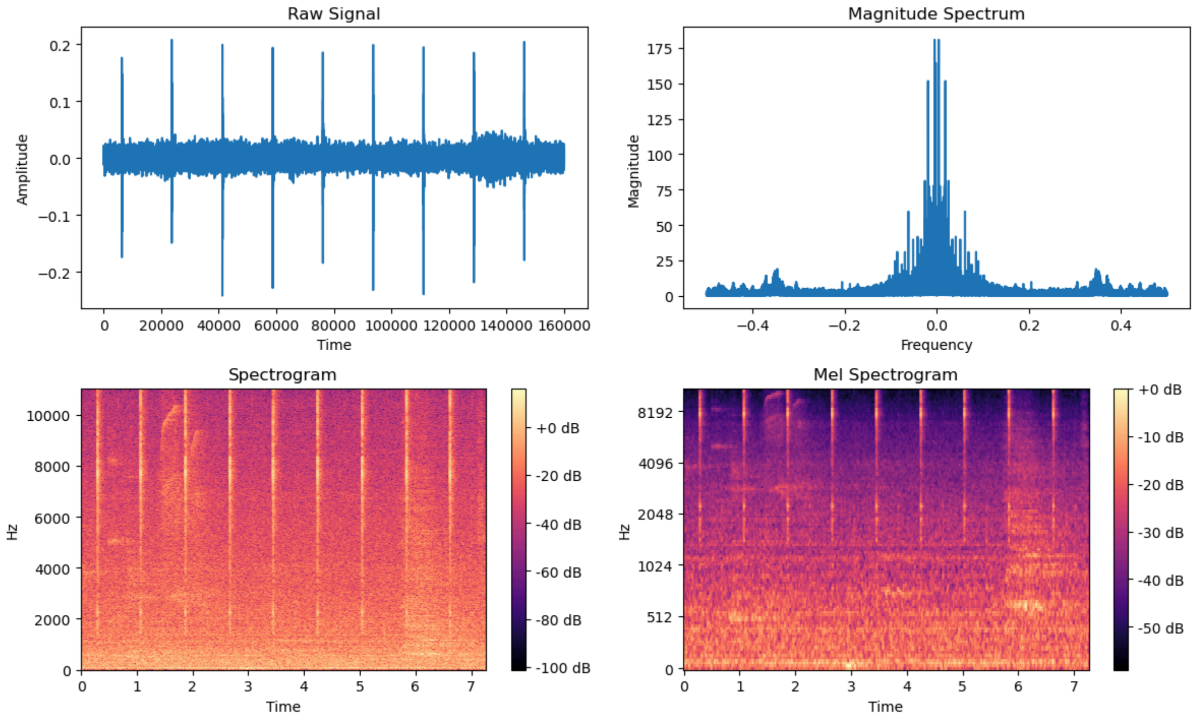
Here we show the raw signals of the first 12 valve test items. We suspect there are a lot of anomalies, for example sample plot 9 where the center magnitude is greater at the beginning but then shrinks after sample 80000.



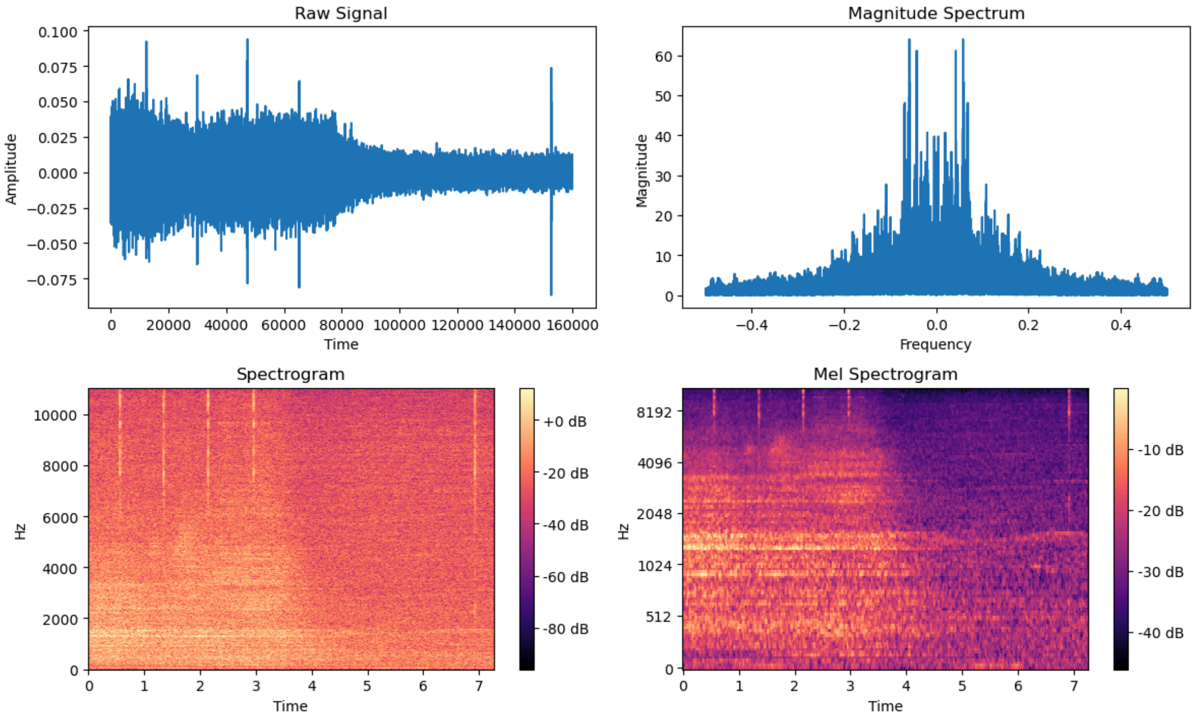
By looking at the raw test dataset we found many abnormal signals (1, 2, 4, 6, 7, 8, 9, 18, 19, 22, 25, 26, 27, 28, 29, 30, 31, 32, ...), usually characterized by spikes that stand out less (smaller peaks in the raw signal amplitude), which likely is because there is some other noise. We show example plots for a "normal" and an "abnormal" sample.

We visualize the signal, spectrum, spectrogram of a normal (0) and abnormal sample (9).

Valve TEST Set Item 0 - Normal



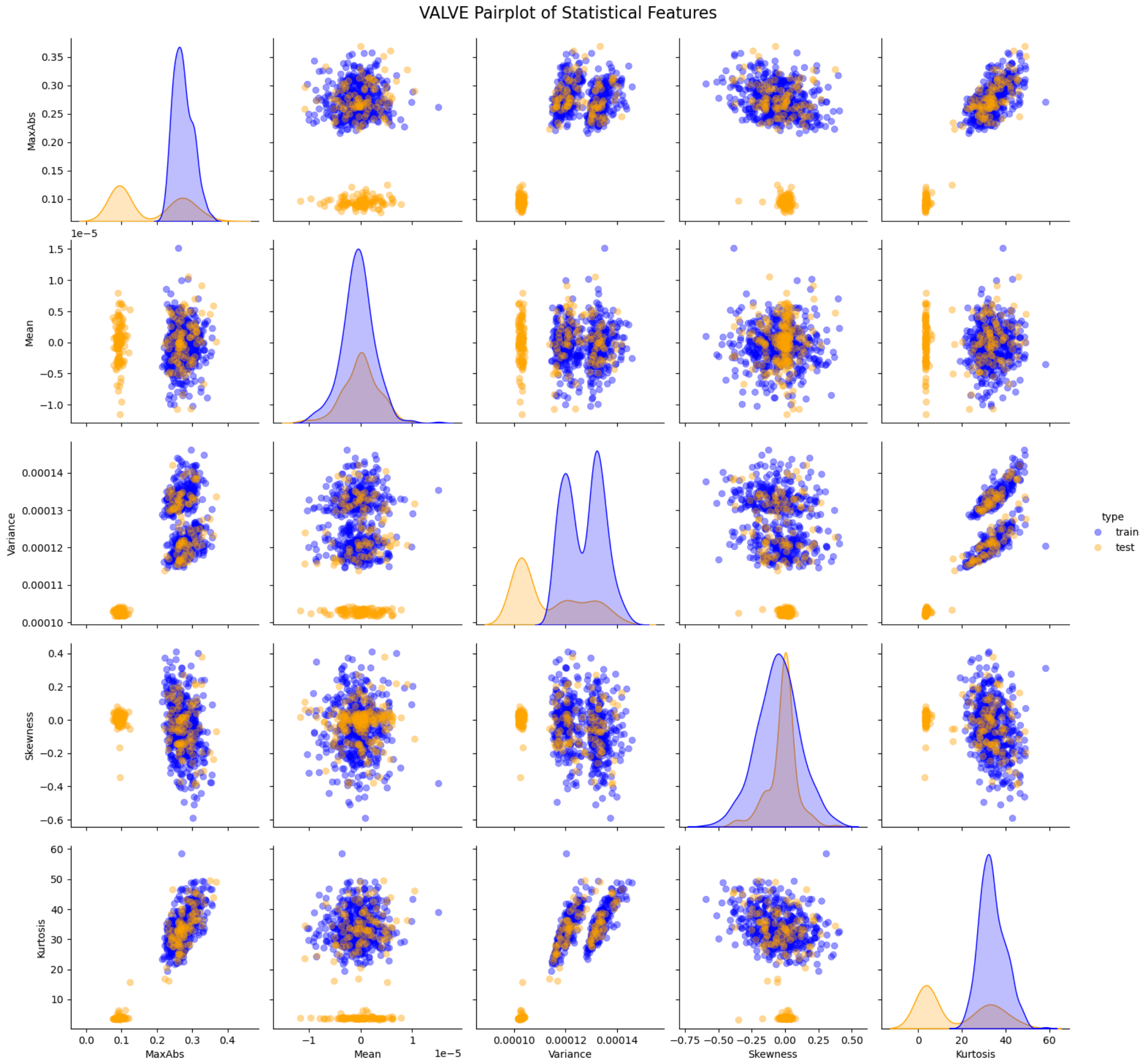
Valve TEST Set Item 9 - Abnormal



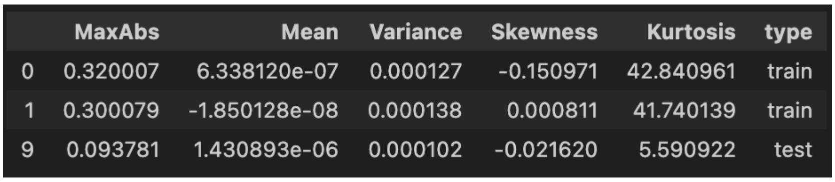
Item 9 seems abnormal because the band through the center amplitude is very wide at the beginning instead of staying constant width like in item 0. The magnitude spectrum also shows abnormalities, with 2 peaks instead of 1. As compared to item 0, the spectrogram for item 9 shows noise in the quadrant of low frequency, low time.

Q3 - Compute basic statistical features (mean, variance, skewness, and kurtosis) for both the training and test datasets of the pump and valve. Are there any abnormal signals that you can detect?

Here we show a pairplot of statistical features. The features Kurtosis, MaxAbs, and Variance show a clear separation between the train and test data. This likely indicates a source of anomalies. Perhaps the scaling is different, so we can use the threshold method next to find anomalies. The mean value also appears to be different (in amplitude), which is related to our earlier observation that there were many, seemingly abnormal samples in the test dataset with smaller spikes.



Earlier, we suspected valve item 9 was abnormal. Indeed, as shown on the below table, there seems to be about an order of magnitude difference between the MaxAbs and Kurtosis of the train items (which are healthy) and the test item 9. These large differences seem to confirm that test item 9 is abnormal. We will have a more broad view of anomalous samples in Question 4.

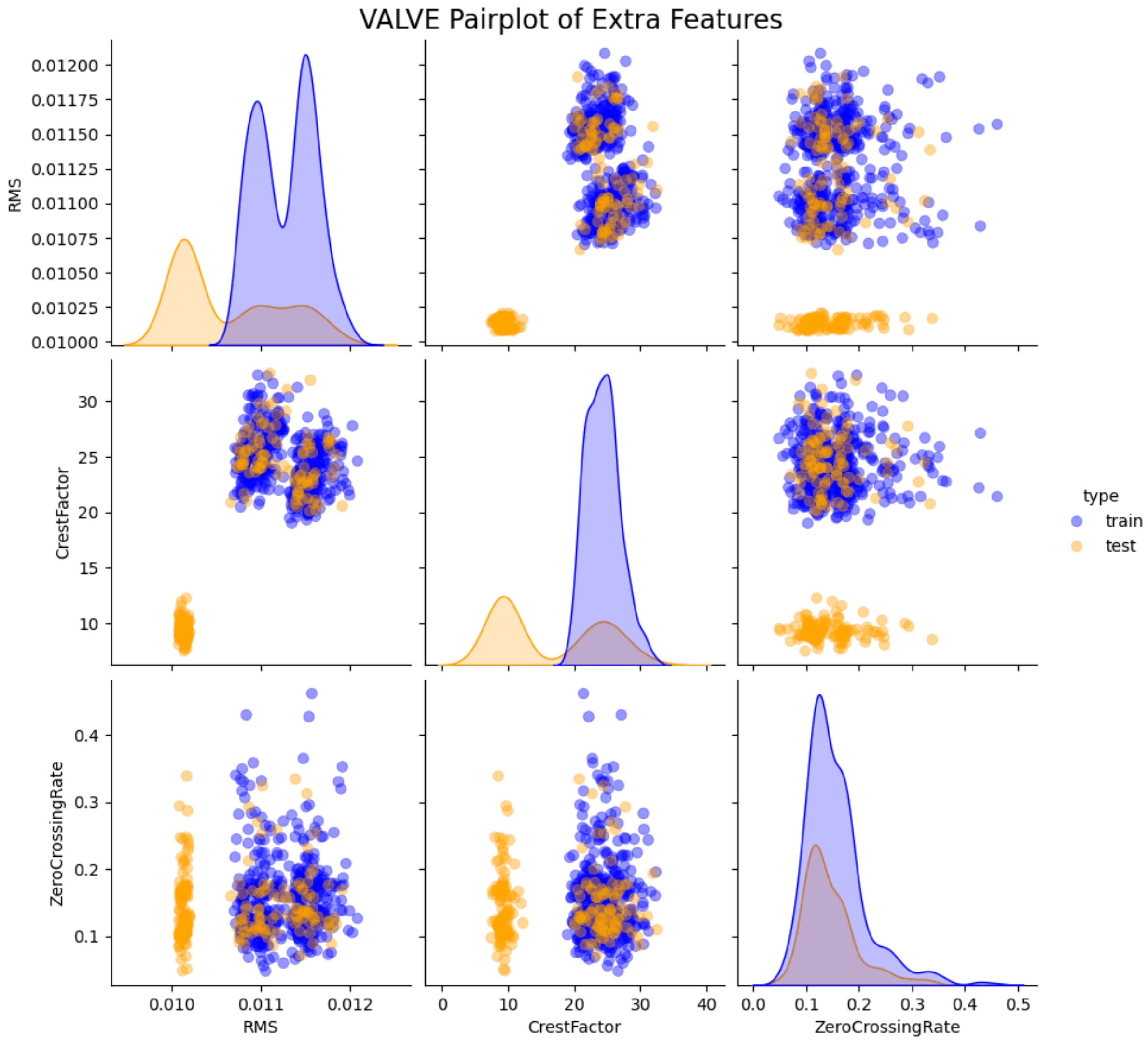


	MaxAbs	Mean	Variance	Skewness	Kurtosis	type
0	0.320007	6.338120e-07	0.000127	-0.150971	42.840961	train
1	0.300079	-1.850128e-08	0.000138	0.000811	41.740139	train
9	0.093781	1.430893e-06	0.000102	-0.021620	5.590922	test

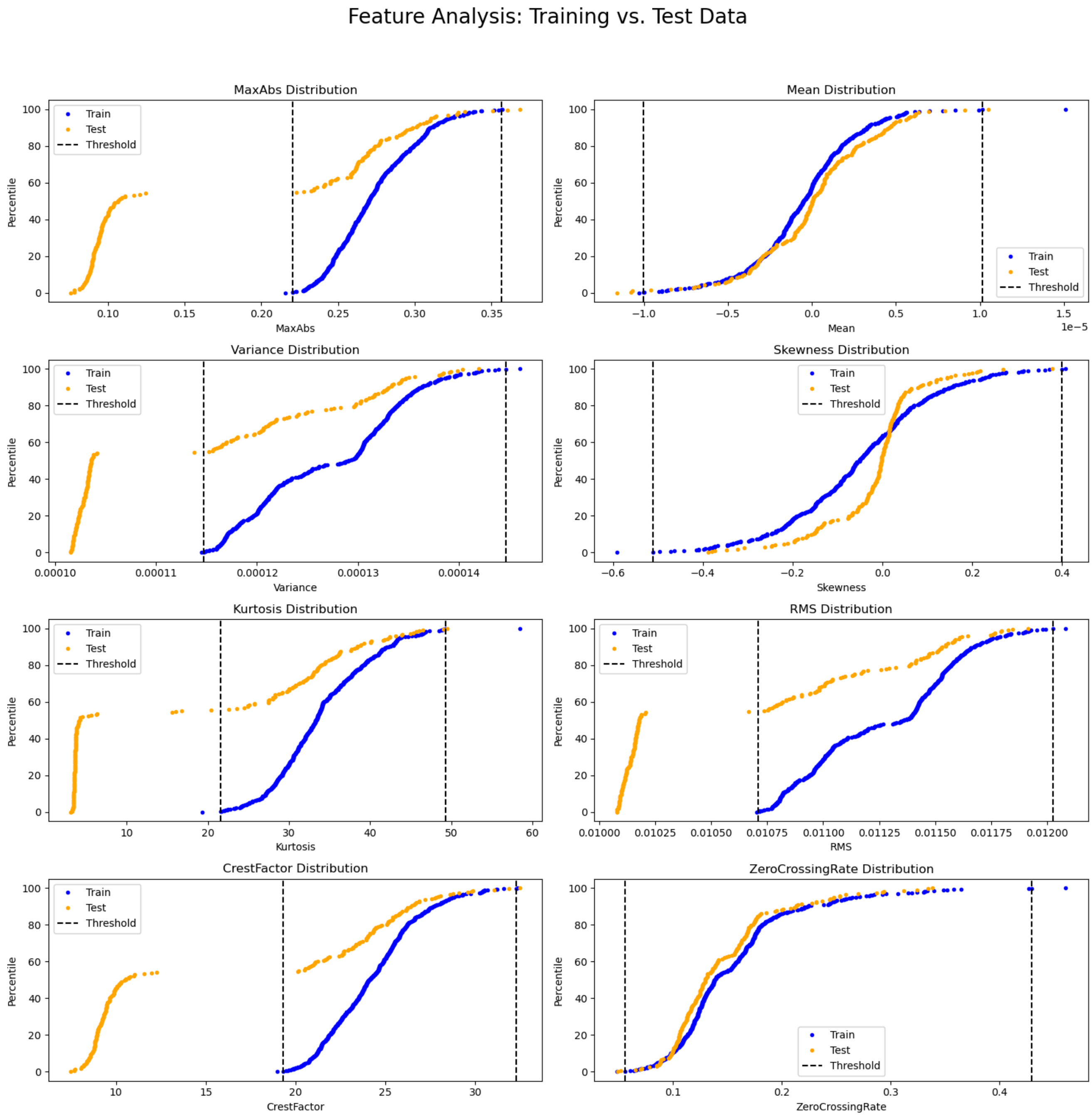
Q4 - Find by yourself a feature or a combination of features that help to uncover signals with abnormal behavior. Analyze whether the selected features trigger alarms for similar behavior or if some are specific to particular anomalies.

In the previous question we used MaxAbs, Mean, Variance, Skewness, and Kurtosis. Now, we look at:

1. The root mean square (RMS), which is commonly used to measure the "energy" of a signal
2. The crest factor (which compares the ratio of the RMS to the max absolute value, one of our most telling features based on the pairplot in Q3)
3. The zero-crossing rate (ZCR), which could be useful in determining whether a signal is more noisy than before (and generally if they have more higher frequency elements than before) by calculating how many times it crosses "zero"



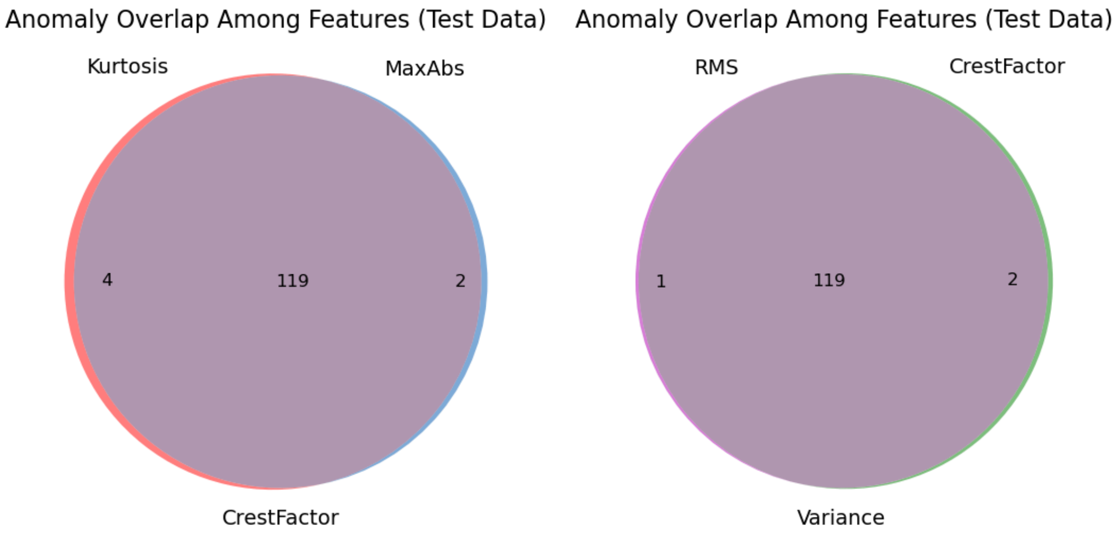
We see clear peaks in the histograms of RMS and CrestFactor, which suggests that these features are meaningfully different for the healthy data and test data. This is promising, as it suggests that our features will hold some discriminatory power. Our next step is to to create a classifier using thresholds, like used in the toy example.



It seems that MaxAbs, Variance, Kurtosis, RMS, and Crestfactor are able to detect anomalies, while the other features do not seem informative. Thus we removed the ZCR, Skewness, and Mean from our feature set.

In order to determine whether the selected features trigger alarms for similar behavior or if some are specific to particular anomalies, we counted how many anomalies were flagged by different selection methods. Our results show that in the test set, 125 samples are found as anomalous by at least 1 selection method, and 119 samples are found as anomalous by all 5 selection methods. This means that there’s great overlap, so samples that are anomalous in one metric are usually anomalous in another.

This means that theoretically using only one of these features could be enough to identify the vast majority of anomalies. However, having multiple features flag the same sample could be useful in order to build stronger conviction around some classification (anomalous/healthy). There are also a small number of samples which were only caught by one metric or another, which suggests that in practice using as many features as possible would allow us to potentially avoid false negatives.

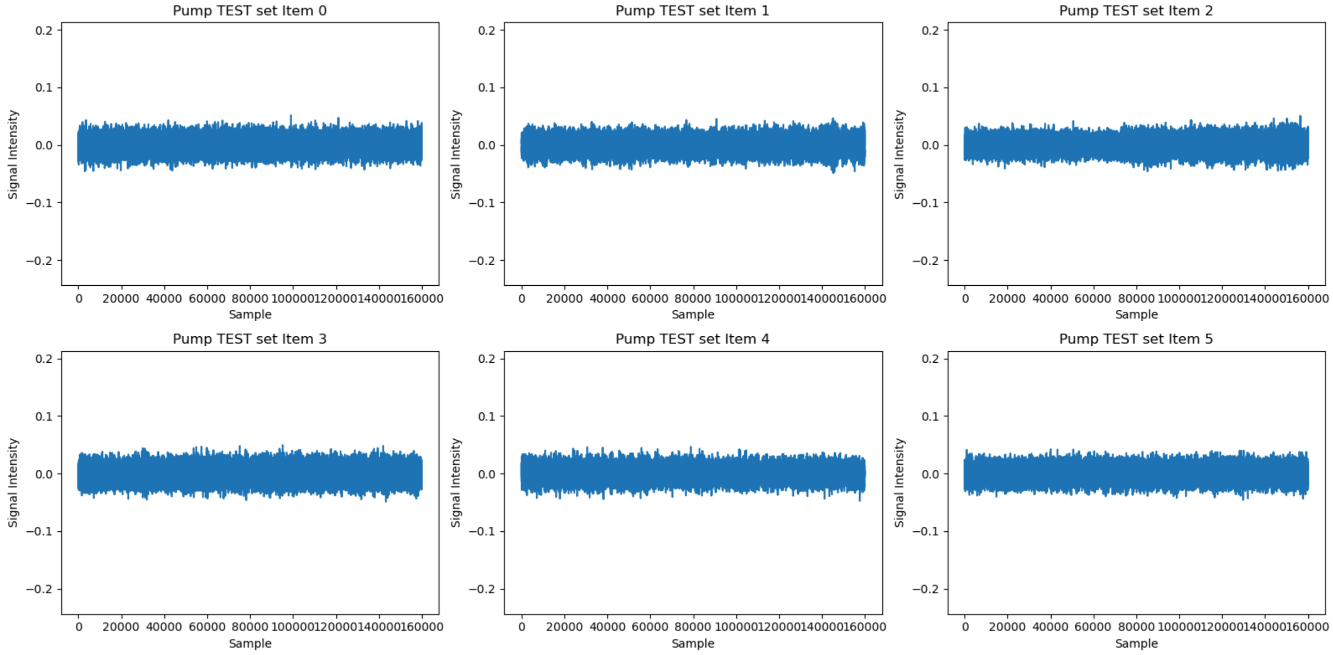


These venn diagrams showed that around 119-120 samples are consistently overlapping. This strengthens our claim from the anomaly analysis that most of the anomalies show up using most of our features. We can thus confidently classify at least 119 samples as anomalies for the valve (as measured by our model) and there are around 6 which should be looked at more closely to understand if they are anomalous.

PUMP

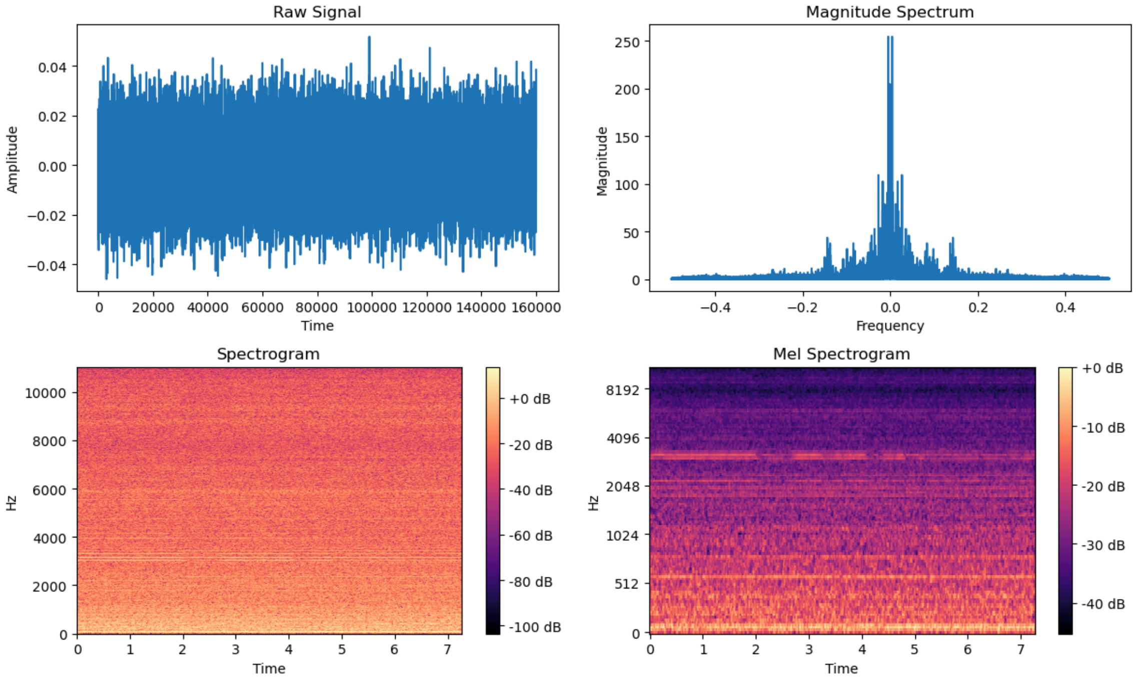
Q2 - Visualize the raw signals, spectrum, and spectrograms of the test dataset for the pump/valve dataset. Are there any signals that appear abnormal to you?

It’s harder to discern anomalies here, but by visualizing all raw signals in the test dataset, we found several anomalies (12, 34, 48, 57, 60, 80). Here we show the first 6 raw signals. (The rest are in the code).



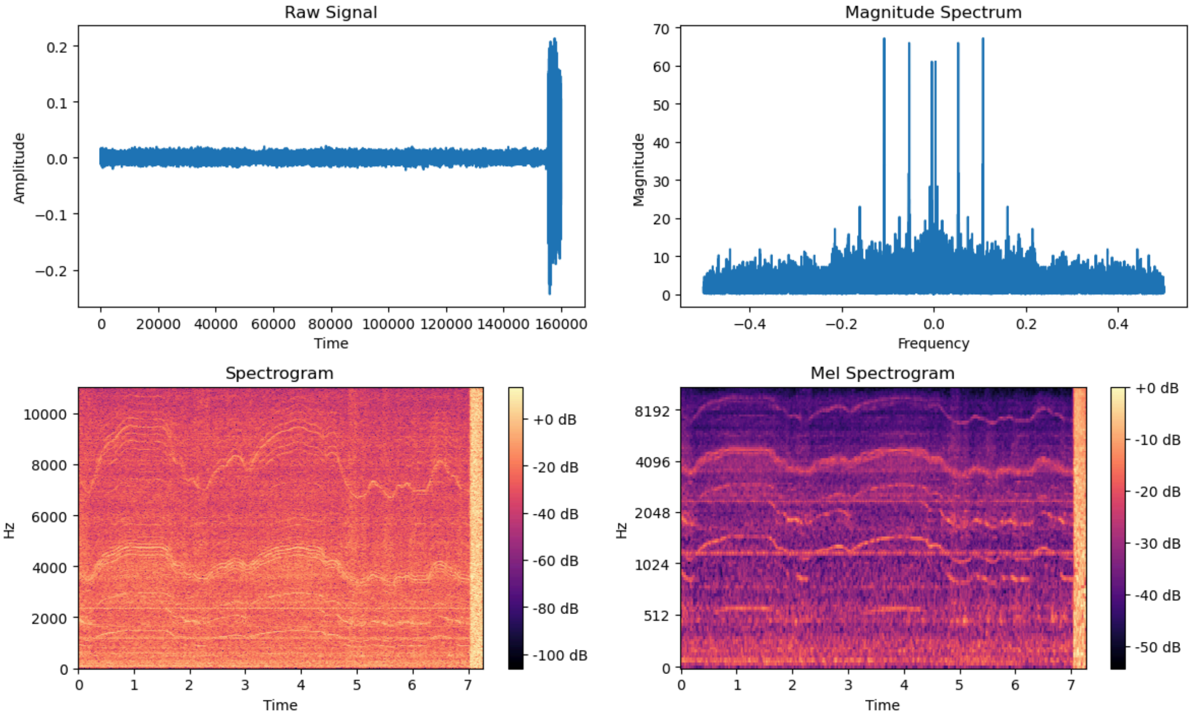
The following plot shows a closer analysis of the first raw signal, which we thought was normal.

Pump TEST Set Item 0 - Normal

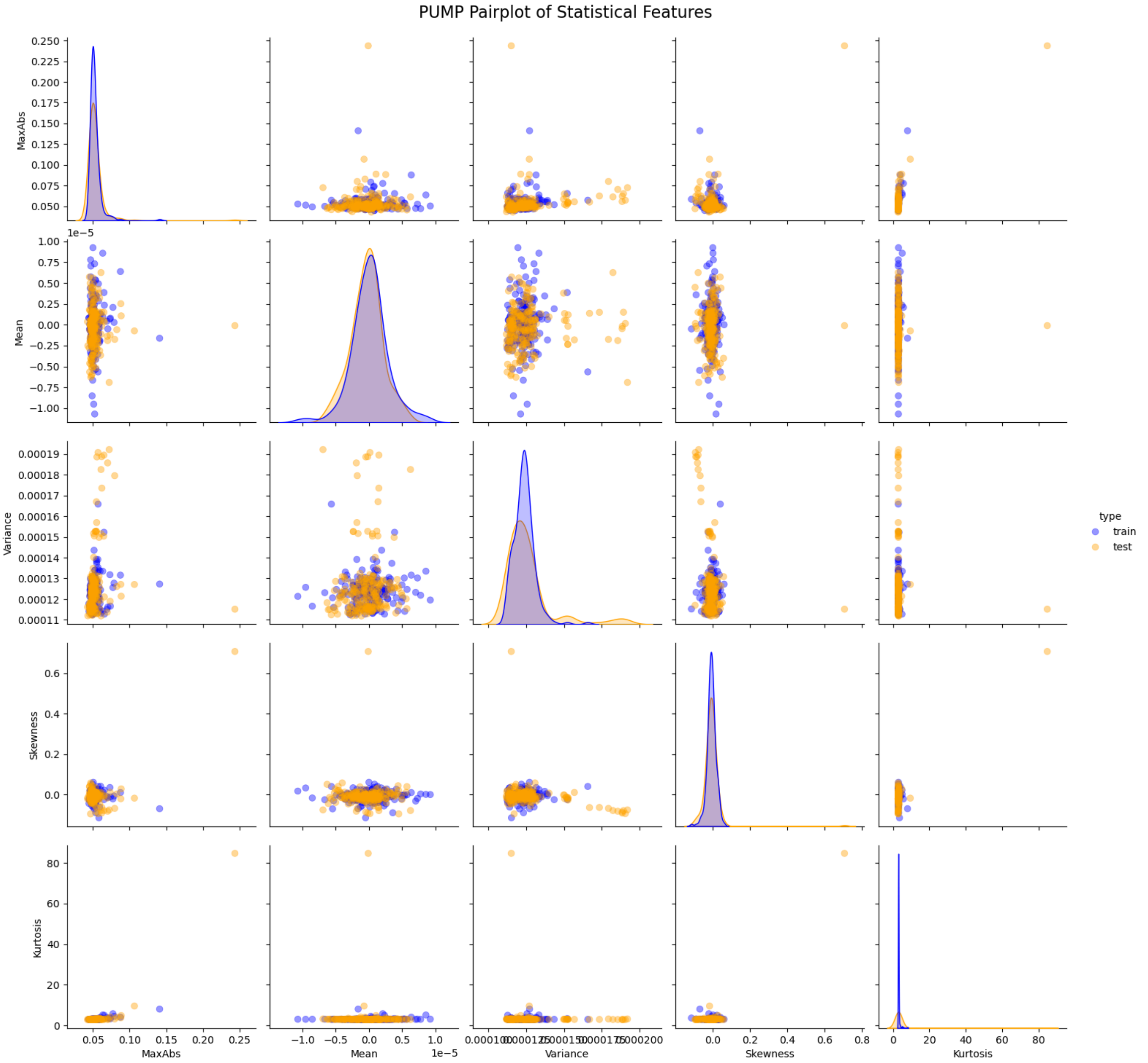


And a closer analysis of item 12, which from the raw signal we thought was abnormal. This is clearly abnormal because of the raw signal showing a huge amplitude at the end of the sample. The magnitude spectrum also has many peaks instead of a normal single peak, and the spectrograms have waves instead of constant lines in the plot.

Pump TEST Set Item 12 - Abnormal

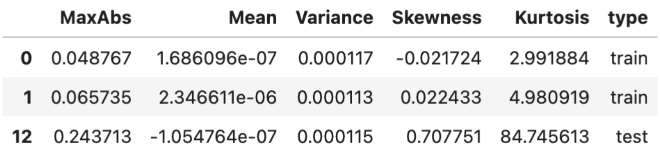


Q3 - Compute basic statistical features (mean, variance, skewness, and kurtosis) for both the training and test datasets of the pump and valve. Are there any abnormal signals that you can detect?


PUMP Pairplot of Statistical Features

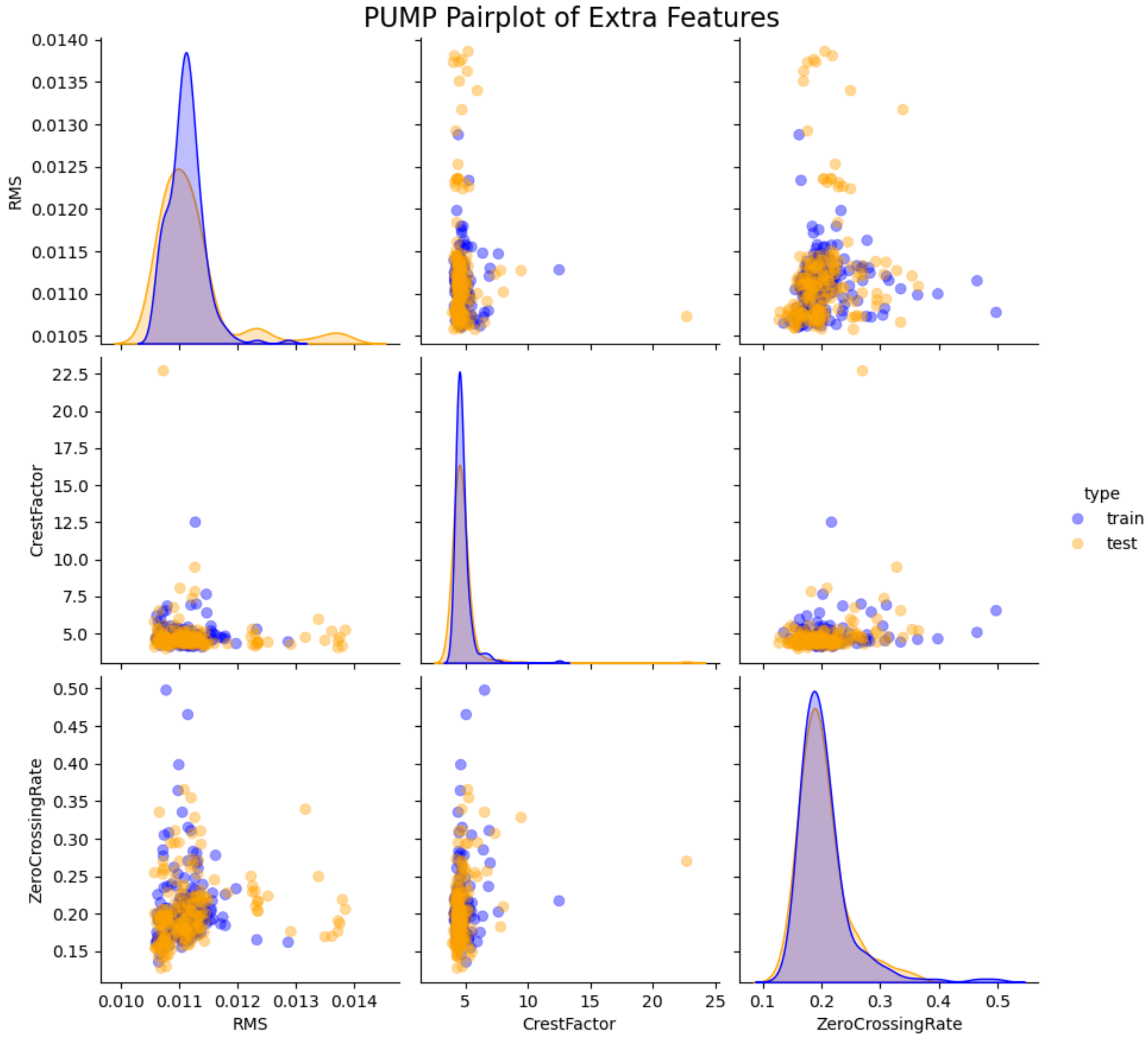
From the above figure, we can more or less confirm our theory from Question 2, that it is harder to discern anomalies for the pump than the valve. The selected features are not so effective at discerning them. Nonetheless, the Variance, and Kurtosis do not overlap perfectly, which suggests there may be some discerning power with these features (although this is definitely less clear than for the valve).

We examined item 12 specifically, and found that it is meaningfully different to the healthy samples in the train dataset. MaxAbs is larger, Skewness is also much larger (although seeing the distribution of skewness in the test dataset, sample 12 might be one of few outliers), and the kurtosis is also meaningfully different in the sample.

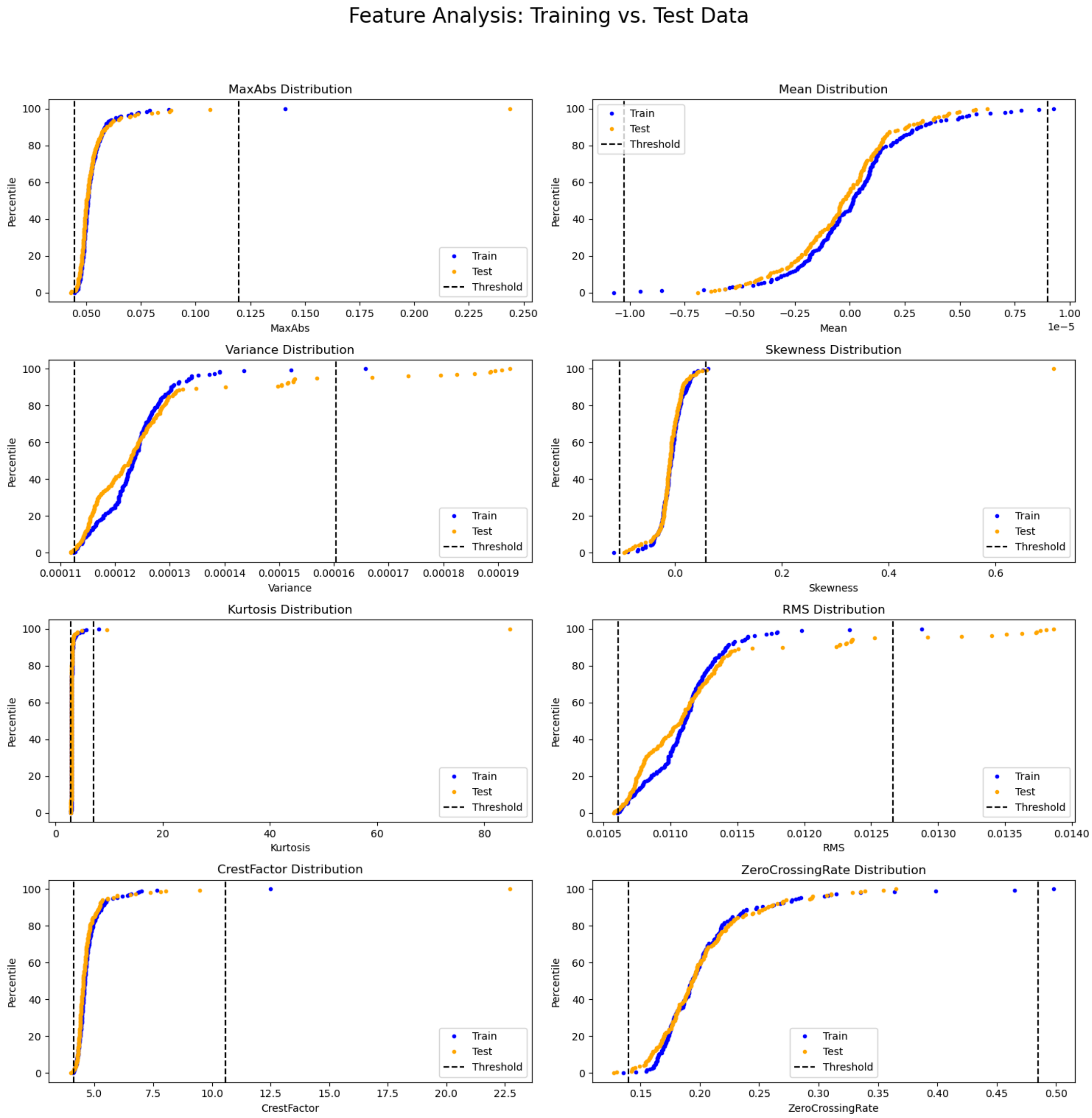

	MaxAbs	Mean	Variance	Skewness	Kurtosis	type
0	0.048767	1.686096e-07	0.000117	-0.021724	2.991884	train
1	0.065735	2.346611e-06	0.000113	0.022433	4.980919	train
12	0.243713	-1.054764e-07	0.000115	0.707751	84.745613	test

Q4 - Find by yourself a feature or a combination of features that help to uncover signals with abnormal behavior. Analyze whether the selected features trigger alarms for similar behavior or if some are specific to particular anomalies.

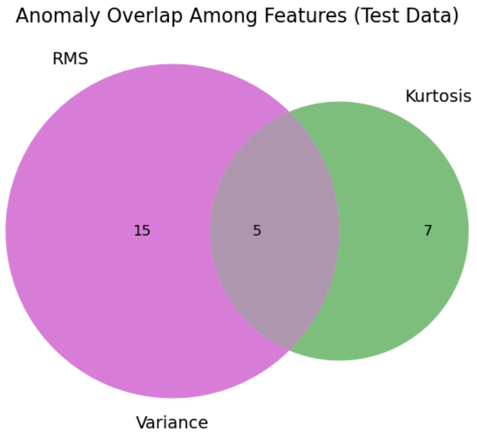
Similarly for the valve, we used RMS, crest factor, and zero crossing rate as new features. The following figure shows a pairplot of these features. RMS shows slight distinction, but crest factor and ZCR don't offer much information.

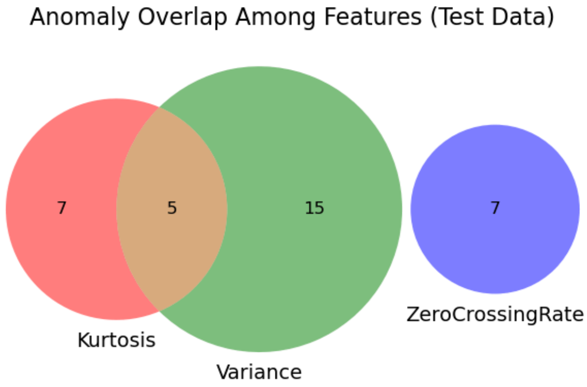


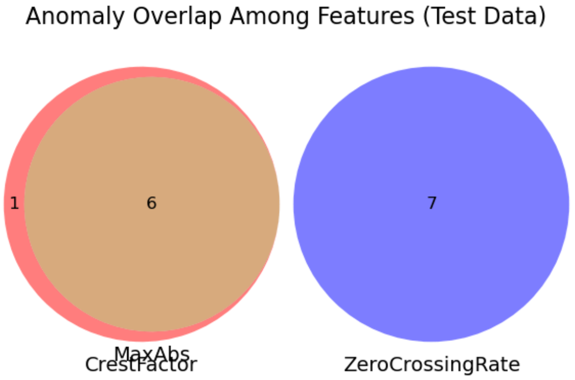
In order to find outliers, we plotted thresholds based on the 0.2 and 99.8 percentile of training data. This plot shows that only the variance and RMS hold any sort of true discriminatory power.

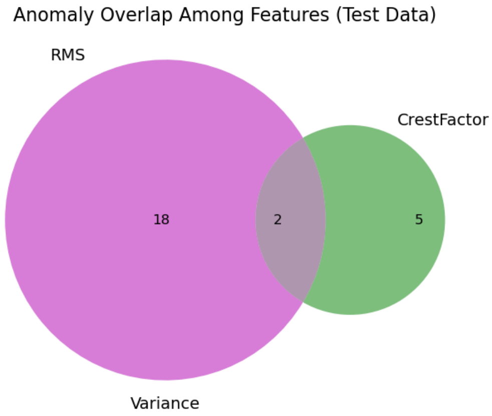


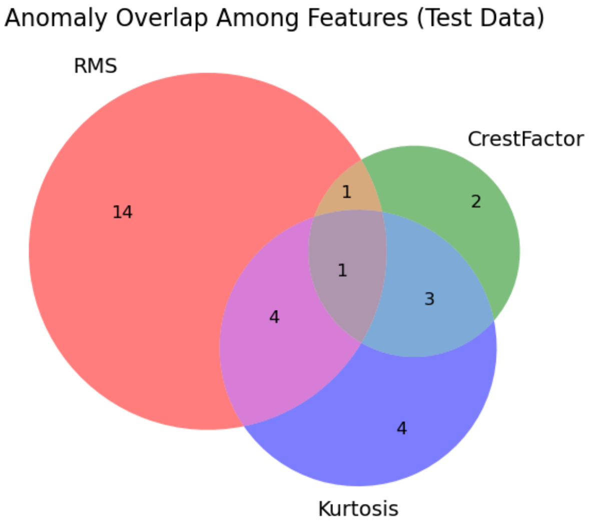
Unlike the valve, it seemed as if some features were specific to certain anomalies, without much overlap. All methods aside from the Mean have found some anomalies. However, only 9 samples have been found as anomalous by at least 3 methods, which suggests that there is not a perfect overlap. This is shown by the following venn diagrams.











Variance and RMS detect signals with the same characteristics, and as previously mentioned, seem to flag anomalies the most often. As seen in the 3rd venn diagram, MaxAbs and CrestFactor also seem to flag the same samples. Kurtosis has some overlap with the Variance/RMS anomalies, but MaxAbs/CrestFactor, as well as the

ZeroCrossingRate, seem to flag entire other sets of samples (plots 2, 3 and 4). Kurtosis also seems somewhat unrelated to both CrestFactor, Zero Crossing Rate, and RMS.

Thus, our conclusions are as follows: in order to minimize the number of False Negatives, which signal an undetected fault in the pump, it is probably safest to include several of the features mentioned. Thus, a feature set of RMS (or Variance), CrestFactor (or MaxAbs), ZeroCrossingRate and Kurtosis is probably the most optimal in order to minimize false negatives, although some further investigation is likely needed to validate, as mentioned previously, whether outliers in some of these signals are actual anomalies (like sample 12), or simply what seems like background noise (like sample 80)

General Questions

Q5 - What are some potential limitations of the method suggested in this exercise for anomaly detection? Answer with at least 3 limitations.

Throughout our work, we noticed several challenges that might arise in using the methods explored in this exercise for anomaly detection. We will explore 5 of our biggest doubts and what we believe are the biggest limitations.

1. **Sample quality**

One of our ground assumptions in our method is that all the samples were collected in roughly the same conditions. That is: using similar (or the same) machines, in roughly the same environment, etc. Unfortunately, as we noticed when listening to some “abnormal” audio files, it sometimes seems like the “anomalies” are in fact just background noise (as is probable in sample 87 of the pump test dataset, for example), due to the fact that the samples were probably taken in an industrial setting where many different noises can be heard.

2. **Statistical assumptions**

Another big assumption that we have made is that the measures we have used are somehow predictive of anomalies, and also are independent from each other. However, we know that some of these statistical features will be correlated with each other (as for example the crest factor is the ratio of the max amplitude and the RMS, two other features we have used). Thus, it is very difficult to actually assess the true performance of our model, and have a risk of overfitting to the training data if we have too many correlated features (which may “support each other” in “wrong” claims of anomaly)

3. **Scalability**

Another challenge is that we have approached this problem by trying to think of interesting statistical measures, and have mostly selected them on whether they are able to separate the training and test dataset well. However, there is a near infinite choice of similar features as the ones we have been using. For instance, one could look at the statistical moments (and other properties) of the Fourier Transforms of the signals, look at more time-domain metrics (peak-to-peak, geometric means, correlation coefficients, …) and we can also imagine all of the interaction terms…

This problem of feature engineering quickly becomes unscalable. Thus, it may be beneficial to do further research into the performance of specific features with labeled data, similar studies that have been done on the subject and the type of features that they have used, and robust success metrics (do we optimize for recall? Precision?)

4. **Difficulty in parameter tuning**

For our exercise we have assumed the same thresholds for all our features. However, finding “the best” threshold for each feature is non-trivial, especially as the problem and the dataset scales.

Q6 - Now that you have developed a set of features and thresholds for a valve/pump, imagine applying them to a different valve/pump. Would the discriminative power change? If so, how do you propose to mitigate it? Justify and provide concrete Details.

It is very probable that the discriminative power will change, because even if the equipment is of the same type, it may exhibit different characteristics due to differences in manufacturing, installation, environment, and usage. Sensor calibration and location will also play a role. As a result, raw statistical measures like kurtosis, time-domain features like MaxAbs, and even frequency domain features (although we have not used them explicitly in our case, values like the Spectral Skew/Kurtosis will also change).

In order to mitigate this problem, there are many possible options (and the optimal solution likely brings together multiple of them). We will explore some below.

1. **Feature standardization**

In order to be able to compare numbers between the machines (the absolute values of which may not be meaningful), it may be more beneficial for our analysis to standardize. This could be especially the case since it seems like most of the features follow (somewhat) gaussian patterns, however this is a relatively strong assumption and if some feature is heavy-tailed this may introduce bias in our features.

Normalisation may help, but since we are especially interested in outliers, it may lead to information loss if there are many outliers.

2. **Use Adaptive Thresholding**

If we want to keep our method of calculating thresholds for a given feature set, it may be needed to implement some sort of adaptive thresholding: that is, to constantly recalculate thresholds based on recent data trends (using a rolling window for example). However, this does mean that this new machine needs to collect its own data, and there is a ramp-up period for new machines before enough new data can be collected. Thus while this may be suitable for a very small network if we deem the method used in this exercise efficient enough, for larger networks it is better to look at option 4 (using ML algorithms using data from many machines).

3. **Manual recalibration of thresholds using new data**

Taken alone, it would require us to put in the same amount of work for the second machine as we did for the first, which is not scalable and not optimal. However, in order to generalise our model to hold its discriminative power for most pumps and valves, we will need the data from many different machines. Thus, this is likely an unavoidable step. Combined with other methods (such as switching to ML techniques like transfer learning or clustering algorithms), collecting a lot of new data can allow us to eventually create what could become a “generalisable” algorithm.

4. **Collect data from many machines and use ML methods (clustering, transfer learning)**

If we are considering scaling our classifier to many valves and pumps, it may be unavoidable to have to complexify our model to not just manually compute thresholds, but switch to new classes of methods like clustering algorithms (such as k-NN or DBSCAN) or transfer learning (which could allow us to fine-tune a pre-trained model on our specific machine). However, it must be noted that for these algorithms to be implemented, much more data will be needed (from a “representative set” of all valves and pumps of interest, for example of a given model)

Q7 - Now imagine you're in a scenario where there are no anomalous samples available. How would you tackle this problem? Answer with an overview of your proposed approach and then provide concrete details regarding the method.

If we do not have access to anomalous samples, it may be challenging to create a model that effectively captures anomalies. It then becomes a problem of modeling “normal” behavior, and then using some unsupervised learning method that can flag any deviation from this “normal data” as potential anomalies. There are several ways to do this, we will explore a couple below.

1. **Using K-Means**

One possible way is to create k clusters using the “normal” data. Then, when gathering new data we can assess the distance with the cluster centroids and beyond a certain threshold, can classify them as anomalies.

2. **Using DBSCAN**

Another way which could be more robust depending on the distribution of samples and what the data “looks like” is to use the DBSCAN algorithm. This has the distinct advantage that if there are some non-convex shapes in the data (non-gaussian patterns can appear depending on our feature set). However, if we have many features this algorithm does not scale well, and we could use method 3.

3. **Using PCA (reconstruction)**

If we are using many features, it may be best to use PCA to try and detect anomalies. This can be done by computing the principal components of our “normal data” and getting a good understanding of the typical “reconstruction error” of the normal data (that is, the information loss when applying PCA and then going back to the higher dimensional space, can be measured by the L2 norm of the difference of the reconstructed data point and the initial feature values for example). Then, we can compute this reconstruction error for all new data points and if we exceed some threshold (eg. some multiple of the mean/median error of the normal data), then we can classify the data as a potential anomaly.

(Disclaimer: got the idea to use PCA reconstruction from ChatGPT, but found it quite enticing and it made a lot of sense to me considering what we spoke about in class)